An Automated Test Assembly Approach Using Item Response Theory to Enhance Evidence of Measurement Invariance

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE
UNIVERSITY OF MINNESOTA
BY

Allison Ward Cooperman

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Dr. Niels G. Waller, Adviser

June 2022

Acknowledgements

If I have learned anything during my graduate school career, it is that it takes a village to complete a doctoral degree. Completing a degree during a global pandemic adds even more levels of complexity. I am indebted to those who have supported me through this journey. Without their guidance, kindness, and humor, I would not be the person that I am today.

First, I am extremely grateful for the support of my advisor, Dr. Niels Waller. I have learned more than I could imagine while working with him, and he has helped me grow as a researcher, teacher, and writer. This dissertation would not have been possible without his guidance. Dr. David Weiss has also been an incredibly important figure in my graduate school career. His mentorship allowed me to explore new research areas and better understand the research process. I also want to express my sincerest gratitude to my other preliminary and final exam committee members — Dr. Seungwon Chung, Dr. Nathaniel Helwig, and Dr. Katerina Marcoulides — who provided invaluable feedback on this dissertation. I am honored to not only have worked with these professors during the dissertation process, but also on other projects during my time at the University of Minnesota. I am a better researcher because of their support.

Next, my time in Elliott Hall has been more special because of my fellow QPM students. I thank Alec Nyce and Justin Kracht for being my unofficial mentors in the program. I am also indebted to Lauren Berry and Kelly Duffy, who have been beside me (literally and figuratively) throughout my graduate school career. Moreover, I want to express my appreciation for Joseph DeWeese, Kayle Donner, Hoang Nguyen, Gretchen Saunders, Matthew Snodgress, King-Yiu Suen, Ming-Him Tai, Raj Wahlquist, Joy

Wang, and Ziming Zhou. These fantastic individuals never failed to brighten my spirits and expand my understanding of quantitative methods.

Numerous other folks at the University of Minnesota have been key factors in my success as a graduate student. I am grateful to Dr. Moin Syed, who helped me build confidence as a scholar and educator. Those in the Psychology graduate student workshop have been instrumental in my PhD journey. I also thank David Olsen, who kindly helped me set up my dissertation project on the LATIS computing cluster. Moreover, I appreciate Rachel Goeller, Amy Kranz, Silke Moeller, Dr. Pat Frazier, and Dr. Alicia Mohr for their support and willingness to answer my many questions. There are many other graduate students, faculty, and staff who I am honored to have worked with at the University of Minnesota — too many to list in the limited space here.

I am also very fortunate to have had numerous mentors guide me through my undergraduate education and to graduate school. Dr. Julia Strand was my first research advisor at Carleton College. She is a fantastic mentor for women in STEM, and I know that I would not be here today without her support. I am next grateful to Drs. Noah Berman and Elise Chor, who helped me find and cultivate my interests in quantitative psychology. They taught me how to be a compassionate and rigorous researcher. I also thank Drs. Daniel Geller, Sabine Wilhelm, Lindsay Chase-Lansdale, Terese Sommer, and Terri Sabol for their guidance, kindness, and investment in my future.

I am forever lucky to have a wonderful support network outside of the academic sphere. One of the great joys of graduate school journey has been developing friendships in and outside of the Twin Cities. These friends have helped me destress, supported my endeavors (whether academic, fitness-related, or otherwise), and have made me happier

than they could ever know. I particularly want to thank Amanda Bridges. She has been my cheerleader from the initial graduate school applications to defending the dissertation. I am a better person for knowing her.

This dissertation would not be possible without the unwavering support of my family: Mary and Jon Cooperman, Sarah Cooperman, Ward and Richard Griffiths, Ruth and Jack Cooperman, Ward Griffiths, Elaine and Avi Betan, Fayanne Betan and Meny Har, and Joan MacDonell and Maury Strand. They taught me to aim high, to work hard, and were always available to talk me through the most difficult moments. My family has been my rock and I am so lucky to have them.

Most importantly, I am eternally grateful to my husband, Max Strand. While I pursued a degree in quantitative psychology, he pursued a degree in *supporting* a graduate student. We can all agree that he graduated at the top of his class! There are no words strong enough to express all that he has done, and all that he means to me. This is for you, Max.

Dedication

For the strong women in my life, who taught me to speak up, to stand out, and to know my worth.

For my grandfather, Dr. Jack Cooperman.

Abstract

When creating a new test to measure a latent trait, test developers must select items that together demonstrate desirable psychometric properties. Automated test assembly (ATA) algorithms allow test developers to systematically compare possible item combinations based on the test's goals. ATA algorithms afford flexibility to incorporate various psychometric criteria for evaluating a new test. However, few algorithms have integrated analyses for item- and test-level bias, particularly within the item response theory framework. This dissertation proposes an approach that balances common indices of test score precision and model fit while simultaneously accounting for differing measurement models between two groups.

Three Monte Carlo studies were designed to evaluate the proposed method (termed "Unbiased-ATA"). The first study found that in many testing scenarios, Unbiased-ATA appropriately constructed tests with evidence of measurement invariance (MI), item fit, and test information function alignment. Importantly, Unbiased-ATA's performance depended on the accuracy of both the DIF detection method and item parameter estimation. The second study revealed that differentially weighting the Unbiased-ATA objective function criteria did not substantially affect the method's performance. The final study found that Unbiased-ATA produced tests with stronger psychometric properties than an objective function based solely on test score precision. Yet adding a criterion for item-level MI did not noticeably improve tests' psychometric strength above and beyond a criterion for test-level MI. Future directions for integrating ATA, test bias, and test fairness more broadly in psychological and educational measurement are discussed.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Classical and Modern Test Theory Approaches for Test Development

Most psychological and educational tests are designed to measure one or more socioemotional or cognitive traits. In these cases, the trait is latent, meaning that it is unobservable and cannot be directly measured. For example, a clinician might gauge an individual's level of depressive symptoms during a therapy session, or a researcher might evaluate personality trait levels among various groups. After determining the test's goal, test developers generate a large set of items that are thought to characterize the trait of interest. This item bank is used to identify an optimal combination of items that best addresses the test's purpose (Crocker & Algina, 2008).

How do test developers operationalize this "optimal" combination of test items? Psychometricians advocate that test items, and the corresponding test scores, be evaluated in terms of two broad characteristics. First, test scores should be precise. Assuming no changes in the underlying trait, individuals should obtain relatively similar scores on each test administration. Test score precision thus signifies the extent to which variation in test scores is a function of true variation on the underlying trait rather than measurement error (Crocker & Algina, 2008; Embretson & Reise, 2000). Second, test scores should be valid, such that the item content and test score structure appropriately reflect the test developer's conceptualization of the trait. For example, the latent structure of a test that purports to measure the Big Five personality factors (Costa & McCrae, 2008; L. R. Goldberg, 1993) should reveal five relatively distinct dimensions. Methods to evaluate both test score precision and validity have largely developed from either classical test

theory (CTT) or item response theory (IRT) frameworks (Crocker & Algina, 2008; Embretson & Reise, 2000; Lord & Novick, 1968; McDonald, 2013).

Test Score Precision

In CTT, an examinee's test score (X) is conceptualized as the sum of their true score (T) and error (E). The true score represents the examinee's expected score across (infinitely) repeated test administrations with no memory effects, and error represents random fluctuations from this expected score. Test score precision in CTT is then assessed with the reliability coefficient,

$$\rho_{XT}^2 = \frac{\sigma_T^2}{\sigma_X^2}, \tag{1}$$

or the proportion of observed score variance in the test that is attributable to the true scores (Crocker & Algina, 2008). In practice, the true reliability coefficient cannot be computed for a given set of test scores. Instead, researchers use a variety of estimates to gauge test score reliability, including Cronbach's α (Cronbach, 1951), McDonald's ω (McDonald, 2013), and test-retest estimates (Crocker & Algina, 2008).

Although CTT reliability estimates are popular in psychometric research (Cortina et al., 2020; Dunn et al., 2014), they are largely limited by their dependence on a given sample (Embretson & Reise, 2000). Specifically, "reliability is a property of the scores on a test for a particular group of examinees" (Crocker & Algina, 2008, p. 144; see also Embretson & Reise, 2000; Rindskopf, 2001; Schroeders et al., 2016). This group-level analysis reduces the generalizability of test score precision, such that a high test score reliability estimate in one sample does not imply a similar reliability estimate in another sample (e.g., Hambleton & Jones, 2005; Rindskopf, 2001). Relatedly, all examinees in a sample will have the same standard error of measurement regardless of their true trait

value (De Champlain, 2010; Embretson & Reise, 2000). A constant standard error is largely unrealistic; for example, recent research illustrates how trait estimation accuracy often decreases for examinees with extreme trait values (De Champlain, 2010).

IRT methods' alternative approach to test score precision largely overcomes CTT's limitation of group-level estimation. IRT uses latent variable models to measure the probability of a particular item response as a function of an examinee's underlying ability (Birnbaum, 1986; Embretson & Reise, 2000; Lord & Novick, 1968). In IRT, test score precision is evaluated using the level of statistical information that an item provides about the latent trait (Embretson & Reise, 2000). Information is typically quantified using Fisher information, computed as the variance of the first derivative (with respect to the latent trait) of the log-likelihood function (DeGroot & Schervish, 2012, p. 515). Assuming that items are locally independent (i.e., an individual's responses to two or more items are unrelated after controlling for the latent trait), test information can be computed as the sum of the item information values. Importantly, information is proportional to the standard error of measurement, so higher precision at a given trait value corresponds to a smaller standard error (Embretson & Reise, 2000; Hambleton & Jones, 2005).

A major distinction between CTT and IRT is that in the latter framework, information highlights individual-level precision. Information can vary along the latent trait continuum, meaning that the test might provide more precise test scores and smaller standard errors for individuals with different latent trait values. Moreover, IRT methods place items and examinees on a common metric, which facilitates the equating of test scores for examinees assessed by different test versions (Embretson & Reise, 2000;

Whitely, 1983). Test score precision estimates are then not limited to a single sample of examinees. This common metric also allows test developers to explicitly build tests that provide maximal precision at specified trait values, such as a cut-off value for a classification test. Therefore, using IRT methods affords numerous advantages over CTT for measuring test score precision during test development (Embretson & Reise, 2000).

Test Score Validity

Whether using a CTT- or IRT-based test construction process, test score precision is not the only important metric for evaluating a given test. Sufficient evidence of test score validity is also necessary to garner robust inferences regarding an individual's latent trait value. In psychometrics, test scores are typically evaluated using construct validity (Cronbach & Meehl, 1955; Kane, 2001; Loevinger, 1957; Messick, 1989), which at its most basic level answers the following question: Does the test actually measure the trait(s) that the test purports to measure?[1] Evidence of construct validity is necessary whenever the underlying trait cannot be uniquely operationalized, a problem that is particularly prevalent in social science research (Cronbach & Meehl, 1955; Gorin, 2007; Loevinger, 1957). Notably, test score precision is necessary, but not sufficient, evidence for construct validity.

Using terminology from Loevinger (1957), construct validity generally comprises three main categories: (a) substantive validity, (b) structural validity, and (c) external validity. First, substantive validity is the extent to which the item content comprehensively covers all aspects of the hypothesized trait (Loevinger, 1957). Cronbach and Meehl (1955) denoted this as content validity, proposing that a high degree of

[1] Note that this definition is debated in the construct validity literature (e.g., Borsboom et al., 2004)

content validity indicates that the items represent the "universe of possible items" reflecting the trait (p. 282). Second, structural validity is the extent to which the psychometric structure aligns with the theorized trait structure. Structural validity also encompasses test score precision, evaluating the degree to which measurement and method error influence the test scores (Loevinger, 1957). Finally, external validity (also called criterion-oriented validity; Cronbach & Meehl, 1955) is the extent to which test scores relate in hypothesized directions and magnitudes to other measures (Loevinger, 1957). Evidence of this type of validity might include strong positive relationships with tests thought to measure a similar trait (termed "convergent validity"), negative relationships with tests thought to measure a dissimilar trait (termed "divergent validity;" Campbell & Fiske, 1959), and predictive power for a criterion administered at a future occasion (termed "predictive validity;" Cronbach & Meehl, 1955).

As an example, consider the development of a new test to measure adolescent Agreeableness, considered one of the five global dimensions of personality in the Big Five model (Costa & McCrae, 2008; L. R. Goldberg, 1993). Some researchers theorize that Agreeableness is comprised of six facets, such as trust and altruism (e.g., Costa et al., 1991; Piedmont & Weinstein, 1993). Using this trait conceptualization, the test shows evidence of substantive validity if the items broadly address all six facets. Moreover, a factor analysis of the sample of test scores should uncover six related dimensions (one for each facet) to indicate evidence of structural validity. The test developers might have their adolescent sample also complete the NEO-PI-R (Costa & McCrae, 2008) and a self-report measure of aggression. Correlating the three sets of test scores should reveal a positive relationship between the new Agreeableness measure and the NEO-PI-R

Agreeableness subscale, as well as a negative relationship between the new Agreeableness measure and the aggression measure (Gleason et al., 2004). This pattern of correlations provides evidence for external validity.

An important consideration when evaluating test score validity is that validity only applies to a particular inference or test goal (American Psychological Association et al., 2014; Cronbach & Meehl, 1955; Kane, 2001; Loevinger, 1957; Messick, 1989). Specifically, evidence supporting the use of test scores for one type of inference does not automatically translate to evidence supporting a new use of the test scores (Cronbach & Meehl, 1955; Flake et al., 2017; Kane, 2013; Mosier, 1951). For example, a test might have been designed to measure anxiety among college-age students. If researchers want to use the test among a sample of middle school students, they need to reevaluate their evidence for construct validity using this new sample. Without a new validity study, there is no guarantee that the inferences drawn from the test scores of the middle school students accurately reflect anxiety. In this way, validity is considered a property of the test score inferences rather than a property of the test itself (American Psychological Association et al., 2014; Cronbach & Meehl, 1955; Kane, 2013; Messick, 1995).

Since Cronbach and Meehl (1955) and Loevinger's (1957) seminal papers on construct validity, researchers have refined and disputed numerous components of this important concept. One issue in operationalizing construct validity concerns the extent to which test development procedures should consider the consequences of test scores (Kane, 2001; Newton & Baird, 2016; Shepard, 2005). Some researchers (e.g., Popham, 2005; as cited in Kane, 2001) have argued that a test's ability to accurately measure a latent trait should be evaluated separately from the decisions made using the test scores

(Kane, 2001). However, many others (e.g., Kane, 2001, 2006, 2013; Moss, 2013; Shepard, 2005) instead contend that these decisions are central to the concept of construct validity. In fact, Messick (1989) differentiated between the "evidential" and the "consequential" aspects of psychological and educational tests, the former referring to the psychometric properties of the test and the latter referring to the usage of the test scores in practical applications. This differentiation highlights the importance of confirming that a test "achieves its goals without unacceptable negative consequences" (Kane, 2006, p. 15; as cited in Moss, 2013, p. 92). Such negative consequences might include incorrect categorizations that disproportionately harm individuals with certain characteristics.

Although beyond the scope of the current review, there remain numerous avenues for future research regarding the role of construct validity in psychometrics. For example, unanswered questions include (a) the degree to which researchers should uncover the causal mechanisms connecting the (presumed-to-exist) trait to the test responses (Borsboom et al., 2004; Embretson & Gorin, 2001; Whitely, 1983), (b) the role of personal ethics in validity definitions (Newton & Baird, 2016), (c) the philosophical underpinnings of measuring latent variables (Borsboom et al., 2004, 2009; Michell, 1990, 2021), and (d) ways for researchers to communicate with practitioners regarding the intended usage of test scores (Gebru et al., 2020; Moss, 2013). However, these debates do not negate the critical role that the validation process plays when constructing a new test, or when seeking to use a test for a new purpose (Newton & Baird, 2016). Validity permeates all components of the test development process, from the initial trait conceptualization to the decisions made based on the test scores (Smith, 2005). Therefore, construct validity must be established when developing tests of latent traits to

garner plausible inferences about what the test is attempting to measure (Cronbach & Meehl, 1955; Flake et al., 2017; Loevinger, 1957; Messick, 1995).

Automated Test Assembly

After establishing the criteria used to evaluate a new psychological or educational test, researchers must compare a wide range of possible tests generated from different item combinations. For instance, a test developer might seek to create a test from an item bank with n items. Using the binomial coefficient, there are $n!/[k!\,(n-k)!]$ unique ways to select a k-item test. With a 50-item bank, this works out to 10,272,278,170 possible item combinations for a 10-item test. Clearly, given a large item bank and sufficiently long test, it quickly becomes impractical (if not impossible) for researchers to manually evaluate each possible item combination (Schroeders et al., 2016).

Instead, researchers have turned to automated test assembly (ATA) approaches to more efficiently select an "optimal" item combination for a new test (van der Linden, 2005). In the machine learning framework, item selection with ATA is a type of combinatorial optimization problem. Specifically, the algorithm searches across a finite set of item combinations with the goal of maximizing a predefined criterion subject to a set of constraints (Dawande et al., 2000; Kellerer et al., 2004; Luo, 2020; Schultze & Eid, 2018; van der Linden, 1998). ATA has been compared to the "knapsack problem" (Kellerer et al., 2004), where the goal is to choose a collection of objects that can fit into a knapsack, while not exceeding the maximum knapsack volume (Kellerer et al., 2004; Schultze & Eid, 2018, p. 177). In the context of test construction, researchers want to choose the combination of items that produces the most favorable psychometric

properties (as defined by the test developer) while controlling characteristics like test length and plausible pairwise item combinations.

Generally, ATA algorithms comprise three main components (Cor et al., 2009; Luo, 2020; van der Linden, 1998, 2005). First, ATA requires a set of decision variables to identify a given item combination. Typically, these are a set of binary variables $x_1, \ldots, x_n$ where $x_j = 1$ if the j^{th} item is included in the test. The second ATA component is the objective function, which is a mathematical expression quantifying the important psychometric properties that the test should fulfill. The objective function might comprise either one criterion or multiple criteria as a weighted linear combination (e.g., Stocking et al., 1998; Yarkoni, 2010). The ATA algorithm then seeks to maximize (or minimize) the value of the objective function to identify the "best" test. Third, ATA can incorporate a set of constraints that ensure that the resulting test matches the predetermined test blueprint (van der Linden & Adema, 1998). For instance, test developers might restrict the number of items for a certain content area (e.g., Stocking et al., 1998), or remove item combinations that engender improper factor analytic solutions (Jankowsky et al., 2020).

Scanning the ATA components, the objective function arguably provides psychometricians with the greatest control over the new test's measurement quality. ATA algorithms from an IRT framework often incorporate objective function criteria related to item and test information (e.g., Armstrong et al., 1998; Diao & van der Linden, 2011; Harel & Baron, 2019; Huitzing et al., 2005; Ishihara et al., 2019; Levis et al., 2016; Luecht, 1998b; Martín-Fernández et al., 2021; van der Linden, 1998; van der Linden & Adema, 1998). Recent studies with structural equation modeling (SEM) have instead used model fit indices to compare competing test forms (e.g, Browne et al., 2018; Raborn

et al., 2020; Schroeders et al., 2016; Schultze & Eid, 2018). For instance, Raborn and authors (2020) employed the comparative fit index (CFI; Bentler, 1990) in a comparison of three common ATA algorithms. Other objective functions with ATA have incorporated the sum of unique item variances in a structural equation model (to quantify reliability, or external validity if a second measure is predicted; Browne et al., 2018), an adjusted R^2 when regressing a total sum score on the individual items (Gonzalez, 2020; Yarkoni, 2010), or a content validity index (Martín-Fernández et al., 2021).

Linear Programming

Given a linear objective function, ATA algorithms can be solved using linear programming (LP) optimization (Finkelman et al., 2010; Luo, 2020; van der Linden, 1998, 2005; van der Linden & Adema, 1998). LP is a type of convex optimization (Yang, 2018) and encompasses a wide range of methods that differ based on the type of decision variables. For example, numerous researchers (e.g., Harel & Baron, 2019; Luo, 2020; Martín-Fernández et al., 2021; van der Linden, 1998; van der Linden & Li, 2016) have employed mixed integer programming (MIP), requiring that at least some of the decision variables are integers. However, these methods are often further simplified to 0-1 LP algorithms because the decision variables in ATA are typically binary (indicating whether or not an item is included in a given test; Finkelman et al., 2010; van der Linden & Adema, 1998).

Briefly, LP optimization searches across the space of possible item combinations and selects the combination with the maximum (or minimum) objective function value. These methods are advantageous due to their flexibility (van der Linden & Li, 2016) and the guarantee that the algorithm will find the global optimal solution if one exists (Luo,

2020; van der Linden, 1998). Although numerous solvers are available (e.g., Boyd & Vandenberghe, 2004; Diao & van der Linden, 2011; Luo, 2020), LP methods can quickly become complex and time-intensive, especially with millions of possible test forms under consideration. Specifically, these optimization problems are NP-Hard, meaning that "a slight change to the model or data (e.g., adding a cognitive constraint, number of forms, test length, or different item bank sizes) can dramatically alter the running time of the [LP] method" (P.-H. Chen, 2017, p. 230).

Metaheuristic Algorithms

To reduce the computational burden of LP optimization, researchers have instead drawn on heuristic algorithms for ATA (e.g., P.-H. Chen, 2016; Luecht, 1998b; Raborn et al., 2020; Schroeders et al., 2016). Heuristic algorithms use a dynamic neighborhood of plausible item combinations. This neighborhood substantially reduces the search space for the algorithm, and thus the computation time. After selecting the "best" item combination (as denoted by the objective function value) within a given neighborhood, the algorithm then updates the neighborhood and continues its search. For example, the neighborhood might shift by incorporating tests with a one-item difference from the current solution (Luo, 2020; G. A. Marcoulides & Drezner, 2004; K. M. Marcoulides, 2018, 2020; Talbi, 2009). Although heuristic algorithms might exhibit faster convergence times, the trade-off is that they are not guaranteed to arrive at the global optimal solution (if one exists; Leite et al., 2008; Luo, 2020; van der Linden & Li, 2016).

Heuristic algorithms are generally formulated for a specific problem, such as the construction of a single test type. Alternatively, researchers can use metaheuristic algorithms, which are applied to a wider collection of optimization goals (Talbi, 2009).

Common metaheuristic algorithms in ATA research include ant colony optimization (ACO; Colorni et al., 1991; Dorigo & Stützle, 2004), Tabu search (Drezner et al., 1999; Glover, 1986; G. A. Marcoulides & Drezner, 2004), genetic algorithms (Fraser, 1957; D. E. Goldberg, 1989), and simulated annealing (Cerny, 1985; Kirkpatrick et al., 1983). For these algorithms, researchers can modify the objective function and related constraints for their own test construction goals. Again, metaheuristic algorithms are not guaranteed to find the optimal solution as indicated by the objective function value (if one exists). Still, there is growing evidence that these methods perform well in combinatorial problems for item selection (e.g., P.-H. Chen, 2017; Gonzalez, 2020; Raborn et al., 2020; Schroeders et al., 2016; Schultze & Eid, 2018).

Using either 0-1 LP or metaheuristic algorithms, ATA offers numerous advantages over manual item selection processes. For one, ATA algorithms can more effectively search across a greater number of item combinations than would be feasible with a manual approach (Cor et al., 2009; K. M. Marcoulides & Falk, 2018; Yarkoni, 2010). As a result, the test scores from an ATA algorithm are likely to demonstrate more desirable psychometric properties when applied to a new sample (i.e., show strong generalizability; K. M. Marcoulides, 2018). Additionally, ATA allows researchers to incorporate multiple criteria into the objective function (van der Linden, 1998; Veldkamp, 1999). For instance, researchers have integrated numerous aspects of the validation process directly into automatic item selection (e.g., Browne et al., 2018; Martín-Fernández et al., 2021; Raborn et al., 2020; Schroeders et al., 2016). However, an important limitation of the extant research is that relatively few ATA algorithms have

explicitly evaluated whether the proposed test is measuring the latent trait(s) equally well for all intended examinees.

Test Bias and Fairness

Test Bias

Test developers often look for statistical evidence that examinees' characteristics (unrelated to the trait of interest) do not systematically influence the measurement and predictive power of test scores. When the test goal is to identify individual differences on a latent trait, this statistical evidence is commonly evaluated using measurement invariance (MI) analyses. MI indicates that the underlying measurement model—relating the item responses to the hypothesized latent trait—is equivalent across all intended groups of examinees. In other words, factors unrelated to the conceptualized trait should not differentially impact the test scores among examinees from separate groups (Mellenbergh, 1989; Meredith, 1993; Meredith & Millsap, 1992).

As a concrete example, consider a test that purports to measure one aspect of math ability and is written in English. The test might include word problems that require examinees to read a long item stem before developing a response. For such items, test developers want to ensure that students' English language ability does not substantially influence the test scores (that are believed to reflect a singular latent ability for the type of math skill). The test developers might then compare the underlying measurement model for the test between a group of native English speakers and students for whom English is a second language. Equivalent measurement models suggest that any comparison of the math test scores between groups reflects differences in the examined math ability rather

than a conflation of math, English language, and potentially other latent abilities (Meredith, 1993; Millsap, 2010; Widaman & Reise, 1997).

MI is not the only type of test bias that is routinely explored during test development. Rather, psychometricians often seek evidence of prediction invariance for tests that are used to infer individuals' performance in the future (see Kuncel & Klieger, 2012, or Millsap, 1997, for a review). If a test demonstrates prediction invariance, then two groups have the same estimated equation when regressing scores on a criterion on the test scores (Borsboom et al., 2008, p. 76; Millsap, 1997). Furthermore, test developers might seek evidence of selection invariance for tests that choose candidates based on a predetermined cut-point. Selection invariance means that the test's classification accuracy (e.g., sensitivity, specificity) is equivalent for subgroups (Borsboom et al., 2008, p. 77).

The three types of invariance—measurement, prediction, and selection—are considered evidence against test bias. Together, these three perspectives focus on the psychometric properties of a given test (Kline, 2013), and have facilitated numerous statistical analyses for comparing item parameters and test structures among groups (Camilli, 2013; Zwick, 2019). For example, educational testing companies routinely assess whether individuals from two groups with the same underlying trait score have similar probabilities of responding to a given item (Dorans, 2013). Significant research, particularly in the fields of Industrial/Organizational Psychology and machine learning, has also evaluated differences in predictive accuracy as a function of group membership (e.g., Hutchinson & Mitchell, 2019; Jones et al., 2020; Kuncel & Klieger, 2012).

Test Fairness

However, a growing literature advocates for a broader construct of test fairness that extends beyond the psychometric properties related to test bias. Specifically, there has been a shift "from a technical view of fairness (as bias) to an embedded view of fairness (as situated within contexts)" (Poe & Elliot, 2019, p. 3; see also Dorans, 2011). Test fairness in this perspective not only encompasses statistical evidence of equivalent measurement and prediction models, but also examines the impact of test score inferences (Baharloo, 2013; Karami, 2013; Kline, 2013; McNamara & Ryan, 2011; Messick, 1998) and tests' relationships to the sociocultural systems affecting examinees (Dorans, 2011; Poe & Elliot, 2019). The ways that test fairness is evaluated can depend on the test's goals. For instance, test developers might address whether examinees have been given appropriate access to learn and demonstrate knowledge of the trait for cognitive assessments (American Psychological Association et al., 2014; Xi, 2010), and also address the relevance of items to different examinees for personality assessments (Randall, 2021; Randall et al., 2022). Regardless of the test type, test developers and users should consider the role of equity both before and after test administration, beyond that which can be evaluated using statistical methods (Elliot, 2016; Haney & Hurtado, 1994; Xi, 2010).

Messick's (1989, 1998) decomposition of test score meaning helps to further clarify what is meant by the term "test fairness" (Kline, 2013). As previously mentioned, Messick (1989) distinguished between the evidential and consequential aspects of test scores. The evidential aspect concerns the measurement quality of the test, referring to "[c]onstant or systematic error due to group membership or some other nominal variable

in the estimation of scores on psychological tests or on performance criteria." On the other hand, the consequential aspect "refers to intended or unintended consequences of test use that can be evaluated from social perspectives about distributive justice" (Kline, 2013, p. 204). Kline (2013) proposed that the evidential aspect denotes *test bias*, whereas the consequential aspect denotes *test fairness*. Importantly, test fairness not only considers bias (i.e., the measurement properties of a test), but also how (and why) tests are being implemented and interpreted to ensure that certain groups are not disproportionately harmed in the process (Kline, 2013; Messick, 1989, 1998). A similar distinction has been made in the machine learning literature, separating statistical from societal bias (Mitchell et al., 2021). Some researchers have even further expanded upon this dichotomization to promote the explicit integration of social justice into the conceptualization of test fairness (see McNamara & Ryan, 2011).

Fairness plays an integral role not only in the analysis and usage of test scores, but also starting from the initial conceptualization of the latent trait (Randall, 2021; Slomp, 2016). Whether knowingly or not, researchers instill their own perceptions, biases, and misconceptions when defining a psychosocial trait (Gorin, 2007; Haney & Hurtado, 1994; Randall, 2021). Researchers have thus argued that developing fair tests begins with a construct validity framework that promotes equity and inclusion (Elliot, 2016; Randall, 2021; Randall et al., 2022; Slomp, 2016). To do so, one must ensure that different cultures, experiences, and values are represented in the construct definition (Bennett, 2022; Elliot, 2016; Randall, 2021) and are not deemed "construct-irrelevant." Specifically, removing context does not necessarily create a "neutral" test, but can implicitly preference historically privileged groups (Randall, 2021). Relatedly, construct

definitions should incorporate the external social and political factors that might affect examinees' development (Mislevy, 2018; as cited in Oliveri et al., 2020).

Although researchers have attempted to outline the multifaceted components of test fairness, individual perceptions of what constitutes equity preclude a concise definition of this construct (Camilli, 2013; Elder, 1997; Mitchell et al., 2021; Xi, 2010; Zwick, 2019). For instance, personal values, ethics, and biases can affect every step of test development and administration (Elder, 1997). What one person considers to be a "fair" test may easily differ from another person's opinion, as showcased by ongoing disagreements between psychometricians and non-statisticians regarding equity in psychometric testing (Zwick, 2019). As a result, whereas researchers may be able to operationalize and measure test *bias*, operationalizing the more subjective construct of test *fairness* is more difficult (Kuncel & Klieger, 2012), and decades of research have not produced a universal understanding of this construct (Poe & Elliot, 2019). In fact, Kuncel and Klieger (2012) posit that fairness "is not necessarily measured or measurable…[or needs] to be rationally or empirically based" (p. 2). Some researchers thus contend that it is "impossible for a test to be perfectly fair for the intended use(s)" (Xi, 2010, p. 148).

Another perplexing issue in the test fairness literature concerns the relationships among the three types of invariance used to establish that a test is "unbiased." Specifically, researchers have shown that under certain common conditions, tests that demonstrate MI simultaneously violate the requirements for prediction invariance (Millsap, 1997, 2007) or selection invariance (Borsboom et al., 2008). Inconsistencies between MI and prediction or selection invariance can have substantial consequences when establishing evidence of test fairness (Borsboom et al., 2008). For example, can one

justify using a measurement-invariant test to decide whom to admit to a program if that test has differential selection accuracy by groups? The extent to which the invariance relationships will affect test fairness investigations depends on the test's goals. Namely, test developers seeking to solely measure individual differences on a latent trait (and not for use in selection decisions) might focus on MI at the expense of prediction or selection invariance. It is therefore imperative that test developers clearly outline the appropriate uses of the resulting test scores.

Researchers also disagree on whether and how test fairness should fit into the larger construct validity framework. Xi (2010) outlined three relationships between fairness and construct validity that have emerged from the literature: (a) fairness and validity as separate entities (e.g., Haney & Hurtado, 1994; Tierney, 2014), (b) fairness as comprising validity (e.g., Kunnan, 2000, 2004, 2007), and (c) validity as comprising fairness (e.g., American Psychological Association et al., 2014; Poe & Elliot, 2019; Sireci & Rios, 2013; Willingham, 1999; Willingham & Cole, 2013; Xi, 2010). The third perspective, which is commonly adopted in the literature, explicitly asserts that test fairness is necessary for evidence of test score validity. Conceptualizing fairness as an integral component of validity adds to a growing literature that advocates for understanding the consequences of test administration when seeking evidence of construct validity (American Psychological Association et al., 2014; e.g., Kane, 2013; Messick, 1998; Shepard, 2005). Yet even these three seemingly comprehensive perspectives have received pushback. For example, some authors (e.g., Davies, 2010) consider fairness and validity to be so intertwined that it is unnecessary to even differentiate between the two concepts.

Scanning the extant literature on test fairness and psychometrics, test fairness remains a nebulous and developing topic. To clarify the competing definitions, the current project uses the following terminology. First, drawing from Kline (2013), *test bias* refers to statistical evidence of differing measurement models as a function of examinees' characteristics that are unrelated to the trait(s) of interest. Test bias thus reflects the test's psychometric properties (Kline, 2013; Kuncel & Klieger, 2012; Meredith, 1993; Millsap, 1997). Second, *test fairness* refers to the more holistic conceptualization that expands beyond test bias to additionally consider the equitable access to, usage of, and decisions drawn from psychometric testing (American Psychological Association et al., 2014; Karami, 2013; Kline, 2013; McNamara & Ryan, 2011; Messick, 1989, 1998).[2]

Measurement Invariance

Considering these definitions, not all aspects of test fairness are within a psychometrician's control when applying ATA. For example, test developers might imbue their conceptualization of the measured trait with systematic biases concerning idealized ability and psychosocial characteristics (Gorin, 2007; Haney & Hurtado, 1994, p. 240; Randall, 2021). Such biases can then impact the item content or theoretical latent trait structure. Moreover, even if psychometricians do their best to ensure comprehensive construct representation and clearly delineate the appropriate usage of their test, it is not guaranteed that the test scores will be used or interpreted in an equitable manner.

[2] It bears repeating that test fairness is a subjective, "unmeasurable" construct (Kuncel & Klieger, 2012). The proposed definition serves to distinguish bias from fairness to orient the subsequent analyses, rather than definitively argue for a particular position within the test fairness literature.

Therefore, ATA is most amenable to addressing test bias analyses that evaluate different item combinations (after a bank of items and a trait structure have been defined). These analyses are often included in the test development process (Cole & Zieky, 2001; Dorans, 2013; Millsap, 1997; Zwick, 2019), and can be modified for inclusion in an ATA objective function.

Because the current project focuses specifically on constructing tests that seek to measure individual differences on a latent trait, test bias is considered here in terms of MI. As previously noted, MI indicates that the test measures the same trait, in the same way, for different groups (Kline, 2013; Meredith, 1993; Millsap, 1997, 2007; Widaman & Reise, 1997). Specifically, MI holds when (Mellenbergh, 1989; Millsap, 2007, p. 463)

$$P(\mathbf{X}|\mathbf{W},\mathbf{V}) = P(\mathbf{X}|\mathbf{W}), \tag{2}$$

where $\mathbf{X}$ is a random vector of observed variables, $\mathbf{W}$ is a vector of latent variables, and $\mathbf{V}$ is a vector of group-defining characteristics. In other words, after controlling for the latent traits, the observed variable responses and the group characteristics are statistically independent. Without evidence for MI, test score differences might reflect variability in measurement error, or the effects of factors not represented in the defined construct, rather than true differences on the latent trait (Haladyna & Downing, 2005; Meredith, 1993; Millsap, 1997, 2007).

In the SEM literature, researchers have proposed four levels of MI that can each be empirically tested when analyzing covariance matrices. First, configural invariance indicates equivalent factor structures across groups, meaning that the same item sets are salient markers of a given latent trait. With configural invariance, the factor loadings, intercepts, and unique item variances can differ among groups. Second, weak (metric)

invariance indicates equivalent factor structure and loadings across groups. Third, strong (scalar) invariance indicates equivalent factor structure, loadings, and intercepts across groups. Finally, strict invariance adds the constraint of equivalent unique item variances. For each of these four MI levels, the presence of one level implies that all lower levels hold. In other words, evidence of strong invariance implies evidence of both weak and configural invariance (Kline, 2013; Meredith, 1993; Widaman & Reise, 1997). Partial invariance is also possible, wherein some parameters (e.g., factor loadings when testing weak invariance) are constrained across groups and others freely estimated (Widaman & Reise, 1997). Determining the level of MI within the data thus often requires an iterative process of model comparisons, applied either at the item or the full test level.

Methods for examining MI have developed in tandem within both the SEM and IRT literatures. In the SEM framework, MI is typically evaluated using multiple-group confirmatory factor analysis (MGCFA) models (Meredith, 1993; Widaman & Reise, 1997). Although numerous MGCFA models are available, such as the multiple indicators, multiple causes model (MIMIC; Finch, 2005; Jöreskog & Goldberger, 1975; as cited in Y.-W. Chang et al., 2015), multiple-group categorical CFA models are arguably most appropriate for the analysis of ordinal data (for a review, see Y.-W. Chang et al., 2015; E. S. Kim & Yoon, 2011). These models allow researchers to specify a CFA model for each group and constrain certain parameters to be equal across the models based on the four MI levels. Then, likelihood ratio tests or changes in model fit indices (e.g., the CFI) help identify the MI model that best fits the data (Widaman & Reise, 1997). With this approach, MI is commonly evaluated at the full model (i.e., test) level. In other words, MGCFA most often identifies whether a set of full models, with all corresponding item

parameters constrained to be equal, fits the data well. Post-hoc analyses can then be used to pinpoint misfitting items.

MI can also be evaluated using IRT models (E. S. Kim & Yoon, 2011; Stark et al., 2006). In this framework, differential item functioning (DIF) analyses examine whether the item response functions, or the probability of answering an item in a certain way for a given latent trait value, are equivalent across groups (Holland & Wainer, 1993). Connecting the terminology, lack of DIF among the items in a test would be considered evidence of MI. Moreover, given the relationship between factor analysis and some IRT models (Kamata & Bauer, 2008; Takane & De Leeuw, 1987), DIF methods can also be considered in terms of the aforementioned MI levels in certain contexts. For example, the difficulty and discrimination parameters in the two-parameter logistic model (2PLM; Lord & Novick, 1968) are related to the intercepts and factor loadings, respectively, in a CFA model. In this case, DIF methods can essentially test for strong and weak invariance by constraining the difficulty and discrimination parameters, respectively (Widaman & Reise, 1997).

In comparison to MGCFA, DIF methods take an explicitly item-level approach. Systematically examining MI at the item level allows psychometricians to more easily review and potentially remove problematic items in the test construction process. Researchers have proposed numerous parametric and nonparametric methods for identifying differentially functioning items within a test. For example, Raju et al. (1995) proposed two indices for quantifying differences in item response functions through their Differential Functioning of Items and Tests (DFIT) framework. Within this framework, the non-compensatory DIF (NCDIF) index “reflects the average squared difference

between the focal group and reference group item-level true scores" (Raju et al., 2002, p. 523). Importantly, NCDIF looks at differential functioning for an item irrespective of the samples' responses to other items (Raju et al., 1995, 2002). Then, the compensatory DIF (CDIF) index extends the NCDIF index to account for differential functioning in the other items (Raju et al., 2002).

Other common DIF detection methods include the Mantel-Haenszel test (e.g., Holland & Thayer, 1988; Mantel & Haenszel, 1959), logistic regression (Swaminathan & Rogers, 1990), and likelihood ratio tests for IRT model comparisons (IRT-LRT; Thissen et al., 1988, 1993). For instance, the backward IRT-LRT method compares a set of models, where each set has one model with all item parameters constrained to be equal across groups, and another model with only one item's parameters freely estimated (see E. S. Kim & Yoon, 2011 for an overview). More recently, researchers have proposed regularization methods that use penalty functions to identify DIF without requiring iterative analyses across items (Bauer et al., 2020; Belzak & Bauer, 2020; Magis et al., 2011). Numerous simulation studies comparing these methods are available in the extant literature (e.g., Finch, 2005, 2016; Rogers & Swaminathan, 1993; Woods, 2011).

Researchers can also examine MI at the full test level with IRT models using methods for differential test functioning (DTF; Raju et al., 1995). DTF analyses evaluate differences in expected test score functions across groups. Relating to the DFIT framework, DTF is conceptualized as the sum of compensatory DIF (Kleinman & Teresi, 2016; Raju et al., 1995). However, there is no direct relationship between NCDIF and DTF (Oshima & Morris, 2008; Raju et al., 1995). Therefore, in many cases, evidence of differential functioning at the item level does not indicate evidence of differential

functioning at the test level (Chalmers et al., 2016, p. 115). Rather, both DIF and DTF analyses are necessary to fully understand the scope of MI within the data (Chalmers et al., 2016; Raju et al., 1995).

Despite the aforementioned difficulties with defining fairness within the test construction process, there is widespread agreement (e.g., American Psychological Association et al., 2014; Millsap, 1997; Widaman & Reise, 1997, 1997; Xi, 2010) that test bias evaluations, such as analyses to establish MI, are critical for ensuring "comparable validity" across relevant groups (Willingham, 1999; Willingham & Cole, 2013; Xi, 2010). ATA procedures are not necessarily less likely to select test forms that violate MI than manual test development approaches. For example, without explicit constraints, ATA cannot make up for potentially non-invariant item parameters that might reside within the item bank (Mitchell et al., 2021; Yarkoni, 2010). When implementing ATA, many researchers might evaluate MI after the "optimal" item combination has already been found. Instead, can researchers improve both the efficiency and psychometric strength of their test construction methods by explicitly incorporating MI analyses into ATA algorithms?

Previous Research Integrating ATA Algorithms and Test Bias

In early ATA research, test bias was infrequently considered during the automated test construction process. A handful of studies examined ATA algorithms with constraints for minimizing test-level impact (see Stocking et al., 1998, 2002). In these studies, impact was defined as differences in mean observed scores among groups of interest (in contrast to other MI research, where impact is evaluated using latent scores). However, assessing impact in this way has numerous statistical limitations, namely that

impact based on observed scores does not necessarily indicate bias at the latent score level (Luecht, 1998a).

More recently, researchers have applied metaheuristic algorithms to develop short-form psychosocial scales with evidence of MI (Jankowsky et al., 2020; Olaru et al., 2018, 2019; Olaru & Danner, 2021; Olaru & Jankowsky, 2021; Schroeders et al., 2016; Schultze & Eid, 2018). This work has primarily applied the ACO algorithm to personality or psychosocial functioning scales. For example, Olaru and colleagues (2018) used ACO to generate a short-form scale of the German NEO-PI-R that demonstrated evidence of both satisfactory model fit and MI across age groups. Additionally, Jankowsky and colleagues (2020) used a related method to develop short-form scales from the IPIP-NEO 300 that showed evidence of MI across culturally similar and dissimilar countries. Generally, results from this work have indicated that the ACO can create short-form scales with evidence of high reliability (as measured by McDonald's ω), satisfactory model fit, and weak to strong MI among different demographic groups.

Across these studies, MI has been incorporated into the ACO algorithm in multiple ways. In some methods, MI was analyzed directly within the objective function. For instance, Olaru et al. (2018), Jankowsky et al. (2020), and Olaru and Jankowsky (2021) incorporated the change in CFI when comparing models with either weak or strong invariance. Similarly, Olaru and Danner (2021) combined the CFI and root mean square error of approximation (RMSEA) model fit indices for a strong MI structural equation model. Alternative algorithms instead either evaluated MI after test assembly (Kerber et al., 2020; Martín-Fernández et al., 2021), or assumed there was MI within the item bank during the test construction process (Schultze & Eid, 2018).

Although this work indicates great promise for integrating MI with ATA, numerous limitations were identified. First, this work has largely relied on (a) empirical analyses of large-scale personality and psychosocial functioning assessments, and (b) one type of metaheuristic algorithm. A simulation study would provide important insight into the performance of ATA with MI across a broader variety of testing conditions, including with different test and algorithm types (Schultze & Eid, 2018). Second, this research has also focused solely on group-based analytic approaches, whether for test score reliability (with McDonald's ω) or test-level MI (with MGCFA models). It is possible that either an item-level or an individualized approach to test score precision would improve the psychometric properties of the resulting test, both when conducting analyses for the full sample and when differentiating among groups of interest. Therefore, questions remain regarding ATA's ability to efficiently construct unbiased psychological tests.

Research Aims

The current study builds upon the extant methods by proposing and evaluating an "Unbiased-ATA" approach. Specifically, an objective function was first developed to explicitly evaluate MI across the test's intended subgroups. The performance of this objective function was then assessed across a variety of testing scenarios, examining the extent to which incorporating MI analyses in this new way produced test scores with desirable psychometric properties. The new objective function was also compared to a subset of ATA methods that have been proposed in the IRT and SEM literatures.

The Unbiased-ATA method differs from previous integrations of ATA and MI in two specific ways. First, Unbiased-ATA exclusively uses an IRT approach for test construction. Compared to group-level SEM analyses, IRT facilitates a more

individualized evaluation of test score precision (Embretson & Reise, 2000; Schroeders et al., 2016), as well as a more common incorporation of item-level test bias analyses (e.g., examining DIF; Holland & Wainer, 1993; Thissen et al., 1993). Since IRT is often used in test construction, it is beneficial to examine its performance with ATA and MI.

The second distinction between Unbiased-ATA and previous methods is that the Unbiased-ATA objective function incorporates both item- and test-level MI analyses. In the IRT framework, these correspond to DIF and DTF analyses, respectively. It is important to examine both types of MI because DIF does not necessarily imply DTF, and vice versa (Chalmers et al., 2016; Raju et al., 1995). Moreover, DIF analyses are often an integral component of test development procedures (e.g., Dorans, 2013) so that psychometricians can flag and review potentially problematic items. It therefore remains to be seen whether incorporating item-level analyses improves upon the MI of the resulting test compared to algorithms that exclusively use test-level analyses.

Three Monte Carlo studies were designed to evaluate the performance of Unbiased-ATA. Study 1 gauged the psychometric properties of tests constructed by Unbiased-ATA across a variety of testing conditions. Study 2 then examined whether differential weighting of the criteria in the objective function affected Unbiased-ATA's performance. Finally, Study 3 explored the extent to which Unbiased-ATA improved upon previous ATA algorithms by comparing the proposed method to two alternative objective functions.

Chapter 2: The Unbiased-ATA Method

The Unbiased-ATA method is defined by an objective function, comprising key properties that test developers might want their resulting test to feature. Specifically, the objective function is a linear, equally weighted combination of different psychometric criteria. These criteria broadly represent indices of (a) item-level MI, (b) test-level MI, (c) test score precision, and (d) item fit. The Unbiased-ATA objective function can be solved with both metaheuristic algorithms and linear programming techniques. Additionally, the most time-consuming analyses (e.g., IRT model estimation) are completed prior to the algorithm's implementation, ensuring that the addition of new objective function criteria does not substantially increase the method's computational burden.

Importantly, the Unbiased-ATA method focuses on building static test forms. This procedure is applicable to scenarios wherein researchers are either (a) building a test from a larger item bank, or (b) creating a short-form scale from a previously developed long-form assessment. One might instead consider using a computerized adaptive test (CAT), which sequentially administers items to an examinee based on their responses to previous items (Weiss, 1982, 2004). Whereas examinees all answer the same item set in a static test form, examinees instead answer an individualized item set in a CAT based on their ability levels. Compared to static test forms, adaptive testing can provide increased measurement efficiency and test score precision (Weiss, 1982, 2004; Weiss & Kingsbury, 1984). However, static test forms remain popular in psychological research. Additionally, relatively less is known about ways to evaluate MI with adaptive tests (where response matrices can be sparse; although see Gierl et al., 2013; Zwick, 2009). Although not

addressed in the current research, integrating ATA algorithms and MI in a CAT context is a fruitful area for future research.

Objective Function Criteria

Item-Level MI

Researchers have proposed numerous methods for evaluating DIF within IRT test construction. A common parametric method is to compare a series of models with IRT-LRT (Thissen et al., 1988, 1993) wherein an individual item's parameters (e.g., difficulty and discrimination in the 2PLM) are sequentially constrained across groups. Specifically, the IRT-LRT for item j compares two models: (a) M_{j0}, where item j's parameters are equivalent for all groups, and (b) M_{j1}, where item j's parameters are freely estimated for all groups (E. S. Kim & Yoon, 2011, p. 217). In these models, typically all other item parameters are also constrained to be equal (Belzak & Bauer, 2020; E. S. Kim & Yoon, 2011). If M_{j1} is found to provide significantly better model-data fit than M_{j0} (as determined using large-sample hypothesis tests with a likelihood ratio test statistic), then item j is considered to demonstrate DIF. This model comparison is then repeated for all other items of interest (E. S. Kim & Yoon, 2011; S.-H. Kim & Cohen, 1998; Thissen et al., 1988, 1993). An extension of IRT-LRT is the Wald test (Lord, 1980; as cited in Teresi et al., 2021), which uses a χ^2 test statistic for the model comparisons (for a review, see Teresi et al., 2000, 2021).

IRT-LRT has been shown to work well at identifying DIF in many contexts (E. S. Kim & Yoon, 2011; Lei et al., 2006; Woods, 2011). However, even when adjusting for multiple testing, IRT-LRT can demonstrate unacceptable Type I error rates (Belzak & Bauer, 2020; Finch, 2005; Stark et al., 2006). Furthermore, researchers typically ensure

model identification by selecting a set of anchor items; these items are assumed to be invariant across groups and are used to identify the scales of the groups' latent distributions. Yet researchers typically do not know which items are truly invariant, and mistakenly choosing a differentially functioning item as an anchor can substantially bias the IRT-LRT results (Belzak & Bauer, 2020; Magis et al., 2011). Although numerous strategies to carefully determine anchor items or "purify" the item set have been proposed (e.g., Sireci & Rios, 2013; W.-C. Wang & Yeh, 2003), the risk of misidentifying anchor items in IRT-LRT remains (Kopf et al., 2015; Teresi et al., 2021).

Recently, DIF researchers have proposed penalized regression techniques to identify DIF through non-iterative methods. Regularized DIF (Belzak & Bauer, 2020; Magis et al., 2011) combines the log-likelihood of the IRT model with the least absolute shrinkage and selection operator (lasso) penalty (Tibshirani, 1996). The penalty conducts a variant of variable selection, whereby the coefficient magnitudes for non-DIF items are reduced towards zero. Regularized DIF with the 2PLM has shown desirable statistical properties in numerous conditions when compared to IRT-LRT, including lower false positive rates while maintaining strong true positive rates (Belzak & Bauer, 2020).

To illustrate the regularized DIF method from Belzak and Bauer (2020), consider the 2PLM model (Belzak & Bauer, 2020, Equation 3):

$$P\left(Y_{ij} = 1 \middle| \theta_i\right) = \frac{1}{1 + \exp\left(-c_j + a_j\theta_i\right)}, \tag{3}$$

where Y_{ij} is the response to item $j \in \{1, \ldots, n\}$ by examinee $i \in \{1, \ldots, N\}$, c_j is the intercept parameter for item j, a_j is the corresponding slope parameter, and θ_i is the latent ability value for examinee i. Note that Equation 3 uses the slope-intercept parameterization rather than the traditional IRT parameterization defined by the

discrimination (a_j) and the difficulty (b_j) parameters. These two parameterizations are mathematically related with the 2PLM, such that $c_j = -a_j \times b_j$ (Belzak & Bauer, 2020; Lord & Novick, 1968).

With two groups (denoted the reference and focal groups), Equation 3 can be re-parameterized as (Belzak & Bauer, 2020, Equation 7)

$$P\left(Y_{ij} = 1 \middle| \theta_i, z_i\right) = \frac{1}{1 + \exp\left[-\left(c_{0j} + c_{1j} z_i\right) - \left(a_{0j} + a_{1j} z_i\right)\theta_i\right]}, \quad (4)$$

where c_{0j} and a_{0j} are the "baseline" intercept and slope parameters, respectively, for the reference group, c_{1j} and a_{1j} are the differences in parameter values between the reference and focal groups, z_i is a binary variable indicating if examinee i is in the focal group, and all other terms are as defined in Equation 3. In this model, c_{1j} and a_{1j} are denoted "DIF parameters." By using this parameterization, the regularized DIF procedure can "penalize the [DIF] parameters during model estimation" such that "if the [DIF] parameters are removed, then the item is an anchor, otherwise the item expresses DIF." In other words, if an item is deemed invariant, these DIF parameters will be zero and the two groups will have equivalent intercept and slope parameters for the item (Belzak & Bauer, 2020, p. 677).

Using the re-parameterized model in Equation 4, the regularized DIF likelihood function is then (Belzak & Bauer, 2020, Equations 8-9)

$$l(\boldsymbol{\gamma})_{\text{Reg-DIF}} = \log\left(\prod_{i=1}^{N} \int p(\mathbf{Y}_i|\theta_i, z_i; \boldsymbol{\omega}) \varphi(\theta_i|z_i; \boldsymbol{\pi}) d\theta_i\right) - \tau \sum_{j=1}^{p} \left[\left|c_{1j}\right| + \left|a_{1j}\right|\right]. \quad (5)$$

Here, the first quantity is the log-likelihood function of the 2PLM in Equation 4, with item parameters $\boldsymbol{\omega}$ and latent distribution parameters $\boldsymbol{\pi}$. The second quantity is a variant

of the lasso penalty function, essentially penalizing the DIF parameters in Equation 4 (Belzak & Bauer, 2020, p. 677). Beyond the 2PLM, researchers have also applied regularized DIF to the graded response model (Belzak, 2021), the generalized partial credit model (Schauberger & Mair, 2020), moderated nonlinear factor analysis (Bauer et al., 2020), and logistic regression models (Magis et al., 2011). Regularized DIF was also recently extended to multidimensional IRT models (C. Wang et al., 2022).

In the lasso penalty, τ is a tuning parameter where larger values of τ result in "greater shrinkage of DIF parameters in $\boldsymbol{\gamma}$, with the goal being to remove these for anchor items that are free of DIF" (Belzak & Bauer, 2020, p. 677). The method for selecting τ is important for balancing Type I errors and power (commonly referred to as false and true positive rates, respectively). With a relatively small item bank, inflated Type I error rates (i.e., estimating non-zero DIF parameters for an invariant item) can be costly in that there are fewer replacement items (Belzak & Bauer, 2020). Alternatively, decreased power, and thus increased Type II errors, are also problematic because truly non-invariant items may be retained. The current study mirrors previous work (Bauer et al., 2020; Belzak & Bauer, 2020) by comparing models across a grid of plausible τ values and selecting the model with the minimized Bayesian information criterion (BIC; Schwarz, 1978). Although the BIC tends to be more conservative in identifying DIF compared to the AIC (Belzak & Bauer, 2020; Magis et al., 2011), previous simulations (e.g., Belzak & Bauer, 2020) have demonstrated that regularized DIF with the BIC produces adequate false and true positive rates for sample sizes examined in the current studies (as described further below).

Regularized DIF facilitates the identification of differentially functioning items using a dichotomous categorization process (i.e., there is evidence that an item is either differentially functioning or invariant). Such DIF detection methods can be further supplemented with a DIF effect size index to better understand the practical significance of the differential functioning. One effect size index recently proposed in the literature is the weighted area between the expected score curves (wABC; Edelen et al., 2015; Hansen et al., 2014). To understand the wABC, first define the expected score curve for item j as (Chalmers et al., 2016, Equation 3; Edelen et al., 2015)

$$S_j(\theta, \boldsymbol{\omega}_j) = \sum_{c=0}^{C-1} c \times P(y = c|\theta, \boldsymbol{\omega}_j), \tag{6}$$

where C is the number of item categories (e.g., two for a dichotomous item) and $P(y = c|\theta, \boldsymbol{\omega}_j)$ is the item response function for item j evaluated at a given θ value with an associated vector of item parameters $\boldsymbol{\omega}_j$. Thus, $S_j(\theta, \boldsymbol{\omega}_j)$ represents "what an individual's expected observed item response value would be when given a person's θ value and the item parameters" (Chalmers et al., 2016, p. 117). For a dichotomous item, the expected score curve reduces to the probability of a correct response, commonly referred to as the item response function (IRF).

Next, the wABC quantifies the difference between $S_j(\theta, \boldsymbol{\omega}_j)$ for two groups (e.g., the reference and focal group). Two differences are computed, with each weighted by the group-specific distribution (e.g., the standard normal distribution). These weighted differences are then averaged, and further weighted by the sample size split between the groups. Altogether, the wABC for item j is (Edelen et al., 2015, pp. 97–98)

$$\mathrm{wABC}_j = \left(\int_\theta \left| S_{j\mathrm{R}}(\theta, \boldsymbol{\omega}_j) - S_{j\mathrm{F}}(\theta, \boldsymbol{\omega}_j) \right| \varphi_{\mathrm{F}}(\theta)\, \mathrm{d}\theta \right) \times (N_{\mathrm{F}}/N)$$

$$+ \left(\int_\theta \left| S_{j\mathrm{R}}(\theta, \boldsymbol{\omega}_j) - S_{j\mathrm{F}}(\theta, \boldsymbol{\omega}_j) \right| \varphi_{\mathrm{R}}(\theta)\, \mathrm{d}\theta \right) \times (N_{\mathrm{R}}/N), \quad (7)$$

where R and F denote the reference and focal group, respectively, $\varphi_{\mathrm{R}}(\theta)$ and $\varphi_{\mathrm{F}}(\theta)$ are the group-specific distributions, N_{R} and N_{F} are the group-specific sample sizes, and N is the total sample size. In practice, the integrals are computed using quadrature across a range of θ values.

In the current method, regularized DIF is applied once to the n_{bank} items prior to algorithm implementation. To obtain the penalized DIF parameter estimates, regularized DIF compares models across a grid of 100 τ values and selects the model with the minimal BIC value. The regularized DIF results are then used to categorize the n_{bank} items into three groups based on increasing DIF severity: (a) anchor items with null DIF parameters, (b) "small DIF" items, and (c) "large DIF" items. Specifically, dichotomous "small DIF" items have both non-zero DIF parameters from the regularization method and wABC values less than or equal to 0.20 (Belzak & Bauer, 2020). Additionally, dichotomous "large DIF" items have non-zero DIF parameters and wABC values greater than 0.20. Supplementing regularized DIF with the wABC metric provides numerical evidence as to whether statistically significant DIF translates into practical significance (Edelen et al., 2015; Hansen et al., 2014). Here, an item is flagged as demonstrating DIF if any of its parameters are deemed to be non-invariant by regularized DIF. In other words, both uniform DIF (differences in only the intercept parameters) and non-uniform DIF (differences in either the slope or both the slope and intercept parameters) are considered.

The DIF severity categorizations for the n_{bank} items then serve as input for calculating the Unbiased-ATA objective function. Recall that an ATA algorithm compares numerous proposed tests with varying combinations of n_{test} items (where $n_{\text{test}} < n_{\text{bank}}$) to find the test with the maximum (or minimum) objective function value. Here, the objective function criterion for item-level MI is a weighted sum of the proportion of items in the proposed combination that are categorized as either anchor, "small DIF," or "large DIF" items. Let $\boldsymbol{\omega}_{1j}$ be the vector of DIF parameters for item j. In regularized DIF, this will be a null vector for items found to be invariant across the two groups. Furthermore, define w as the threshold differentiating small and large DIF based on the wABC value (i.e., $w = 0.2$ for the 2PLM). The objective function is then written as

$$\gamma_{\text{DIF}} = 0.25\left[\frac{\sum_{j=1}^{n_{\text{bank}}} \mathbf{1}(\boldsymbol{\omega}_{1j} = \mathbf{0})x_j}{n_{\text{test}}}\right] + 0.25\left[1 - \frac{\sum_{j=1}^{n_{\text{bank}}} \mathbf{1}(\boldsymbol{\omega}_{1j} \neq \mathbf{0} \cap 0 < \text{wABC}_j \leq w)x_j}{n_{\text{test}}}\right]$$
$$+ 0.5\left[1 - \frac{\sum_{j=1}^{n_{\text{bank}}} \mathbf{1}(\boldsymbol{\omega}_{1j} \neq \mathbf{0} \cap \text{wABC}_j > w)x_j}{n_{\text{test}}}\right], \quad (8)$$

where $\mathbf{1}()$ is the indicator function and $\mathbf{0}$ is a null vector with length equal to the length of $\boldsymbol{\omega}_{1j}$ (e.g., two in the 2PLM). In Equation 8, x_j is a binary variable indicating whether item j is included in the proposed test. The proportions of small and large DIF (the second and third quantities in Equation 8, respectively) are subtracted from 1.0 because the Unbiased-ATA algorithm seeks to maximize the objective function. In this case, lower proportions of DIF items in the test are associated with higher objective function values. Moreover, the three proportions are weighted such that greater weight is placed on minimizing the number of "large DIF" items (based on previous research suggesting

that small DIF items are not practically problematic in test construction applications; Edelen et al., 2015; Hansen et al., 2014). Importantly, both the *w* threshold and the proportion weights in Equation 8 are easily modified based on a test developer's specific test goals.

An alternative method would be to simply remove all DIF items from the bank prior to implementing the ATA algorithm. DIF items are retained in the item bank for two related reasons. First, items are often not immediately removed if they are found to have DIF. Rather, items flagged as demonstrating significant DIF are typically examined afterward using "qualitative interpretations" (Sireci & Rios, 2013, p. 172). In many cases, these items might need to be rewritten or restructured rather than fully removed. Unbiased-ATA still arguably improves the efficiency of this process: rather than review all bank-level DIF items, researchers might only need to review the subset of DIF items that are included in the algorithm's proposed "best" test. Second, Unbiased-ATA seeks an item set that maximizes a *combination* of psychometric criteria (described further below), not only item-level MI. DIF items might be retained in a scale due to other desirable properties (e.g., high information), with an accompanying note for future researchers to be careful if they decide to draw group-level comparisons with the item (e.g., Varni et al., 2014). This perspective is not applicable in all testing settings (e.g., in high-stakes testing or job hiring processes), so Unbiased-ATA might be better suited for clinical and research purposes.

It also merits comment that the factors engendering DIF can differ depending on the test, item type, and trait of interest. For example, an item might have different meaning to examinees with different characteristics (Varni et al., 2014), or non-invariant

item parameters can result from the disproportionate impact of other latent factors (Sireci & Rios, 2013). By integrating item-level MI analyses directly into the objective function, Unbiased-ATA can identify, and attempt to filter out, items with small to problematic DIF from the proposed test. However, this method cannot differentiate among the various causes of DIF. This limitation is not specific to Unbiased-ATA but is shared among ATA problems more broadly. Specifically, ATA is not intended to be the final step of the test development process, but rather an efficient mechanism for identifying a plausible set of items that can be further reviewed by test developers. Employing group-level analyses also allows test developers to determine whether certain items should be flagged or removed when making certain group comparisons (e.g., Varni et al., 2014).

Test-Level MI

Beyond understanding differences in item-level functioning across groups, it is necessary to also explore the extent to which group's expected test score functions differ (Chalmers et al., 2016; Raju et al., 1995). For instance, sets of items might be biased toward groups in different ways, such that evidence of DIF does not necessarily translate to evidence of differential functioning at the full test level. Additionally, seemingly small bias at the item level can combine to demonstrate substantial bias at the test level (Chalmers et al., 2016, p. 118). Therefore, the objective function expands beyond item-level MI analyses to additionally incorporate a measure of DTF.

Unbiased-ATA uses an effect size proposed by Chalmers and colleagues (2016) to quantify the extent to which expected test score functions "have a large degree of overall separation on average" (p. 119). This effect size improves upon Raju et al.'s (1995) DTF index in numerous ways. For instance, a limitation of Raju et al.'s (1995)

index is its dependence on a sample's θ estimates (Chalmers et al., 2016, pp. 122–123). The effect size uses the expected test score function, defined as (Chalmers et al., 2016, Equation 4)

$$T(\theta, \boldsymbol{\omega}) = \sum_{j=1}^{n} S_j(\theta, \boldsymbol{\omega}), \tag{9}$$

where $S_j(\theta, \boldsymbol{\omega})$ is the expected item score curve (i.e., the IRF for a dichotomous item) from Equation 6 and n is the number of test items. In other words, $T(\theta, \boldsymbol{\omega})$ represents an individual's expected observed test score across all n items.

Drawing from Equations 6 and 9, the unsigned DTF effect size (uDTF) is then written for ATA with two groups as (Chalmers et al., 2016, Equation 7)

$$\text{uDTF} = \int \left| \sum_{j=1}^{n_{\text{bank}}} S_j(\theta, \boldsymbol{\omega}_{\text{R}})\, x_j - \sum_{j=1}^{n_{\text{bank}}} S_j(\theta, \boldsymbol{\omega}_{\text{F}})\, x_j \right| \varphi(\theta) d\theta\,, \tag{10}$$

where R and F refer to the reference and focal group, respectively, and $\varphi(\theta)$ is "a weighting function with the property that $\int \varphi(\theta) d\theta = 1$" (based on the latent trait density; Chalmers et al., 2016, p. 120). To make the estimation more feasible, the integral in Equation 10 is replaced with a summand across a large number of quadrature points (Chalmers, 2012; Chalmers et al., 2016). uDTF thus "captures the average area between the two [expected test score] curves, indicating absolute deviations in item properties that have been aggregated over the whole test" (Chalmers et al., 2016, p. 120). It merits comment that uDTF differs from the wABC in that the former examines the area between the expected *test* score functions, whereas the latter examines the area between *item* score functions.

The uDTF effect size in Equation 10 requires sets of item parameters for both the reference and focal groups. In Unbiased-ATA, these item parameters are based on a multiple-group IRT model, fit using all n_{bank} items and estimated using an Expectation-Maximization (EM) algorithm with 61 quadrature points (as implemented in the R package *mirt*; Chalmers, 2012). To identify the model, the invariant items selected by regularized DIF are set as anchors. Using the estimated item parameters, the item score function values across 100 quadrature points are easily computed for all n_{bank} items. Then, the objective function criterion for test-level MI comprises the uDTF, transformed to a proportion to convert the index to a $[0,1]$ scale:

$$\gamma_{\text{DTF}} = 1 - \frac{\text{uDTF}}{\text{TS}}, \tag{11}$$

where TS represents the "highest possible test score" for the proposed test (Chalmers et al., 2016, p. 120). For example, $\text{TS} = 20$ for a test with 20 dichotomous (0/1) items. Again, the proportion uDTF is subtracted from 1.0 so that lower values of uDTF are associated with higher objective function criterion values.

Test Score Precision

Test score precision is arguably the psychometric property most often incorporated into psychological and educational ATA. Within an IRT framework, test score precision is typically evaluated using the test information function (TIF), and ATA objective functions often seek to maximize the TIF across a range of desired latent trait values (e.g., Boekkooi-Timminga, 1990; P.-H. Chen et al., 2012; Finkelman et al., 2010; Huitzing et al., 2005; Levis et al., 2016; Martín-Fernández et al., 2021). Assuming local independence of items (i.e., an examinees' responses to different items are statistically independent after controlling for θ), the test information is calculated as the sum of the

item information values for a given value of θ. Specifically, the TIF is (Embretson & Reise, 2000, p. 184)

$$\mathrm{TIF}(\theta) = \sum_{j=1}^{n} I_j(\theta)\,, \tag{12}$$

where θ is the latent trait value, n is the number of test items, and I_j is the item information value for item j. The TIF can be decomposed as the sum of the expected or the observed item information values (Magis, 2015). In the current study, Unbiased-ATA uses the expected (Fisher) information (DeGroot & Schervish, 2012, p. 515), although note that observed and expected information values are equivalent with the 2PLM (Bradlow, 1996; Magis, 2015).

Maximizing Equation 12 across a range of intended θ values can produce TIFs with unfamiliar or unwanted shapes. For example, test developers might seek a TIF that is relatively high and flat across the θ continuum. Yet an item combination that maximizes Equation 12 might instead have high, multimodal peaks due to a handful of items with relatively higher information. Therefore, the objective function criterion for test score precision in Unbiased-ATA instead maximizes the deviation of the estimated TIF from a target TIF (Ali & van Rijn, 2016; Armstrong et al., 1998; P.-H. Chen, 2016; Luecht, 1998b; van der Linden & Adema, 1998). This criterion is

$$\Delta_{\mathrm{TIF}} = \sum_{k=1}^{K} \left| \sum_{j=1}^{n_{\mathrm{bank}}} I_j(\theta_k) x_j - \mathrm{TIF_T}(\theta_k) \right|, \tag{13}$$

where K is the number of θ values and $\mathrm{TIF_T}$ is the target TIF (see Equation 3 in Armstrong et al., 1998). Estimating the TIF deviation at a sequence of discretized θ

values, rather than integrating across θ, makes the computation more feasible. Note that in Equation 13, TIF deviations are equally weighted across the range of θ values.

Target TIFs have been frequently used in previous ATA research, such as when creating parallel test forms (e.g., van der Linden & Adema, 1998). In the current study, the target TIF is based on the TIF for the full item bank. This choice implicitly assumes that the TIF for the full item bank is the desired information function shape for the proposed test. This assumption is reasonable in many test development situations, including scale short-form development. However, because the TIF depends upon the number of items, the bank-level TIF will naturally be larger than that for the proposed test. To better match the two TIFs, the target TIF values are multiplied by the ratio of the number of items for the proposed test to the number of items in the item bank (Ali & van Rijn, 2016, p. 170).

To compute the TIF deviation in Equation 13, the vector of item parameters from the full item bank can be used. However, if DIF is identified within the item bank, using one set of item parameters for the full sample will ignore the presence of measurement non-invariance. A plausible alternative is to instead use the estimated item parameters from the multiple-group IRT model, and compute one target TIF per group. The Unbiased-ATA objective function criterion then sums the TIF deviations across the two groups. The TIF deviations are summed rather than averaged to avoid a case wherein one group's large TIF deviation is masked by the other group's small deviation.

In summary, the test score precision criterion for the Unbiased-ATA objective function is expanded as

$$\gamma_{\text{Precision}} = \left[1 - g\left(\sum_{k=1}^{K}\left|\left\{\sum_{j=1}^{n_{\text{bank}}} I_{j\text{R}}(\theta_k)x_j\right\} - \text{TIF}_{\text{TR}}(\theta_k)\right|\right)\right] +$$

$$\left[1 - g\left(\sum_{k=1}^{K}\left|\left\{\sum_{j=1}^{n_{\text{bank}}} I_{j\text{F}}(\theta_k)x_j\right\} - \text{TIF}_{\text{TF}}(\theta_k)\right|\right)\right], \quad (14)$$

where TIF_{T} is again the target TIF, and R and F indicate the reference and focal groups, respectively. In Equation 14, $g()$ is a function that uses min-max normalization to convert the TIF deviation (Δ_{TIF}) to a $[0,1]$ scale. Specifically, this function is

$$g(\Delta_{\text{TIF}}) = \frac{\Delta_{\text{TIF}} - \min(\Delta_{\text{TIF}})}{\max(\Delta_{\text{TIF}}) - \min(\Delta_{\text{TIF}})}. \quad (15)$$

Here, the minimum TIF deviation across the K θ values is zero. The maximum TIF deviation is then

$$\max(\Delta_{\text{TIF}}) = \sum_{k=1}^{K}\left|0 - \text{TIF}_{\text{T}}(\theta_k)\right|. \quad (16)$$

This maximum deviation is akin to having a proposed test with null information at each θ_k value. Although it is highly unlikely that the selected test will have a null TIF, Equation 16 places an upper bound on the plausible Δ_{TIF} values. Note also that the maximum Δ_{TIF} might differ between the reference and focal groups (because the target TIFs may differ). Again, each transformed TIF deviation is subtracted from 1.0 so that the objective function prioritizes smaller differences in the TIF values.

Item Fit

The final psychometric criterion evaluated in Unbiased-ATA concerns the extent to which the specified model approximately fits the data in the proposed test. In IRT, model-data fit is typically evaluated at the item level by using one or more item fit

statistics to determine similarities between the estimated and observed item response functions (IRFs). As Ames and Penfield (2015) describe, "A key assumption of IRT is that each IRF of the scored data accurately reflects the link between an individual's latent ability and item responses." One or more misfitting items may lead to "biased ability and item parameter estimates." Calculating item fit statistics is therefore a critical component of the test construction process in IRT (Ames & Penfield, 2015, p. 39).

Numerous item fit statistics have been proposed in the IRT literature, largely based on either χ^2 or likelihood ratio tests (Ames & Penfield, 2015). One statistic that has demonstrated particularly desirable statistical properties for unidimensional IRT models (in terms of true and false positive rates) is Orlando and Thissen's (2000) $S-\chi^2$ statistic (Davis, 2009; Kang & Chen, 2008, 2011; Orlando & Thissen, 2000, 2003; Stone & Zhang, 2003). The $S-\chi^2$ statistic uses observed rather than latent scores to classify examinees into G groups of similar scores (Kang & Chen, 2008). For binary item j, the $S-\chi^2$ statistic is computed as (Orlando & Thissen, 2000, Equation 13)

$$S-\chi_j^2=\sum_{g=1}^{G-1} N_g \frac{\left(O_{jg}-E_{jg}\right)^2}{E_{jg}\left(1-E_{jg}\right)}, \tag{17}$$

where N_g is the number of examinees in group g, O_{jg} is the observed proportion of correct responses to item j for examinees in group g, and E_{jg} is the corresponding expected proportion. E_{jg} is computed using a "recursive algorithm that builds the joint likelihood for each score group, one item at a time" (Orlando & Thissen, 2000, p. 53). Specifically, this quantity is computed for a binary item as (Kang & Chen, 2008, Equation 4; Orlando & Thissen, 2000, Equation 12)

$$E_{jg} = \frac{\int P\left(Y_j = 1 \middle| \theta\right) S_{g-1}^{*j}(\theta)\varphi(\theta)d\theta}{\int S_g\,(\theta)\varphi(\theta)d\theta}, \tag{18}$$

where $P\left(Y_j = 1 \middle| \theta\right)$ is the IRF for item j, $S_g(\theta)$ is "the number correct score posterior distribution for score [bin g]" and $S_{g-1}^{*j}(\theta)$ is "the number correct score posterior distribution for score [bin g]…without item j" (Orlando & Thissen, 2000, pp. 53–54). In practice, Equation 18 is computed using quadrature. More details about the construction of the $S - \chi^2$ statistic can be found in Orlando and Thissen (2000, 2003). Additionally, Kang and Chen (2008, 2011) provided evidence of this item fit statistic's strong performance in unidimensional, polytomous IRT models (although see Su et al. [2021] for insight into the $S - \chi^2$ statistic's performance in multidimensional polytomous IRT models).

Using the estimated item parameters from the multiple-group IRT model, group-level $S - \chi^2$ statistics indices (for either the reference or focal group) are computed for each item in the bank. Based on the null hypothesis that the IRF model provides sufficient fit to the data, a p-value for binary item j (p_j) can be computed by comparing the $S - \chi^2$ statistic to a χ^2 distribution with $df = n - 1 - p$, where n is the number of items and p is the number of estimated item parameters (e.g., two in the 2PLM; Ames & Penfield, 2015). The Unbiased-ATA objective function criterion for item fit then calculates the proportion of well-fitting items in both the reference and focal group (where well-fitting is operationalized as a p-value greater than α):

$$\gamma_{\text{IF}} = \left[\frac{\sum_{j=1}^{n_{\text{bank}}} \mathbf{1}\left(p_{j\text{R}} > \alpha\right) x_j}{n_{\text{test}}}\right] + \left[\frac{\sum_{j=1}^{n_{\text{bank}}} \mathbf{1}\left(p_{j\text{F}} > \alpha\right) x_j}{n_{\text{test}}}\right]. \tag{19}$$

Here, a higher proportion of well-fitting items is preferred. Again, the item fit indices and corresponding *p*-values are computed once for the full item bank, avoiding unnecessary computational cost that arises from re-estimating the statistics within each algorithm iteration.

The choice of α depends upon a test developer's goals, with higher levels of α resulting in a higher false positive rate (where a "positive" here refers to identifying an item as misfitting). In certain cases, a higher false positive rate is preferred, allowing for a few more false positives to not miss as many true positives. In the current study, α was set to 0.10, as preliminary studies showed that this α level provided a better balance of false and true positive rates. Relatedly, these preliminary studies also showed that *p*-value adjustments for multiple testing were excessively conservative.

Method Summary

Overall, the Unbiased-ATA objective function is composed of four criteria that broadly represent important aspects of test score accuracy and precision from an IRT framework. The following analytic steps are required before the ATA algorithm is initiated. First, regularized DIF (see Equations 3 – 5) and the wABC (see Equation 6 – 7) are applied to the data to identify both DIF and non-DIF items (with the former category further differentiated between "small" and "large" DIF items). Then, using the non-DIF items as anchors, a multiple-group IRT model is fit to the full data set. The estimated item parameters from this model are next used to compute group-level (a) item score functions, (b) item information values, and (c) item fit indices. Both the item score function and information values are computed across a range of discretized θ values to facilitate the quadrature computations in the objective function.

The bank-level item indices serve as input for the Unbiased-ATA objective function, i.e., $f(\text{UATA})$. The objective function is then an equally weighted combination of Equations 8, 11, 14, and 19,

$$f(\text{UATA}) = \gamma_{\text{DIF}} + \gamma_{\text{DTF}} + \gamma_{\text{Precision}} + \gamma_{\text{IF}} \tag{20}$$

subject to

$$\sum_{j=1}^{n_{\text{bank}}} x_j = n_{\text{test}}, \tag{21}$$

where the latter is an algorithm constraint limiting the test length to a single possible value (Green et al., 1988; Jankowsky et al., 2020; Schroeders et al., 2016). ATA problems often incorporate a predetermined test length (e.g., Jankowsky et al., 2020; Raborn et al., 2020; Schroeders et al., 2016; van der Linden, 1998), which has the additional benefit of considerably reducing the algorithm search space. For example, finding an optimal combination of n_{test} items has lower computational burden than finding an optimal combination among tests with $n_{test} - 2$, $n_{test} - 1$, n_{test}, $n_{test} + 1$, and $n_{test} + 2$ items. Additional constraints, such as the number of items per content area or subscale, are not explicitly implemented here (e.g., Stocking et al., 1998; van der Linden, 1998). However, the Unbiased-ATA method can easily be extended to incorporate such constraints.

Chapter 3: Performance Evaluation Across Testing Scenarios

Study 1 evaluated the performance of the proposed Unbiased-ATA method across a variety of test construction scenarios. Using a large simulation design, this study aimed to identify the model and data conditions wherein Unbiased-ATA produced a test with strong psychometric properties.

Simulation Design

Item Bank Generation

Study 1 used simulated tests with dichotomous items (i.e., having two possible answers, such as Yes/No or True/False) measuring a single latent trait (θ). This test type was selected to mirror a clinical symptomatology survey, with binary items indicating whether the respondent is experiencing a symptom. Although only dichotomous items were examined in the current study, the Unbiased-ATA method was created to apply to polytomous items as well (e.g., Likert-scale items analyzed with the graded response model).

The item bank[3] comprised $n_{\text{bank}} = 60$ items. The number of items was chosen to reflect scale lengths in common psychosocial measures, such as the Taylor Manifest Anxiety Scale (Taylor, 1953) and the IPIP-NEO-300 (L. R. Goldberg, 1999; L. R. Goldberg et al., 2006). First, dichotomous items were generated from a 2PLM with scaling constant $\mathrm{D} = 1$, using the traditional IRT parameterization with discrimination (a) and difficulty (b) parameters. Mirroring previous IRT simulation studies, a and b

[3] The term "item bank" is often used in the literature to represent a set of items much larger than 60 (e.g., hundreds of items, as those used for computerized adaptive testing). However, the term is used here to refer to a group of items from which test developers are selecting a test with a length less than the bank size.

parameters for the dichotomous items were each random realizations of specified probability distributions. Specifically, $a \sim \mathrm{N}(1.5, 0.15^2)$ and $b \sim \mathrm{U}[0,2]$ were used to produce an item set with moderate to high discriminations and an item bank with higher information around $\theta = 1$. Similar test or bank information functions are commonly found when a 2PLM is fit to clinical questionnaire data (e.g., Attell et al., 2020; Pattanaik et al., 2020; Svicher et al., 2019).

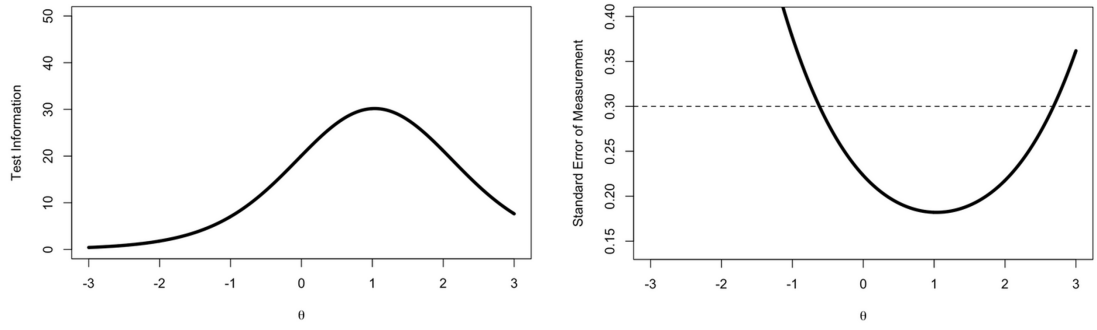
In Study 1, one set of item parameters was generated for the item bank (i.e., the "bank-level parameters") according to the specified probability distributions. Within each simulation repetition, the parameters for item j were then random realizations of uniform distributions centered at that item's bank-level parameter values (e.g., a_j and b_j). Using the process specified by Belzak (2020) and Belzak and Bauer (2020), the discrimination parameters in each repetition were random realizations of $\mathrm{U}[a_j - 0.15, a_j + 0.15]$. A lower bound of zero was added to this data generation process to ensure realistic sampled discrimination values. For dichotomous items, the difficulty parameters were random realizations of $\mathrm{U}[b_j - 0.30, b_j + 0.30]$. The item parameters for a given replication were generated for the full sample, after which group-level parameters were modified to introduce DIF. The amount that the parameters were modified produced a particular wABC value (Edelen et al., 2015) for the item, representing systematically increasing DIF severity levels. Notably, these wABC and parameter modification values were based on the item's bank-level parameters (rather than re-estimated in each replication). This method of item parameter generation therefore kept the DIF magnitudes constant across simulation repetitions while introducing cross-repetition variability into the full-sample item parameters (Belzak, 2020; Belzak & Bauer, 2020).

Table 1 presents summary statistics for the bank-level item parameters. As a point of reference, transforming the discriminations to factor loadings (Kamata & Bauer, 2008; E. S. Kim & Yoon, 2011; Takane & De Leeuw, 1987) produced a range of standardized loadings from $0.57 - 0.73$. Figure 1 displays the test information and standard error of measurement (SEM) functions for the dichotomous item bank, where the SEM is computed as the inverse square root of the TIF. Notice in this figure that the TIF increased around $\theta = 1$, similar to TIFs found in empirical analyses of clinical questionnaires (e.g., Attell et al., 2020; Svicher et al., 2019). Additionally, for $-0.5 \leq \theta \leq 2.5$ the SEMs were below 0.30, a threshold that is commonly used as a stopping criterion in adaptive testing (e.g., Dodd et al., 1993; as cited in Feuerstahler, 2018).

Table 1. Generating Distributions and Summary Statistics for Item Bank Parameters

	Discrimination (a)	Difficulty (b)
Distribution	N(1.25, 0.15^2)	U[0,2]
Mean (SD)	1.52 (0.13)	1.03 (0.52)
Median	1.52	1.04
Range	1.17 – 1.80	0.07 – 1.98

Figure 1. Item Bank Test Information and Standard Error of Measurement Functions



Item responses corresponding to the item parameters were generated using the `simIrt` function (from the *catIrt* package; Nydick, 2014) in R statistical software (R

Core Team, 2021). This function requires a set of item parameters, θ values, and a data-generating model. Here, the data-generating model was the 2PLM, meaning that the models were correctly specified. In addition to the item responses for the two item banks, responses to an external criterion measure were also generated. This 30-item criterion measure was used to evaluate Unbiased-ATA's performance in terms of external validity. Items on the criterion measure reflected a single factor that was designed to positively correlate ($\rho = 0.50$) with the latent trait represented in the item bank. The item parameters for the criterion measure were generated from the same distributions used to create the 60-item bank of interest.

To simulate mean impact on the latent trait underlying the test of interest, the data-generating distributions for the reference and focal groups were $\theta_{\mathrm{R}} \sim \mathrm{N}(0,1)$ and $\theta_{\mathrm{F}} \sim \mathrm{N}(0.5,1)$, respectively (Belzak & Bauer, 2020; E. S. Kim & Yoon, 2011). The data-generating distribution for the external criterion measure was then $\mathrm{N}(0,1)$ for examinees in both the reference and focal groups. Examinees' true θ values, denoting their ability levels for the trait of interest (thought to engender the item bank responses) as well as for the criterion measure, were randomly generated in each simulation replication. Specifically, the two sets of θ values were drawn from a bivariate normal distribution with a correlation of 0.5.

ATA Algorithms

Study 1 compared three ATA algorithms: (a) 0-1 LP (for an overview, see Luo, 2020; van der Linden, 1998, 2005), (b) ant colony optimization (ACO; Colorni et al., 1991; Dorigo & Stützle, 2004; Leite et al., 2008), and (c) Tabu search (Drezner et al., 1999; G. A. Marcoulides & Drezner, 2004; K. M. Marcoulides & Falk, 2018). The

second two algorithms are considered metaheuristic algorithms, meaning that they search across a dynamic neighborhood of plausible models to find a best-fitting (but not necessarily the global optimal) solution. Compared to the metaheuristic algorithms, 0-1 LP is a more comprehensive optimization method that searches across the entire solution space and is guaranteed to find the optimal solution, if one exists (P.-H. Chen, 2017; Luo, 2020; van der Linden, 1998).

Numerous researchers (e.g., Adema et al., 1991; Boekkooi-Timminga, 1990; Luo, 2020; Martín-Fernández et al., 2021; van der Linden, 1998; van der Linden & Adema, 1998) have applied variants of LP optimization to the ATA problem in IRT. The 0-1 LP method can be written mathematically for the ATA problem as

$$\max \mathbf{d}'\mathbf{x} \tag{22}$$

subject to

$$\mathbf{E}\mathbf{x} \leq \mathbf{f}. \tag{23}$$

Here, $\mathbf{x}$ is an $n \times 1$ vector of decision variables ($x_j \in \{0,1\}$) indicating if an item is included in the proposed test. The $n \times 1$ vector $\mathbf{d}$ combines with $\mathbf{x}$ to define the objective function that is maximized (or minimized). Therefore, the objective function is a weighted, linear combination of the decision variables (where the weights are determined by the test developers). Moreover, a set of m constraints (e.g., proposed test length, number of items for each content area) are defined by the $m \times n$ matrix $\mathbf{E}$ and $m \times 1$ vector $\mathbf{f}$ (Diao & van der Linden, 2011, p. 399; Luo, 2020).

Numerous commercial and open-source software are available to solve Equations 22 and 23. These solvers typically use recursive algorithms, like branch-and-bound, to identify the solution with the maximum (or minimum) objective function value within the

specified constraints (Luo, 2020). The present study used the R package *CVXR* (Fu et al., 2020) to formulate the optimization problem. *CVXR* was used in tandem with the R package *Rglpk* (Theussl et al., 2019), which applies the GLPK solver (Makhorin, 2012) to conduct the optimization. The time limit for each 0-1 LP problem was set at five minutes. A time limit was imposed to appropriately manage the simulation design in situations where LP was unable to definitively select an optimal item combination. Relatedly, cases where the time limit was reached provide helpful knowledge for practical applications of 0-1 LP and ATA where time might be an important factor.

ACO is a popular metaheuristic algorithm that mimics the behavior of a group of ants in search of food. As ants discover a food source, they leave a trail of pheromones to alert other ants. The pheromone levels accumulate as more and more ants follow the same path (Leite et al., 2008; Olaru et al., 2019). In the ACO algorithm, sampling weights represent these pheromone levels, which "determine the selection probability of the corresponding items" (Olaru et al., 2019, p. 406). As the algorithm iterates, items associated with test forms possessing better properties (as defined by the objective function) are more likely to be selected (Leite et al., 2008; Olaru et al., 2019).

At the algorithm start, all items begin with equivalent sampling weights. A subset of test forms is randomly selected, and the corresponding objective function values are calculated. The sampling weights for items then "increase depending on the quality of the solution," wherein larger weight is given to items on tests with higher objective function values (Olaru et al., 2019, p. 406). The algorithm might also implement a process called evaporation, where weights decrease by a certain percentage before each sampling weight increase. This process "reduces the influence of the solutions obtained at earlier stages of

the search, when poor-quality solutions are more likely to be selected" (Leite et al., 2008, p. 422). The algorithm then iterates until a stopping criterion is satisfied (Leite et al., 2008; Raborn et al., 2020). Based on previous literature, ACO was implemented with an evaporation value of 0.90, 20 ants (representing the number of test forms; Raborn et al., 2020), and a maximum of 60 iterations (Jankowsky et al., 2020).

Tabu search is another metaheuristic algorithm that has shown promise for ATA applications (e.g., Raborn et al., 2020). The Tabu search algorithm is characterized by a list of test forms that were previously evaluated and found to be inferior to the current optimal solution. This Tabu list "allows the algorithm to effectively search the model parameter space in areas that might be away from a currently best-fitted model" (K. M. Marcoulides & Falk, 2018, pp. 488–489). In addition to the Tabu list, the algorithm uses a flexible neighborhood of potential test forms to move across the search space. For instance, K. M. Marcoulides (2020) used a neighborhood "in which all feasible subsets of a current solution (beginning with the starting solution) are enumerated by including one additional variable into the model, one less variable, or replacing variables in the model by one not in the model" (p. 3).

The Tabu search algorithm begins by defining a starting solution and corresponding neighborhood. The test forms are compared on their objective function value, such that the form with the maximum value is set as the current solution and all others are put on the Tabu list. Another neighborhood is then defined using the new solution. If more than a predetermined number of test forms are added to the Tabu list, then solutions are sequentially removed. Moreover, if a new "best" solution is found, the search restarts with a new Tabu list (G. A. Marcoulides & Drezner, 2004; K. M.

Marcoulides, 2018). Like with the ACO, the algorithm iterates until a stopping criterion is met. The stopping criterion might be defined as a fixed number of iterations or the number of iterations without a change in the test form solution (K. M. Marcoulides, 2020, p. 3). Using algorithm controls based on previous ATA literature, the algorithm stopped after a maximum of 50 iterations (Raborn et al., 2020) and the Tabu list size was set to 10 (Mills et al., 2005).

Without an optimality guarantee, metaheuristic algorithms like ACO and Tabu search are at risk of identifying locally optimal solutions (Jankowsky et al., 2020; Leite et al., 2008; Olaru et al., 2018, 2019; Olaru & Jankowsky, 2021). In other words, the "best" test that the algorithm selects in one run may differ from the "best" test in another run based on the starting item configuration. A similar phenomenon occurs in the rotation of factor analysis solutions (e.g., Hattori et al., 2017; Nguyen & Waller, 2022; Rozeboom, 1992). To account for the possibility of locally optimal solutions, the ACO and Tabu search algorithms were each repeated five times from random starting item configurations (Jankowsky et al., 2020). For each algorithm type, the test with the maximum objective function value across the five repetitions was chosen as the "best" test and used for subsequent analysis.

Both the optimality guarantee and recent advances in computing power suggest that LP optimization methods should uniformly outperform metaheuristic algorithms. Indeed, some researchers (e.g., van der Linden & Li, 2016) have argued that LP solvers should largely replace heuristic and metaheuristic algorithms in ATA. However, metaheuristic algorithms were considered in the current research for three reasons. First, LP "cannot guarantee the obtainment of an optimal or near optimal solution," or to locate

a solution faster than other optimization methods. Rather, the feasibility and computational speed of LP is problem-specific (e.g., relating to the complexity of the objective function and the number of constraints; P.-H. Chen, 2017, p. 228). For a new optimization problem like Unbiased-ATA, it is beneficial to compare 0-1 LP and metaheuristic algorithms to evaluate which method provides better performance in the given context. Second, incorporating two common metaheuristic algorithms in the study design also provides important insight into their performance in an IRT, rather than an SEM, context (see Raborn et al., 2020). Finally, formulating a mathematical problem for 0-1 LP can be difficult for individuals who are unfamiliar with optimization methods. Therefore, metaheuristic algorithms might have practical advantages for test developers in many contexts.

Simulation Design Factors

Study 1 was designed to identify the conditions where Unbiased-ATA performs relatively well when comparing two groups. In total, seven design factors were manipulated to reflect realistic testing scenarios: (a) type of DIF in the item bank, (b) percentage of DIF items in the bank, (c) magnitude of DIF, (d) direction of DIF, (e) total sample size, (f) proportion of total sample in each group, and (g) item parameter generation procedure. The three ATA algorithms—0-1 LP, ACO, and Tabu search—were compared within each combination of these seven design factors.

First, Study 1 examined the effects of four types of DIF in the item bank. In the "no DIF" condition, both discrimination and difficulty parameters were equivalent across the two groups. In the "uniform DIF" condition, only the difficulty parameters differed between the groups. Then, both the discrimination and difficulty parameters differed

between the groups in the "non-uniform DIF" condition. The final level was a "mixture DIF type" condition, wherein half of the non-invariant items demonstrated uniform DIF and the other half demonstrated non-uniform DIF.

The second design factor varied the percentage of non-invariant items in the item bank. In the "small percentage" condition, a sixth of the items in the bank (10 out of 60) were selected for non-invariant item parameters across groups. Similarly, a third of the items (20 out of 60) were non-invariant in the "moderate percentage" condition. In the "large percentage" condition, one-half of the items (30 out of 60) were non-invariant. These percentages align with values used in previous simulation studies (Belzak, 2020; Belzak & Bauer, 2020) as well as values seen in practice (e.g., C. D. Huang et al., 1997; Sheppard et al., 2006). Although the item parameter values were re-generated for each simulation repetition, the non-invariant items remained constant throughout the simulation. For example, in the "small percentage" condition, the parameters for the same items were modified to introduce DIF regardless of the other simulation specifications (Belzak & Bauer, 2020; E. S. Kim & Yoon, 2011). Additionally, items chosen for a given percentage condition were a subset of the items chosen for the next largest level (e.g., items in the "small percentage" condition were a subset of the items for both the "moderate" and "large percentage" conditions; Bauer et al., 2020; Belzak & Bauer, 2020).

DIF magnitude was varied across three levels based on wABC thresholds posited by Edelen and colleagues (2015) and others (e.g., Belzak & Bauer, 2020; Hansen et al., 2014; Stucky et al., 2014). By using the wABC corresponding to the bank-level item parameter values, DIF magnitudes could differ among items (Belzak & Bauer, 2020). In

the “small magnitude” condition, item parameters were modified such that the wABC was approximately 0.1 (Belzak, 2020; Belzak & Bauer, 2020). For the “large magnitude” condition, item parameters were modified to obtain wABC values of approximately 0.20. These latter wABC values are thought to represent potentially “problematic DIF” in practical applications (Edelen et al., 2015; Hansen et al., 2014). The final level for the DIF magnitude design factor, “mixture magnitude,” reflected an equal combination of small and large DIF items.

The fourth design factor in the Study 1 simulation concerned the group for which DIF modifications were made (Suh & Cho, 2014). In the “Focal Group” condition, all DIF modifications were made to the focal group. In the “Both Groups” condition, half of the DIF modifications were made to the reference group, and the other half to the focal group. This latter condition mirrors a scenario wherein DIF effects might “cancel out” and then not exhibit differential functioning at the test level (Chalmers et al., 2016). Importantly, the design factors related to the percentage, magnitude, and direction of DIF were only applied in the “uniform DIF”, “non-uniform DIF” and “mixture DIF type” conditions.

The remaining design factors extended beyond the type and amount of DIF in the item banks. Specifically, three sample sizes were examined, $N_{\text{Total}} = \{500, 1000, 5000\}$. These sample sizes reflected the total number of examinees, ranging from arguably too small to sufficient for accurate item parameter estimation with the 2PLM (Drasgow, 1989; E. S. Kim & Yoon, 2011). The number of examinees per group was either balanced, such that $N_R = N_F = N_{\text{Total}} \times 0.5$ (e.g., Belzak & Bauer, 2020; E. S. Kim & Yoon, 2011), or unbalanced, such that $N_R = N_{\text{Total}} \times 0.7$ and $N_F = N_{\text{Total}} \times 0.3$.

The final design factor compared Unbiased-ATA's performance when item response data were generated from a set of estimated rather than true item parameter values. In the "true parameters" condition, item responses were generated from the bank-level item parameters (using the procedure described in the "Test Types and Item Bank Generation" subsection). Here, DIF was a product of varying group-specific parameter values at the population level. In the "estimated parameters" condition, examinees' item responses were generated from a set of estimated bank-level item parameters. This condition added an intermediary step wherein a set of item responses generated from the bank-level item parameters were the input for marginal maximum likelihood (MML) estimation with an expectation-maximization (EM) algorithm (Bock & Aitkin, 1981). Using the *mirt* package in R (Chalmers, 2012), MML estimation was implemented with the BFGS optimizer, 61 quadrature points, and a model convergence threshold of 0.0001. These estimated item parameters were calibrated using a sample of 1,000 simulees with θ values drawn from a standard normal distribution. Item parameters for each simulation replication were drawn from a narrow, uniform distribution centered at the estimated bank-level parameter values, and then used to generate the item responses for the simulation replication. In this condition, DIF was both incorporated within the bank-level item parameters and might also have been a product of item parameter estimation error.

For conditions with non-invariant items in the item bank (i.e., the "non-uniform DIF", "uniform DIF", and "mixture DIF type" conditions), Study 1 encompassed a total of 3 DIF types × 3 DIF percentages × 3 DIF magnitudes × 2 DIF directions × 3 sample sizes × 2 sample size proportions × 2 parameter generation procedures = 648 simulation conditions. In addition, there were 3 sample sizes × 2 sample size proportions × 2

parameter generation procedures = 12 simulation conditions for the "no DIF" condition. Therefore, a total of 660 simulation conditions were examined. Within each of these conditions, results were compared among the three algorithm types (Morris et al., 2019). Note that several model and data features were held constant across the simulation, including two groups of simulees, group-level latent trait distributions with $\theta_R \sim \mathrm{N}(0,1)$ and $\theta_F \sim \mathrm{N}(0.5,1)$, dichotomous item types, no cases of model misspecification, and a target test length of $n_{\text{test}} = 20$ items. These features were not manipulated in the current study to manage the simulation size.

It merits comment that the given simulation procedure implicitly asserts a set of assumptions about the item responses and DIF type. Specifically, it was assumed that the items still followed the model-implied item response functions even when item parameter values were modified to incorporate DIF. This type of DIF might reflect a scenario where individuals with different characteristics attribute different meaning to an item, but the underlying model remains the same across groups. IRT research commonly simulates DIF in this way (e.g., Belzak & Bauer, 2020; Finch, 2016; E. S. Kim & Yoon, 2011). However, other methods for introducing DIF are possible, including altering the dimensionality of the underlying latent trait for a particular group.

The Study 1 simulation for a given condition proceeded as follows. A 60-item bank was first generated, where each item's parameters were random realizations of uniform distributions centered at the item's (true or estimated) population values. Next, examinee θ values were randomly drawn from a bivariate normal distribution, reflecting the latent abilities for the proposed test and the external criterion measure. The θ and item parameter values were then used to generate item responses. Regularized DIF and

multiple-group IRT estimation were applied to these item responses to compute the bank-level indices (e.g., DIF severity, item information values, etc.). TIF deviations were computed across a sequence of 13 θ values where $\theta \in \{-3, -2.5, \ldots, 2.5, 3\}$. Moreover, the expected score curves used 100 quadrature points ranging from -3 to 3. The bank-level indices served as the input for the ATA algorithms. The "best" test selected by each of the algorithms was then evaluated on a suite of psychometric and algorithm properties (described in the next section). Overall, this simulation procedure was repeated $R = 100$ times for each condition.

One note about the regularized DIF implementation merits comment. With an increasing number of items (n), examinees (N), and τ values, the θ estimation in regularized DIF can be computationally expensive. An option to reduce computation time is to use observed sum scores as a proxy for θ values (Belzak, 2021). A preliminary simulation was conducted to compare false positive rates (FPRs) and true positive rates (TPRs) with and without using the proxy scores across a handful of conditions. Using proxy scores tended to result in higher FPRs when DIF percentages and magnitudes were large. For example, the average FPRs with and without using proxy scores when 1/2 of the item bank included DIF were 0.32 and 0.21, respectively. There were negligible differences in average FPRs across other examined DIF conditions. Importantly, using proxy scores resulted in noticeably higher TPRs across many examined conditions. Given the relative importance of high TPRs for DIF identification, coupled with substantial reductions in computation time (arguably a barrier for practical implementations), regularized DIF used sum scores rather than θ estimation in Study 1.

Performance Evaluation

Numerous variables were used to identify the testing scenarios wherein Unbiased-ATA selected tests with relatively strong psychometric properties. First, the Study 1 simulation evaluated the FPRs and TPRs for regularized DIF across the varying simulation conditions. A false positive referred to a truly invariant item that was estimated to have non-zero DIF parameters. Similarly, a true positive referred to a truly non-invariant item with non-zero estimated DIF parameters. These analyses add to a growing literature on the efficacy of this method (Bauer et al., 2020; Belzak & Bauer, 2020; Magis et al., 2011; C. Wang et al., 2022).

The characteristics of the "best" test selected by 0-1 LP, Tabu search, or ACO were examined in multiple ways. Specifically, the criteria that comprise the Unbiased-ATA objective function were also dependent variables in the Study 1 simulation. For example, each proposed test was evaluated on the number of well-fitting items selected based on the $S - \chi^2$ statistics, and the deviations between the estimated and the target TIFs. Related to test score precision, the overall TIF values (i.e., the sum of the item information values) were also computed for a range of θ values (i.e., $-3 \leq \theta \leq 3$). Furthermore, the correlations between the estimated θ values on the proposed test and the external criterion measure were calculated. For all tests (i.e., both the newly developed test and the criterion measure), θ was estimated using maximum likelihood (ML) estimation according to the data-generating model (i.e., the 2PLM).[4] Correlations were then computed as simple Pearson product-moment correlations, both for the full sample

[4] If item response patterns were not mixed (i.e., all responses were either 0 or 1), the ML estimation bounded the θ estimate at either -6 or 6, following specifications in the *catIrt* R package (Nydick, 2014).

and by group. Summary statistics for the chosen item parameters were also examined, providing insight into whether certain algorithms and simulation conditions resulted in tests with higher or lower discrimination or difficulty parameters.

To measure evidence of MI in the proposed test, the simulation computed the number of selected invariant and non-invariant items. In this context, larger numbers of invariant items, and smaller proportions of non-invariant items, reflected more desirable algorithm performance. In addition to item-level MI, the uDTF effect size was calculated to quantify test-level MI for the proposed test. Finally, the test data were fit to a series of successively more restrictive multiple-group IRT models aligning with the configural, weak, and strong MI levels from the SEM framework. Based on guidelines from Maydeu-Olivares (2015) and Maydeu-Olivares and Joe (2014), both a full-sample RMSEA and group-level standardized root mean squared residual (SRMSR) values were computed for each of these models to identify the best-fitting MI level.

Analysis Plan

The first set of analyses in Study 1 identified the simulation design factors with the largest effects on each of the examined dependent variables. In this design, the DIF characteristics—DIF percentage, magnitude, and direction—were only applicable within the "non-uniform DIF", "uniform DIF", and "mixture DIF type" conditions. In other words, these three DIF characteristic variables were nested within DIF type. This partially-nested experimental design precludes fitting fully-crossed, multiway analysis of variance (ANOVA) models, because not all levels of DIF type are paired with all levels of DIF percentage, magnitude, and direction (Oehlert, 2018; Sahai & Ageel, 2000).

To appropriately analyze this simulation design and calculate effect sizes, a set of three models were constructed for each dependent variable. First, two multiway ANOVAs were fit using the corresponding design factors and their three-way interactions. In these models, psychometric properties of the "best" selected test were used as dependent variables (e.g., TIF deviations, number of invariant items), and dependent variables were averaged across the $R = 100$ simulation repetitions for each condition. In Model Type 1, the dependent variable was regressed on five factors using the full data set: (a) sample size, (b) sample size balance, (c) estimation type, (d) algorithm type, and (e) DIF type. Dummy coding was used for all categorical design factors. Then in Model Type 2, the dependent variable was regressed on all design factors in Model Type 1, plus DIF percentage, magnitude, and direction. Model Type 2 was only fit to the data from conditions with DIF (allowing for a fully crossed design).

Then, based on the selected model, partial η^2 values (η_p^2) for design factor x were calculated as (Cohen, 1973, p. 108)

$$\eta_p^2 = \frac{\mathrm{SS}_x}{\mathrm{SS}_x + \mathrm{SS}_E}, \tag{24}$$

where SS_x is the sums of squares for x and SS_E is the corresponding error sums of squares. The Type II sums of squares (also referred to as "hierarchical" sums of squares) were used to compute η_p^2. To obtain the Type II sums of squares for design factor x, a model with x and all other factors is compared to a model without x and any other terms including x (i.e., respecting the "hierarchy principle"). Cohen's (1992) guidelines for η_p^2 magnitudes were used to identify important design factors, wherein $0.13 \leq \eta_p^2 < 0.26$ was considered a moderate effect and $\eta_p^2 \geq 0.26$ was considered a large effect.

An alternative approach to analyze a partially-nested design is using mixed-effects models (e.g., Oehlert, 2018). Therefore, Model Type 3 conceptualized the nested DIF characteristics—DIF percentage, magnitude, and direction—as random effects nested within DIF type. In Model Type 3, fixed effects included the two-way interactions and lower terms among sample size, sample size balance, estimation type, algorithm type, and DIF type. Two-way interactions among the fixed effects were used to prevent over-fitting when random effects were also included. Model Type 3 was fit using restricted maximum-likelihood estimation, and η_p^2 effect sizes for the fixed-effects were computed using Type II sums of squares and Satterthwaite degree of freedom approximations (the latter of which has shown comparable performance to the Kenward-Rogers approximation in many contexts; e.g., Luke, 2017). Moreover, factor-level intraclass correlations (ICCs) were computed as effect sizes for the random effects. Cicchetti's (1994) guidelines for ICC magnitudes were used, with ICCs greater than 0.40 considered noteworthy (as cited in Hallgren, 2012).

Granted, the three model types each possess unique limitations for the given experimental design. For instance, subsetting the data to only comprise DIF conditions for Model Type 2 prevents drawing comparisons to the "no DIF" conditions. Relatedly, it might not be appropriate to conceptualize the DIF characteristics as random effects with linear mixed-effects models. Given these limitations, important design factors were selected as those with moderate to large effects across all three models. In other words, no one model type or effect size value was emphasized; rather, the models were considered together as a robustness check for substantial effects that warranted further examination.

Using the effect sizes as guidance, figures were constructed to elucidate the performance of Unbiased-ATA across the examined conditions and algorithms. Again, the overarching goal of these analyses was to identify testing scenarios wherein Unbiased-ATA selected tests with relatively strong psychometric properties. Influential simulation design factors, as drawn from the effect size analyses and figures, were then selected to craft the subsequent simulations in Studies 2 and 3.

Software

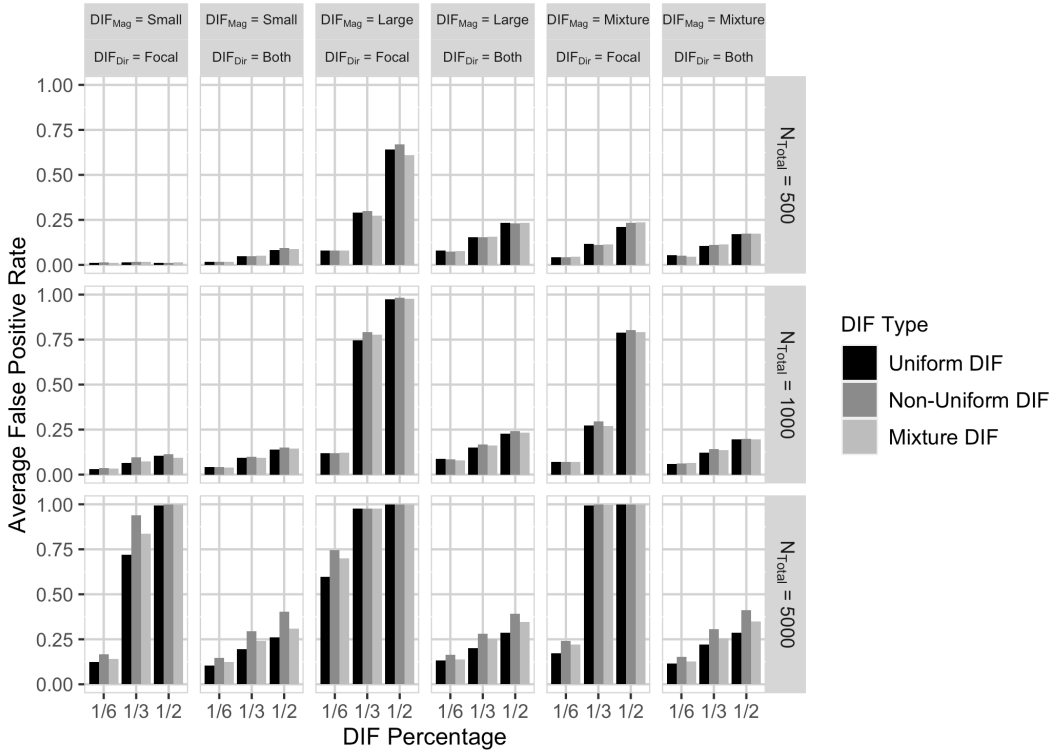
All analyses were conducted in R statistical software, version 4.1.1 (R Core Team, 2021). As previously noted, item response generation was completed using the *catIrt* package (Nydick, 2014). This package was also used to compute the expected item and test information functions, as well for θ estimation. True θ values were drawn from bivariate normal distributions using the *MASS* package (Venables & Ripley, 2002). Regularized DIF analyses were completed using the *regDIF* package (Belzak, 2021), whereas MML estimation, multiple-group IRT model estimation, the uDTF effect size, and IRT model fit indices were computed using *mirt* (Chalmers, 2012). Functions to compute the wABC metric were based on code from Edelen et al. (2015) and Stucky et al. (2014). Moreover, Tabu search was implemented with code modified from Marcoulides and Falk (2018), Raborn and Leite (2018), and Raborn et al. (2020). The ACO algorithm used modified code from Jankowsky et al. (2020) and Raborn and Leite (2018). The *CVXR* (Fu et al., 2020) and *Rglpk* packages (Theussl et al., 2019) were used for 0-1 LP optimization. Finally, the R packages *effectsize* (Ben-Shachar et al., 2020), *lme4* (Bates et al., 2015), *lmerTest* (Kuznetsova et al., 2017), and *ggplot2* (Wickham, 2016) were used for analysis. All code is available from the author by request.

Results

Regularized DIF Performance

Unbiased-ATA begins by identifying invariant and differentially functioning items in the bank using regularized DIF (Belzak & Bauer, 2020). Comparing across the DIF types, average FPRs ($\overline{FPR}$; computed across the simulation replications for each condition) were smallest when the item bank did not contain DIF. In these "no DIF" conditions, $\overline{FPR}$ was extremely small and ranged between 0.000 to 0.002 (Figure A1 in the appendix). Regularized DIF was slightly more likely to erroneously identify items as differentially functioning when the total sample size was large and group-level sample sizes were unbalanced. However, the magnitude of differences between balanced and unbalanced samples was only 0.001. Thus, regularized DIF almost never misidentified an invariant item as differentially functioning when no DIF was present in the item bank.

Figure 2. Average False Positive Rates for Regularized DIF



$\overline{\text{FPR}}$ substantially increased when either 10, 20, or 30 items were simulated as differentially functioning in the 60-item bank. Figure 2 presents the $\overline{\text{FPR}}$ when marginalizing across estimation method and sample size balance. Here, $\overline{\text{FPR}}$ increased with higher DIF percentages, larger DIF magnitudes, and larger sample sizes. For example, holding the other design factors constant, the median $\overline{\text{FPR}}$ was 0.093 for small DIF magnitudes compared to 0.244 for large DIF magnitudes. There were smaller differences in $\overline{\text{FPR}}$ based on DIF type, with median $\overline{\text{FPR}}$ of 0.146, 0.163, and 0.151 for uniform, non-uniform, and mixture DIF types, respectively.

Figure 2 also highlights that $\overline{\text{FPR}}$ differed based on the direction of DIF. When DIF modifications were applied evenly to both the reference and focal groups (i.e., the "both groups" direction; Columns 2, 4, and 6), $\overline{\text{FPR}}$ ranged from 0.012 to 0.433. Yet $\overline{\text{FPR}}$ routinely exceeded 0.50 and often reached close to 1.00 when DIF modifications were applied only to the focal group (i.e., the "focal group" direction; Columns 1, 3, and 5). Here, $\overline{\text{FPR}}$ was largest with one-third or more of the items having DIF and $N \geq 1000$. Given large DIF percentages and sample sizes, $\overline{\text{FPR}}$ exceeded 0.75 even with small DIF magnitudes (Column 1, Row 3).

To better understand the relationship between $\overline{\text{FPR}}$ and DIF direction, consider the item parameter differences between groups. In each item bank with simulated DIF, the absolute value differences between reference and focal group difficulty (b) and discrimination (a) parameters were computed for each item. These differences were then summed across all 60 items in the bank. Table 2 presents the average difference magnitudes for each parameter based on DIF characteristics. Notice that in the "focal group" direction, parameter differences across the full test were exacerbated. On the

contrary, in the “both groups” direction, parameter differences largely "canceled out" when summed across the test. In the first case, the DIF detection method might flag more items as differentially functioning.

Table 2. Average Magnitude of Item Parameter Differences Summed Across Item Banks for Varying DIF Characteristics

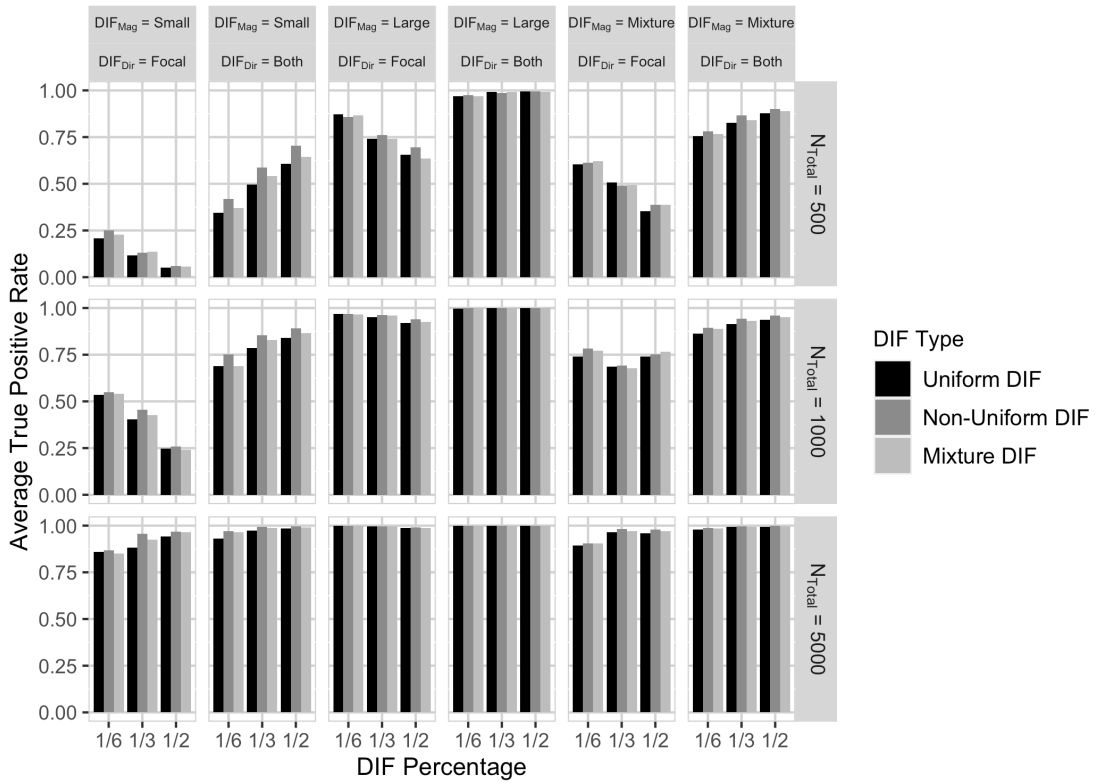
	Small DIF Magnitude		Large DIF Magnitude		Mixture DIF Magnitude	
DIF Percentage	Focal	Both	Focal	Both	Focal	Both
Difficulty Parameter, Uniform DIF						
1/6	4.25	0.49	8.31	0.48	6.29	0.75
1/3	8.84	0.73	16.95	0.71	12.88	1.01
1/2	13.05	0.83	25.22	0.87	19.16	1.22
Difficulty Parameter, Non-Uniform DIF						
1/6	4.89	0.54	8.55	0.53	6.70	0.75
1/3	10.27	0.74	17.53	0.81	13.84	1.06
1/2	14.84	0.96	25.93	0.96	20.43	1.26
Discrimination Parameter, Non-Uniform DIF						
1/6	3.14	0.67	2.49	0.81	2.83	0.79
1/3	6.83	1.02	5.75	1.14	6.23	1.13
1/2	8.65	1.23	7.24	1.48	8.00	1.36

Note. “Focal” indicates that DIF modifications were made only to the focal group. “Both” indicates that DIF modifications were made evenly to both the reference and the focal groups. Difference magnitudes were averaged across the simulation replications for each condition.

Regularized DIF’s high FPRs often translated to high power to identify truly differentially functioning items. Figure 3 demonstrates that TPRs averaged near 1.00 in many examined conditions. The relationship between average TPRs ($\overline{\text{TPR}}$) and DIF percentage depended upon DIF magnitude, DIF direction, and sample size. Notice that $\overline{\text{TPR}}$ was highest when $N = 5{,}000$ and DIF magnitudes were large (Columns 3 – 4, Row 3), and the lowest $\overline{\text{TPR}}$ occurred for small DIF modifications applied only to the focal group with $N = 500$ (Column 1, Row 1). For the “focal group” direction, $\overline{\text{TPR}}$ generally decreased as DIF percentage increased. Alternatively, for the “both groups” direction, $\overline{\text{TPR}}$ increased as DIF percentages increased for smaller sample sizes and DIF

magnitudes. This interaction among DIF percentage and magnitude weakened as sample size increased.

Figure 3. Average True Positive Rates for Regularized DIF



Overall, DIF characteristics in the item bank influenced regularized DIF's ability to accurately categorize invariant and non-invariant items. Regularized DIF often mistakenly identified truly invariant items as differentially functioning, while simultaneously correctly identifying truly non-invariant items as differentially functioning. At times both $\overline{\text{FPR}}$ and $\overline{\text{TPR}}$ neared 1.00, indicating that close to all items in the bank were categorized as differentially functioning.

Algorithm Performance

The Unbiased-ATA objective function was paired with three algorithms—0-1 LP, ACO, and Tabu Search—in Study 1. Comparisons of the three algorithms in relation to

the ATA-selected tests' psychometric properties are detailed in the next section. Here, the algorithms are compared on their speed and performance quality separate from the properties of the selected tests. Across the simulation conditions, the ACO algorithm tended to find solutions with lower objective functions ($Mean_{ACO}$ = 5.80, $Median_{ACO}$ = 5.88, $Range_{ACO}$ = 5.13 – 5.99) than 0-1 LP ($Mean_{0\text{-}1\ LP}$ = 5.86, $Median_{0\text{-}1\ LP}$ = 5.96, $Range_{0\text{-}1\ LP}$ = 5.20-6.00) or Tabu search ($Mean_{Tabu\ Search}$ = 5.86, $Median_{Tabu\ Search}$ = 5.96, $Range_{Tabu\ Search}$ = 5.20 – 6.00).[5] The differences in objective function values across item bank characteristics were relatively small, and generally mirrored the number of items in the bank that regularized DIF categorized as differentially functioning. Although ACO tests often had the lowest objective function values, this algorithm type was consistently the fastest algorithm to find a solution (Mean = 0.06 minutes, SD = 0.005, Range = 0.04 – 0.07). Tabu search followed close behind in terms of speed (Mean = 1.48 minutes, SD = 0.08, Range = 1.14 – 1.65), whereas 0-1 LP was the slowest algorithm on average (Mean = 6.37 minutes, SD = 0.81, Range = 4.79 – 9.58). Note that although the 0-1 LP solver was given a time limit of five minutes, the recorded elapsed time could exceed this limit due to additional time spent during solver set-up. The 0-1 LP algorithm tended to take longer when the item bank did not contain simulated DIF compared to conditions with simulated DIF ($Mean_{No\ DIF}$ = 8.66 minutes; $Mean_{DIF}$ = 6.33 minutes). Otherwise, differences in average times across the simulation conditions were negligible.

The 0-1 LP algorithm was likely the slowest because this algorithm did not always find an optimal solution. Indeed, longer solver times more often occurred in

[5] Note that the maximum objective function value possible was 6.0, since $\gamma_{\text{Precision}}$ and γ_{IF} were both composed of two terms (for the reference and focal group), each ranging from 0.0 to 1.0.

conditions where more simulation trials returned non-optimal solutions ($r = 0.668$). Figure 4 and Figure 5 present boxplots for the proportion of optimal solutions found across various simulation conditions. The distributions for these proportions were often either narrowly centered near 1.00 or noticeably left-skewed, indicating that an optimal solution was discovered in most, if not all, simulation repetitions. However, notice in Figure 4 that the proportion of optimal solutions was substantially smaller in conditions without simulated DIF (Mean = 0.51, Median = 0.54) compared to conditions with simulated DIF (Mean = 0.96, Median = 0.98). In the "no DIF" conditions (Row 1), 0-1 LP less often found an optimal solution when using true compared to estimated parameters. Moreover, Figure 5 highlights that in conditions with simulated DIF, the proportion of optimal solutions was smaller when (a) DIF magnitude and percentage decreased, (b) simulating uniform DIF, and (c) in the "focal group" direction. Interestingly, 0-1 LP also had greater difficulty finding optimal solutions with higher percentages of large DIF in the "focal group" direction with uniform or mixture DIF types (Column 3).

Given that 0-1 LP did not always select an optimal item combination, the psychometric characteristics of the ATA-selected tests were analyzed when including or excluding the non-optimal solutions. In the first case, results were averaged across all 100 simulation trials regardless of the 0-1 LP solution status. In the second case, results were averaged across all simulation trials wherein the 0-1 LP solution status was "optimal." The same trials were selected for all three algorithm types (0-1 LP, Tabu search, or ACO), but the number of trials over which results were averaged could differ by condition. For example, one condition's averaged results might be based on 76 trials

Figure 4. Average Proportion of Optimal Solutions Found by 0-1 LP Across DIF Types, Estimation Type, and Sample Size

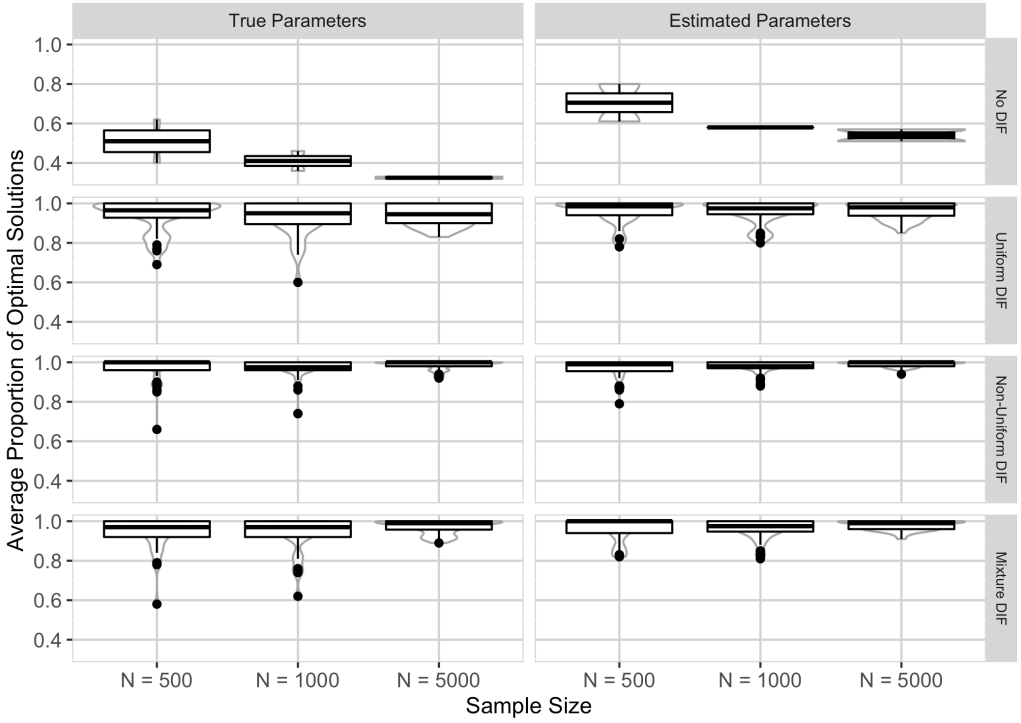
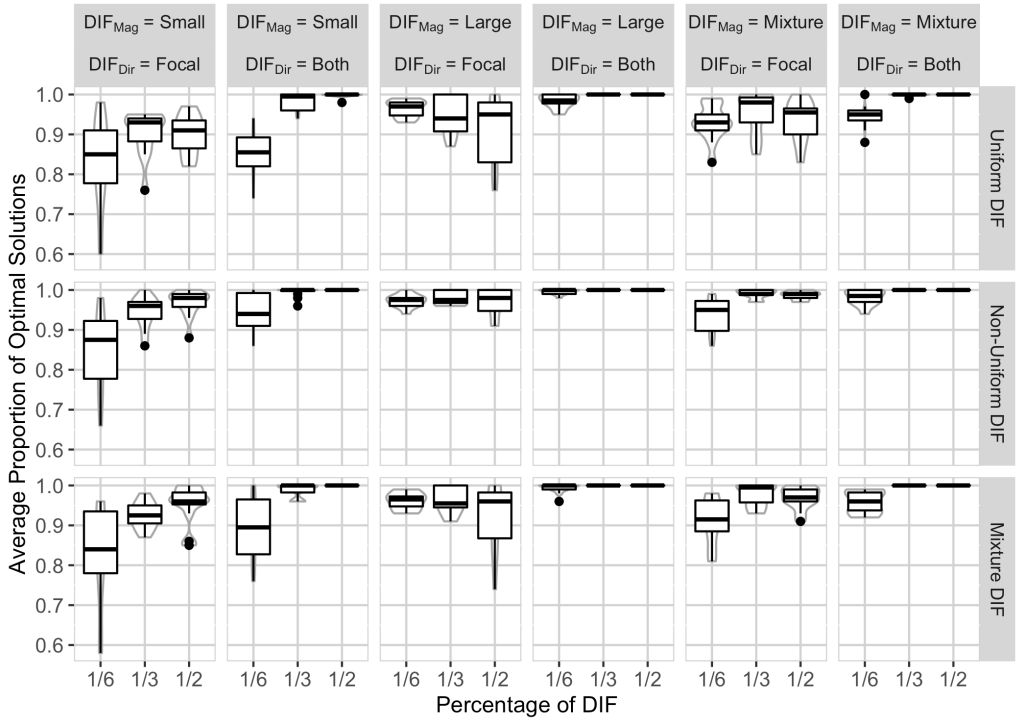


Figure 5. Average Proportion of Optimal Solutions Found by 0-1 LP Across DIF Characteristics



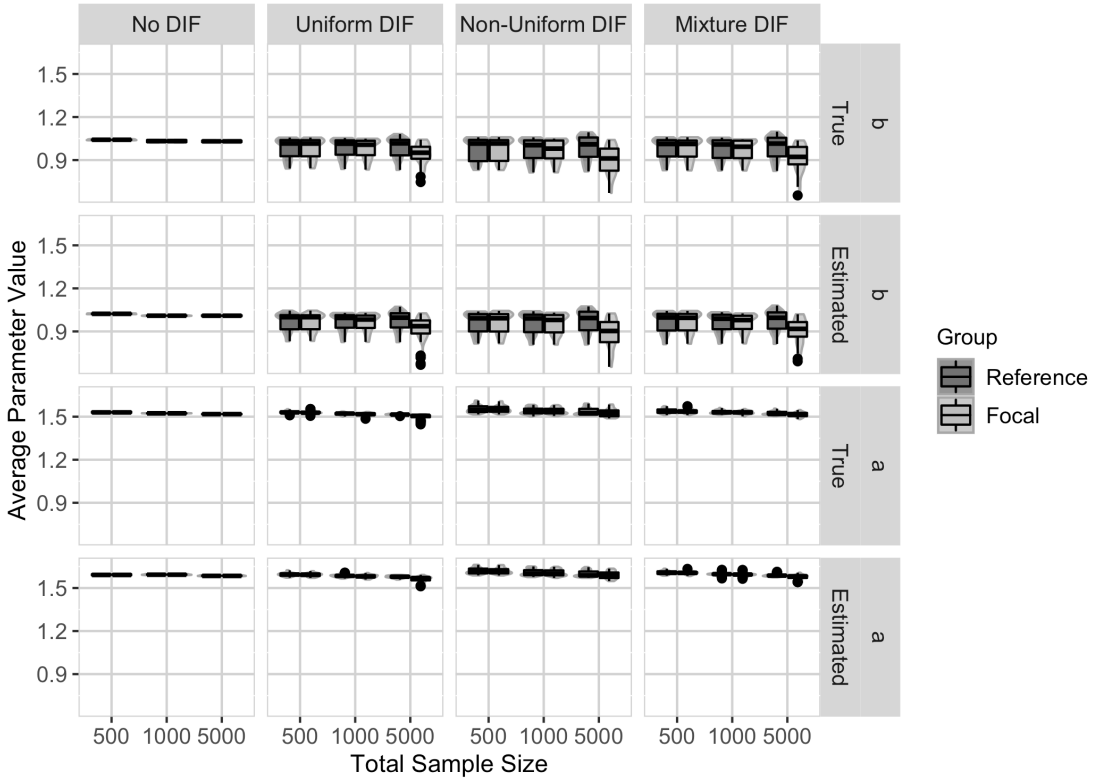
conditions whereas another condition's averaged results might be based on 100 trials. Overall, there were negligible differences in ATA-selected test properties whether including or excluding trials with non-optimal solutions. Thus, the following results are presented with all simulation trials regardless of 0-1 LP solution status. The Study 1 discussion further examines why 0-1 LP returned non-optimal solutions within certain item banks.

Parameter Summaries of the Selected Tests

To evaluate the characteristics of the ATA-selected tests, descriptive statistics for the selected item parameters were first examined. The item parameters described here were estimated when fitting the item bank data to a multiple-group IRT model in the preliminary Unbiased-ATA analyses. Within each simulation repetition, summary statistics (e.g., mean, standard deviation, and range) were calculated for the estimated difficulty ($\hat{b}$) and discrimination ($\hat{a}$) values among the 20 ATA-selected items. The mean $\hat{b}$ and $\hat{a}$ were then averaged across the 100 simulation repetitions (i.e., the following results present "averaged averages").

Figure 6 presents boxplots and overlying violin plots of the average $\hat{b}$ (Rows 1 – 2) and $\hat{a}$ values (Rows 3 – 4) across DIF types, estimation type, and total sample size. Overall, average $\hat{b}$ and $\hat{a}$ remained relatively consistent at approximately 1.00 and 1.60, respectively. Across all conditions, average $\hat{b}$ ranged from 0.00 to 2.00 for the reference group, and from -0.02 to 2.00 for the focal group. Similarly, average $\hat{a}$ ranged from 1.16 to 2.02 for the reference group, and from 1.16 to 2.03 for the focal group. Using a particular algorithm type (0-1 LP, ACO, or Tabu search) was not associated with substantially different item parameter estimates.

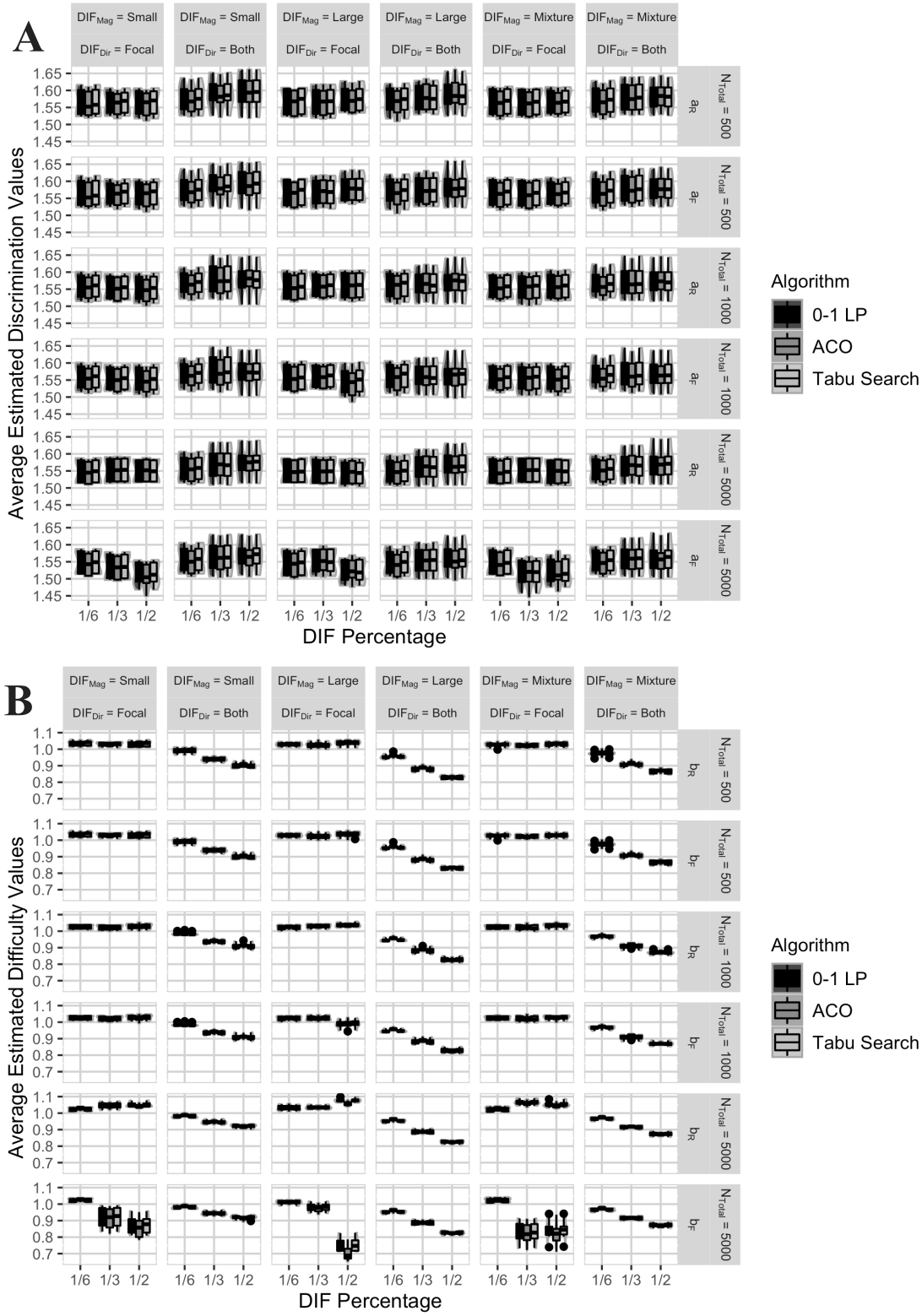
Figure 6. Average Difficulty and Discrimination Values for Selected Tests Across DIF Type, Estimation Type, and Sample Size



These plots also indicate larger variation in average $\hat{b}$ when DIF was simulated, particularly for the focal group at $N = 5000$. Also notice in this figure that average $\hat{a}$ was uniformly higher when using estimated versus true parameters for item response generation. It is likely that the initial MML estimation generally overestimated the a parameters, leading to tests with higher $\hat{a}$ in the "estimated parameter" compared to the "true parameter" conditions. Yet differences in average $\hat{a}$ were not large: the average $\hat{a}$ values hovered around 1.53 when using true parameters and 1.59 when using estimated parameters.

Figure 7 presents the average $\hat{a}$ (Panel A) and $\hat{b}$ (Panel B) values exclusively within DIF conditions. Specifically, Figure 7B elucidates differences in $\hat{b}$ as a function of

Figure 7. Average Discrimination and Difficulty Values for Selected Tests Across DIF Characteristics



DIF direction. In the "focal group" direction, average $\hat{b}_{\mathrm{F}}$ decreased as DIF percentage increased whereas average $\hat{b}_{\mathrm{R}}$ remained relatively consistent; there was also increasing variation in average $\hat{b}_{\mathrm{F}}$ for higher DIF percentages. Yet in the "both groups" direction, both average $\hat{b}_{\mathrm{R}}$ and $\hat{b}_{\mathrm{F}}$ systematically decreased with higher DIF percentages. These trends suggest that the parameter estimation generally reflected the simulated DIF patterns in the item bank. Among item banks with various DIF characteristics, there were negligible differences in average $\hat{a}$.

Psychometric Properties of the Selected Tests

Effect Size Results. A series of fixed-effects and mixed-effects models were first fit to the simulation data to calculate effect sizes for each design factor. Next, η_p^2 values[6] were calculated for the fixed effects in Model Types 1 through 3, and ICCs were calculated for the random effects in Model Type 3. Tables 3 through 5 present these effect sizes for each psychometric property. For simplicity, Table 4 (Model Type 2 with all three-way interactions) presents only design factors and associated interactions that had at least one $\eta_p^2 \geq 0.13$ (Cohen, 1992).

Before reviewing the effect size trends, it is important to note that the standard and adjusted R^2 values for Model Type 1 (the two-way ANOVAs incorporating sample

[6] Partial ω^2 omega-squared values were also calculated. Because the trends of substantive effects were consistent across the two indices, only the η_p^2 values are presented here. It is worth noting that numerous negative ω_p^2 values occurred because of F statistics below 1.00. These negative ω_p^2 values corresponded to negligible to small η_p^2 values.

size, algorithm, estimation, and DIF type) were often extremely low. For six psychometric properties, the R^2 values were less than 0.50 and sometimes less than 0.10 (e.g., when regressing the external validity coefficients or RMSEA). These low coefficients of determination indicate that the included predictors did not account for substantial variation in the data. The exact η_p^2 values from these models should thus be interpreted with caution. Here, design factors with non-negligible effects across the three model types were used to inform further graphical interpretations.

Tables 3 – 5 show that certain design factors consistently displayed moderate to large effects on the psychometric properties of the ATA-selected tests. For example, η_p^2 values for total sample size exceeded 0.13 for most properties, and often exceeded 0.26 (the lower bound for large effect sizes; Cohen, 1992). The sample size effect sizes were especially large for test characteristics related to item- and test-level MI. When modeled as fixed effects in Model Type 2, DIF percentage, magnitude, and direction also influenced the ATA-selected test properties (see Table 4). The η_p^2 values for these characteristics were frequently large when examining test information and MI indices, both individually and when combined with total sample size. However, Table 5 shows that when modeled as nested random effects, the ICCs for the DIF characteristics rarely surpassed Cicchetti's (1994) thresholds for "fair" correlations (ICC ≥ 0.40). Indeed, some linear mixed-effects models returned boundary cases because the estimated random effects were approximately zero. The larger effect sizes in Model Type 2 compared to Model Type 3 for the DIF characteristics might be a product of using subsetted data for Model Type 2, as well as the estimation differences between conceptualizing these factors as fixed versus random effects.

Table 3. Partial η^2 Effect Sizes When Regressing Test Properties on Sample Size, Algorithm, Estimation, and DIF Type

	Measurement Invariance					Information		Item Fit		Validity	
Design Factor	$n_{\text{No DIF}}$	uDTF	RMSEA	SRMSR_R	SRMSR_F	$\Delta_{\text{TIF}_\text{R}}$	$\Delta_{\text{TIF}_\text{F}}$	$n_{\text{Fitting}_\text{R}}$	$n_{\text{Fitting}_\text{F}}$	r_R	r_F
Total Sample Size (N)	**0.23**	**0.19**	0.01	**0.71**	**0.98**	0.12	**0.20**	0.03	**0.14**	0.01	0.02
Group Sample Size (N_{Group})	0.00	0.00	0.01	**0.35**	**0.91**	0.00	0.00	0.00	0.04	0.00	0.01
Estimation	0.00	0.00	0.00	0.00	0.00	**0.78**	**0.78**	0.00	0.00	0.00	0.01
Algorithm	0.01	0.01	0.04	0.00	0.03	0.01	0.02	**0.39**	**0.54**	0.00	0.00
DIF Type	0.02	0.01	0.02	0.03	0.01	**0.21**	**0.25**	0.01	0.01	0.01	0.01
N x N_{Group}	0.00	0.00	0.01	0.02	**0.56**	0.00	0.00	0.00	0.04	0.01	0.01
N x Estimation	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
N x Algorithm	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	**0.18**	0.00	0.00
N x DIF Type	0.01	0.01	0.00	0.00	0.02	0.00	0.01	0.00	0.00	0.00	0.01
N_{Group} x Estimation	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
N_{Group} x Algorithm	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00
N_{Group} x DIF Type	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.01	0.00
Estimation x Algorithm	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Estimation x DIF Type	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Algorithm x DIF Type	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.00	0.00
N x N_{Group} x Estimation	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
N x N_{Group} x Algorithm	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00
N x N_{Group} x DIF Type	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.00
N x Estimation x Algorithm	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
N x Estimation x DIF Type	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
N x Algorithm x DIF Type	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
N_{Group} x Estimation x Algorithm	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
N_{Group} x Estimation x DIF Type	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
N_{Group} x Algorithm x DIF Type	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Estimation x Algorithm x DIF Type	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
R^2	0.247	0.212	0.084	0.751	0.987	0.796	0.807	0.433	0.637	0.046	0.067
Adjusted R^2	0.210	0.174	0.040	0.739	0.986	0.786	0.798	0.405	0.619	0.000	0.022

Note. $n_{\text{No DIF}}$ = Number of "no DIF" items; uDTF = unsigned differential test functioning effect size; RMSEA = root mean square error of approximation; R = reference group; F = focal group; SRMSR = standardized root mean square residual; Δ_{TIF} = TIF deviation; n_{Fitting} = Number of well-fitting items; r = External validity coefficient. Effect sizes greater than or equal to 0.13 are bolded (Cohen, 1992).

Table 4. Partial η^2 Effect Sizes When Regressing Test Properties on Sample Size, Algorithm, Estimation, and DIF Characteristics Among Conditions with DIF

	Measurement Invariance					Information		Item Fit		Validity	
Design Factor	$n_{\text{No DIF}}$	uDTF	RMSEA	SRMSR_R	SRMSR_F	$\Delta_{\text{TIF}_\text{R}}$	$\Delta_{\text{TIF}_\text{F}}$	$n_{\text{Fitting}_\text{R}}$	$n_{\text{Fitting}_\text{F}}$	r_R	r_F
Total Sample Size (N)	**0.87**	**0.78**	**0.24**	**0.98**	**1.00**	**0.51**	**0.80**	0.06	**0.25**	0.01	0.02
Group Sample Size (N_Group)	0.00	0.00	**0.24**	**0.92**	**0.99**	0.02	0.00	0.00	0.07	0.00	0.01
Algorithm	**0.20**	0.10	**0.57**	0.04	**0.22**	**0.88**	**0.38**	**0.53**	**0.71**	0.00	0.00
DIF Type	0.06	0.09	0.08	**0.36**	0.09	**0.35**	**0.13**	0.01	0.00	0.00	0.01
DIF Percentage (Perc)	**0.87**	**0.66**	**0.94**	**0.84**	**0.76**	**0.86**	**0.76**	0.09	0.14	0.05	**0.23**
DIF Magnitude (Mag)	**0.65**	**0.17**	**0.77**	**0.16**	**0.29**	**0.66**	**0.29**	0.07	0.12	0.01	0.06
DIF Direction (Dir)	**0.79**	**0.65**	**0.71**	**0.89**	**0.44**	**0.29**	**0.61**	0.00	0.00	**0.25**	0.01
N x N_Group	0.00	0.00	**0.19**	**0.29**	**0.92**	0.00	0.00	0.00	0.07	0.01	0.01
N x Algorithm	0.00	0.00	0.00	0.00	0.01	**0.20**	0.01	0.08	**0.32**	0.00	0.00
N x Perc	**0.61**	**0.63**	0.20	**0.28**	**0.55**	**0.14**	**0.66**	0.02	0.05	0.01	0.01
N x Mag	**0.20**	0.07	**0.19**	**0.21**	0.10	0.12	0.06	0.01	0.04	0.00	0.00
N x Dir	**0.78**	**0.78**	**0.21**	0.11	0.02	**0.67**	**0.81**	0.00	0.01	0.00	0.00
Algorithm x Perc	0.04	0.02	**0.29**	0.02	0.08	**0.17**	0.01	0.08	**0.13**	0.00	0.00
Algorithm x Mag	0.01	0.01	**0.13**	0.00	0.04	0.00	0.00	0.06	0.10	0.00	0.00
Algorithm x Dir	0.05	0.00	**0.34**	0.03	0.07	0.01	0.02	0.01	0.00	0.00	0.00
Type x Perc	0.01	0.04	0.05	**0.25**	0.05	0.09	0.04	0.01	0.01	0.01	0.01
Type x Dir	0.01	0.07	**0.13**	0.08	0.01	0.04	0.02	0.02	0.01	0.00	0.00
Perc x Mag	**0.25**	**0.30**	**0.68**	**0.22**	**0.38**	**0.21**	**0.31**	0.02	0.01	0.01	0.02
Perc x Dir	**0.51**	**0.50**	**0.35**	**0.80**	**0.15**	012	**0.51**	0.02	0.00	0.08	0.03
Mag x Dir	**0.48**	0.08	0.10	**0.29**	0.06	0.04	0.07	0.00	0.01	0.05	0.01
N x Perc x Mag	**0.53**	**0.35**	**0.46**	**0.31**	0.07	**0.36**	**0.34**	0.04	0.05	0.01	0.02
N x Perc x Dir	**0.44**	**0.64**	**0.45**	**0.21**	0.12	**0.48**	**0.70**	0.02	0.02	0.01	0.01
N x Mag x Dir	**0.21**	0.08	**0.14**	**0.22**	0.05	0.04	0.09	0.01	0.02	0.00	0.01
Algorithm x Perc x Dir	0.03	0.00	**0.17**	0.02	0.02	0.00	0.02	0.01	0.00	0.00	0.00
Type x Perc x Dir	0.02	0.03	**0.14**	0.02	0.05	0.01	0.01	0.02	0.01	0.01	0.02
Perc x Mag x Dir	0.06	**0.28**	**0.21**	**0.21**	0.01	**0.20**	**0.30**	0.01	0.01	0.00	0.00
R^2	0.967	0.948	0.970	0.988	0.999	0.958	0.956	0.672	0.823	0.446	0.433
Adjusted R^2	0.961	0.937	0.964	0.986	0.998	0.950	0.948	0.609	0.789	0.338	0.322

Note. Only predictors with $\eta_p^2 \geq 0.13$ for at least one dependent variable are presented. Effect sizes greater than or equal to 0.13 are bolded (Cohen, 1992).

Table 5. Effect Sizes in Partially-Nested, Mixed-Effect Linear Models

	Measurement Invariance					Information		Item Fit		Validity	
Design Factor	$n_{\text{No DIF}}$	uDTF	RMSEA	SRMSR_{R}	SRMSR_{F}	$\Delta_{\text{TIF}_{\text{R}}}$	$\Delta_{\text{TIF}_{\text{F}}}$	$n_{\text{Fitting}_{\text{R}}}$	$n_{\text{Fitting}_{\text{F}}}$	r_{R}	r_{F}
Partial η^2 for Fixed Effects											
Total Sample Size (N)	**0.45**	**0.29**	0.06	**0.94**	**0.99**	**0.16**	**0.28**	0.04	**0.15**	0.01	0.02
Group Sample Size (N_{Group})	0.00	0.00	0.06	**0.78**	**0.96**	0.00	0.00	0.00	0.04	0.00	0.01
Estimation	0.00	0.00	0.00	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.01
Algorithm	0.03	0.01	**0.20**	0.01	0.07	**0.57**	0.05	**0.44**	**0.57**	0.00	0.00
DIF Type	0.00	0.03	0.00	0.06	0.02	0.03	0.03	0.03	0.00	0.00	0.02
N x N_{Group}	0.00	0.00	0.04	0.12	**0.76**	0.00	0.00	0.00	0.04	0.01	0.01
N x Estimation	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
N x Algorithm	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.06	**0.20**	0.00	0.00
N x DIF Type	0.00	0.01	0.00	0.00	0.04	0.01	0.01	0.00	0.00	0.00	0.00
N_{Group} x Estimation	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
N_{Group} x Algorithm	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00
N_{Group} x DIF Type	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.01	0.00
Estimation x Algorithm	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Estimation x DIF Type	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Algorithm x DIF Type	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ICCs for Random Effects											
DIF Direction (Dir)	0.19	0.16	0.11	0.37	0.11	0.04	0.10	0.00	0.00	0.27	0.00
DIF Percentage (Perc)	0.26	0.07	**0.53**	0.08	0.35	**0.48**	0.15	0.06	0.10	0.00	0.20
DIF Magnitude (Mag)	0.04	0.00	0.09	0.00	0.01	0.14	0.01	0.05	0.07	0.00	0.04
Perc x Dir	0.10	0.15	0.04	0.38	0.05	0.01	0.12	0.04	0.00	0.10	0.05
Dir x Mag	0.09	0.00	0.00	0.02	0.02	0.00	0.00	0.00	0.01	0.03	0.00
Perc x Mag	0.02	0.01	0.10	0.00	0.12	0.01	0.01	0.01	0.00	0.00	0.01
Perc x Dir x Mag	0.00	0.07	0.02	0.04	0.01	0.04	0.06	0.02	0.00	0.01	0.04

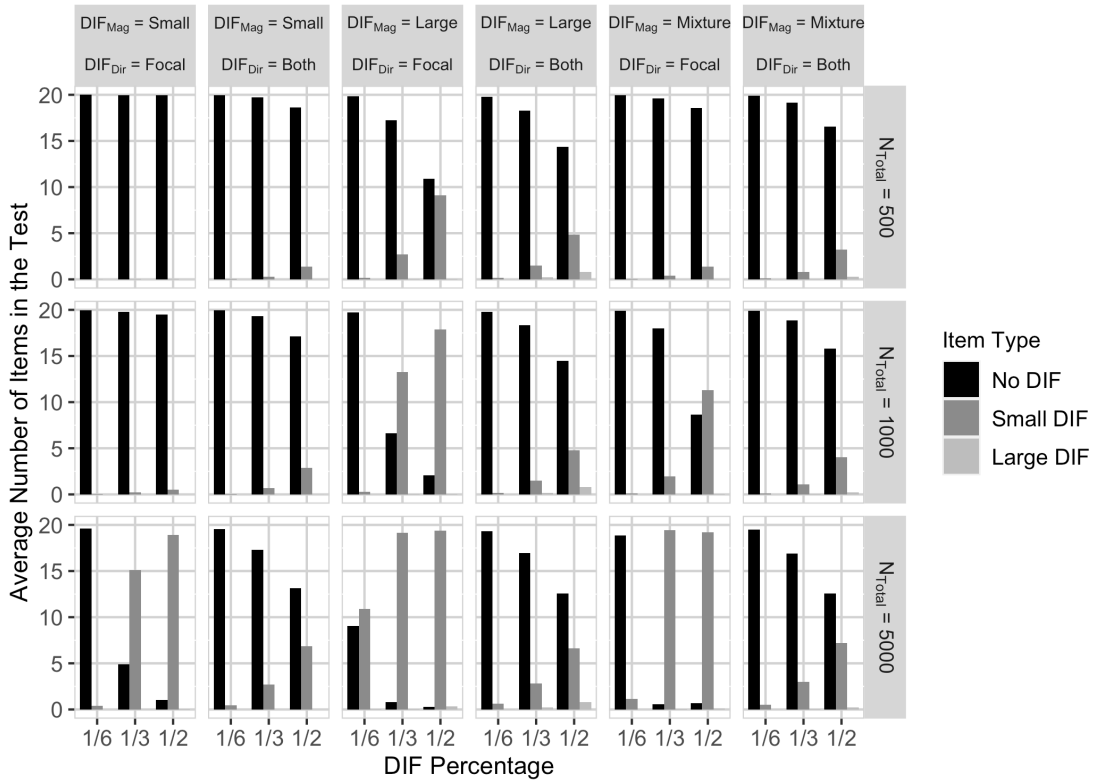
Note. ICC = intraclass correlation coefficient. Partial η^2 effect sizes greater than or equal to 0.13 (Cohen, 1992) and ICCs greater than or equal to 0.40 (Cicchetti, 1994) are bolded.

Furthermore, algorithm type—0-1 LP, ACO, or Tabu Search—most consistently demonstrated large effects for TIF deviations and the number of well-fitting items. In Model Type 2, algorithm type also had moderate to large effects on (a) the number of invariant items in the final test, (b) full-sample RMSEA for fitting the data to a strong MI model, and (c) focal group SRMSR for a strong MI model. Across the models, the interaction between total sample size and algorithm type influenced the number of well-fitting items in the focal group ($0.18 \leq \eta_p^2 \leq 0.32$). However, estimation method, sample size balance, and DIF type most often demonstrated small to negligible effects.

Item-Level MI with Regularized DIF Categorizations. Item-level MI was evaluated by comparing the number of items categorized as "no DIF," "small DIF," and "large DIF" items in each ATA-selected test. Here, the DIF categorizations were operationalized by a combination of regularized DIF's results (i.e., whether an item was identified as differentially functioning or not) and the wABC value. When DIF was not simulated in the item bank, the selected tests only comprised "no DIF" items. Given that the regularized DIF FPRs were very small, Unbiased-ATA understandably could only select among "no DIF" items in the bank.

Figure 8 presents the average number of each item type ("no DIF," "small DIF," and "large DIF") in the ATA-selected tests when the item banks contained DIF. In this figure, item type refers to the categorizations made by regularized DIF and the wABC index, separate from whether the items were simulated to be truly invariant in the simulation design. Comparing Figure 8 to Figure 2 (the $\overline{\text{FPR}}$ for regularized DIF), more differentially functioning items in the ATA-selected tests aligned with higher $\overline{\text{FPR}}$. When marginalizing across other design factors, Unbiased-ATA selected more differentially

Figure 8. Average Number of Differentially Functioning Items in Selected Tests Across DIF Characteristics



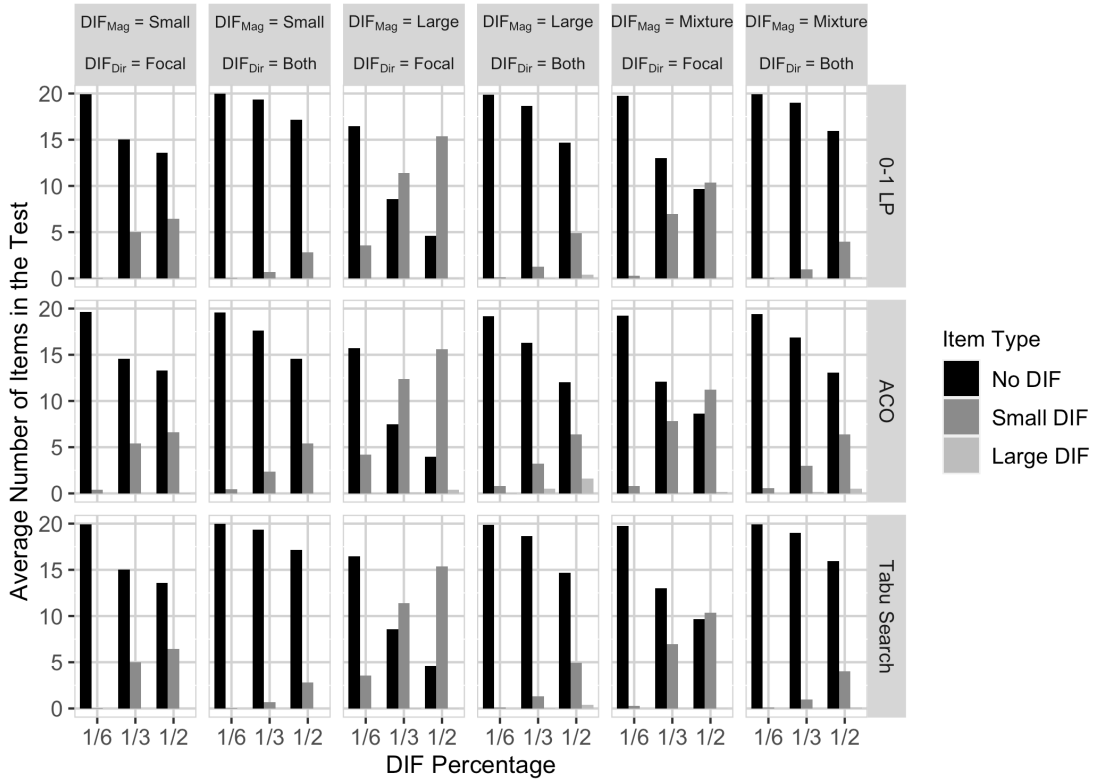
functioning items when (a) DIF magnitudes were large, (b) higher percentages of items in the bank had DIF, (c) DIF modifications were applied only to the focal group, and (d) the total sample size increased. These four design factors—DIF magnitude, percentage, direction, and total sample size—together produced many moderate to large main and interaction effects on the number of "no DIF" items selected.

The correspondence between regularized DIF's performance and the number of selected differentially functioning items can be understood when considering the Unbiased-ATA procedure. In conditions with higher regularized DIF $\overline{\text{FPR}}$ and $\overline{\text{TPR}}$, higher proportions of the item bank were categorized as differentially functioning. Unbiased-ATA then had fewer "no DIF" items to choose from when constructing the test. Indeed, if more than 40 items in the bank were categorized as "DIF," Unbiased-ATA was

required to select from these items to construct a 20-item test. Interestingly, Unbiased-ATA selected predominately "small DIF" items even when large magnitudes of DIF were simulated in the item bank (e.g., Figure 8, Column 3). This latter trend suggests that there were relatively small differences in parameter estimates between groups when fitting the item bank data to the multiple-group IRT model.

Concerning differences among the algorithm types, Figure 9 shows that ACO tended to derive tests with fewer "no DIF" items compared to 0-1 LP or Tabu search (corresponding to $\eta_p^2 = 0.20$ for algorithm type in Model Type 2). ACO either selected more "small DIF" items (e.g., when small DIF magnitudes were simulated), or selected relatively equivalent numbers of "small DIF" items but more "large DIF" items (e.g., when large DIF magnitudes and percentages were simulated). However, differences

Figure 9. Algorithm Differences in the Average Number of Differentially Functioning Items in Selected Tests



between ACO and either 0-1 LP or Tabu Search were small in magnitude. Holding other design factors constant, the average numbers of "no DIF" items in the ATA-selected tests were 15.92, 14.73, and 15.91 for 0-1 LP, ACO, and Tabu search, respectively. Moreover, ACO selected 0.00 to 3.43 fewer "no DIF" items on average than 0-1 LP, and 0.00 to 3.39 fewer on average than Tabu Search. Across the simulation conditions, 0-1 LP and Tabu Search selected similar numbers of differentially functioning items on average.

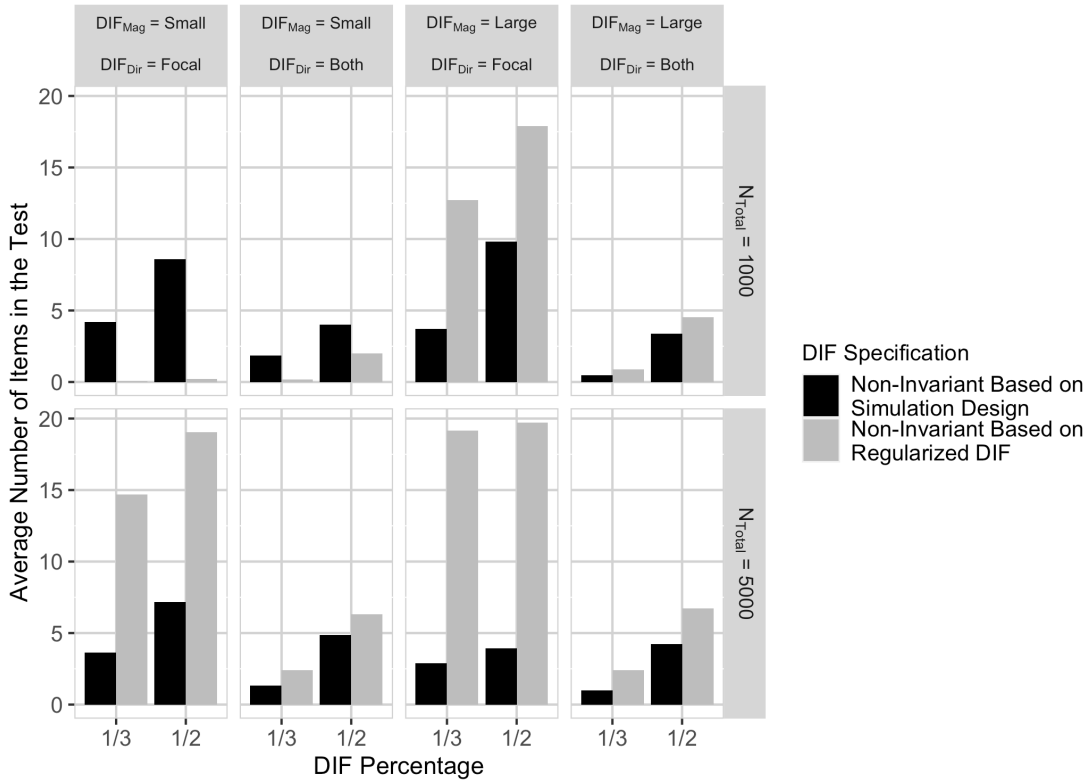
Item-Level MI with True DIF Categorizations. In the results presented thus far, the differentially functioning items in the ATA-selected tests were categorized based on regularized DIF and the wABC index. However, regularized DIF demonstrated exceedingly large FPRs, so many items categorized as "DIF" and selected by Unbiased-ATA were likely actually invariant. A small-scale simulation was constructed to examine whether Unbiased-ATA also selected higher numbers of truly invariant items in conditions where regularized DIF categorized most items as "DIF."

This simulation incorporated 64 conditions from Study 1 that were associated with both high regularized DIF $\overline{\text{FPR}}$ and ATA-selected tests with more "small DIF" and "large DIF" items. Specifically, the proof-of-concept simulation varied six design factors: (a) DIF type (uniform or non-uniform), (b) DIF percentage (moderate or large), (c) DIF magnitude (small or large), (d) DIF direction (focal or both), (e) sample size (moderate or large), and (f) estimation type (true or estimated parameters). Sample size balance was held constant with half of the total sample size in each group. Using the same item and person parameters from the 100 repetitions in the original simulation, Unbiased-ATA was re-run using Tabu search with five random starts. Only Tabu search was used because this algorithm performed similarly to 0-1 LP within a substantially shorter time frame. In

each repetition, the "best" test selected by Unbiased-ATA was evaluated using (a) the number of differentially functioning items as specified by the simulation design (DIF_{True}), and (b) the number of differentially functioning items as categorized by regularized DIF ($\text{DIF}_{\text{RegDIF}}$).

Notice in Figure 10 that in conditions with high $\overline{\text{FPR}}$ (e.g., large DIF magnitudes in the "focal group" direction, see Columns 3 – 4), there were again high proportions of $\text{DIF}_{\text{RegDIF}}$ items in the ATA-selected tests. Importantly, the proportion of DIF_{True} items in these tests was noticeably smaller. For example, when one-half of the item bank had large DIF in the "focal group" direction and $N = 5000$ (Column 3, Row 2), the average number of DIF_{True} items was less than five compared to nearly 20 for $\text{DIF}_{\text{RegDIF}}$ items.

Figure 10. Average Number of Items in the Selected Tests that were Truly Non-Invariant or Categorized as Non-Invariant by Regularized DIF

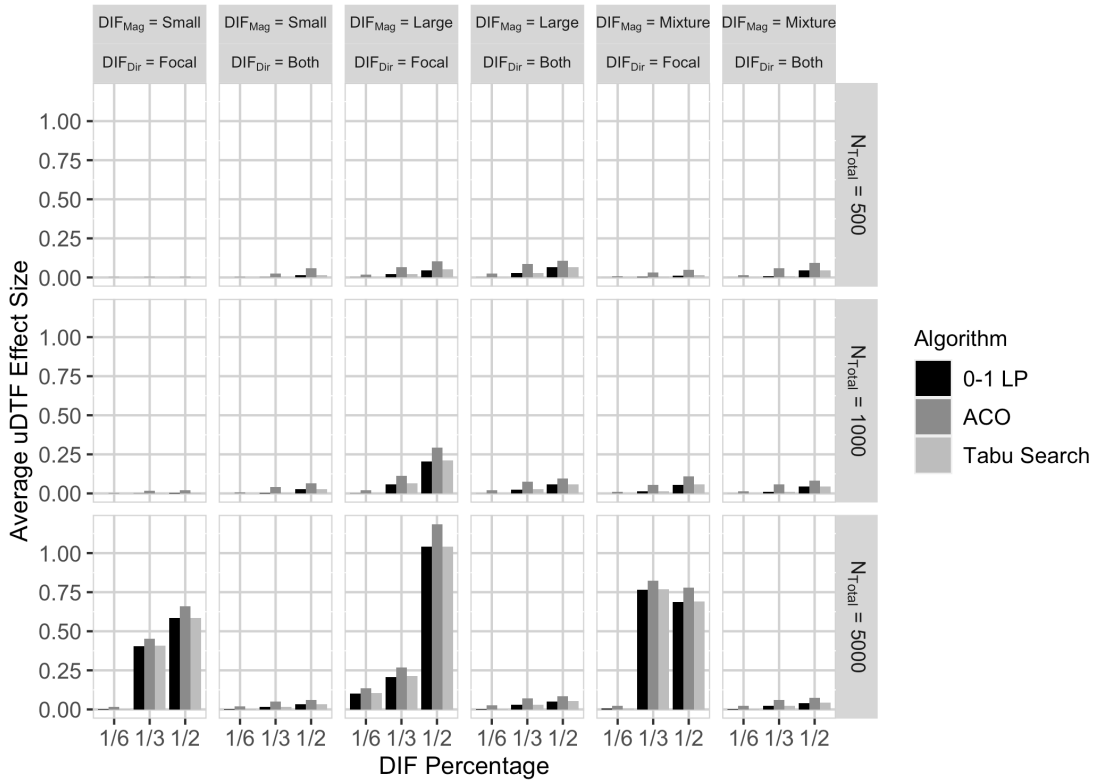


These results suggest that when the DIF detection method erroneously mislabeled items as differentially functioning, Unbiased-ATA could draw upon other item bank characteristics (e.g., information, item fit) during test construction. The average number of DIF_{True} items only exceeded the number of $\text{DIF}_{\text{RegDIF}}$ items when small DIF magnitudes were simulated with $N = 1000$ (Columns 1 – 2, Row 1). Looking back at Figure 2 and Figure 3, these conditions had relatively small false and true positive rates for regularized DIF, resulting in fewer $\text{DIF}_{\text{RegDIF}}$ items in the bank.

Test-Level MI. The uDTF index (Chalmers et al., 2016) was first used to evaluate test-level MI for the ATA-selected tests. The fixed- and mixed-effects models revealed that sample size, DIF percentage, DIF direction, and DIF magnitude had moderate to large effects on uDTF. Figure 11 presents the uDTF values for the selected tests when averaged across the simulation repetitions ($\overline{\text{uDTF}}$). Holding other factors constant, $\overline{\text{uDTF}}$ generally increased with larger sample sizes and when using ACO compared to 0-1 LP or Tabu search. $\overline{\text{uDTF}}$ was also positively associated with DIF percentage and magnitude. Moreover, when one-third or more of the item bank had DIF, there was a stronger relationship between sample size and $\overline{\text{uDTF}}$ in the "focal group" direction (Columns 1, 3, and 5) compared to the "both groups" direction (Columns 2, 4, and 6). This interaction resulted in substantially larger $\overline{\text{uDTF}}$ when DIF modifications were made only to the focal group with $N = 5000$ (Row 3).

Comparing Figure 11 to Figure 8, $\overline{\text{uDTF}}$ paralleled the average number of $\text{DIF}_{\text{RegDIF}}$ items in the selected test. More $\text{DIF}_{\text{RegDIF}}$ items resulted in fewer anchor items for the multiple-group parameter estimation. The reference and focal groups' item parameters then differed across more items and subsequently increased the uDTF index.

Figure 11. Average uDTF Effect Size for Selected Tests Across DIF Characteristics



Still, the maximum $\overline{\text{uDTF}}$ across all conditions was 1.36, translating to a "percent scoring difference for the overall test" of 6.8% (Chalmers et al., 2016, p. 120). Therefore, even when the ATA-selected tests were primarily comprised of $\text{DIF}_{\text{RegDIF}}$ items, the item-level DIF did not necessarily translate to large test-level differences.

Test-level MI was also evaluated by fitting the test data to successively constrained multiple-group IRT models to identify whether the data best aligned with a configural, weak, or strong MI model. Both full-sample RMSEA and group-level SRMSR values were calculated for each MI model (Maydeu-Olivares, 2015; Maydeu-Olivares & Joe, 2014) using the *mirt* R package (Chalmers, 2012). These fit statistics were then averaged across the simulation repetitions. The best-fitting model in each condition was selected as the most restrictive model for which all three averaged model

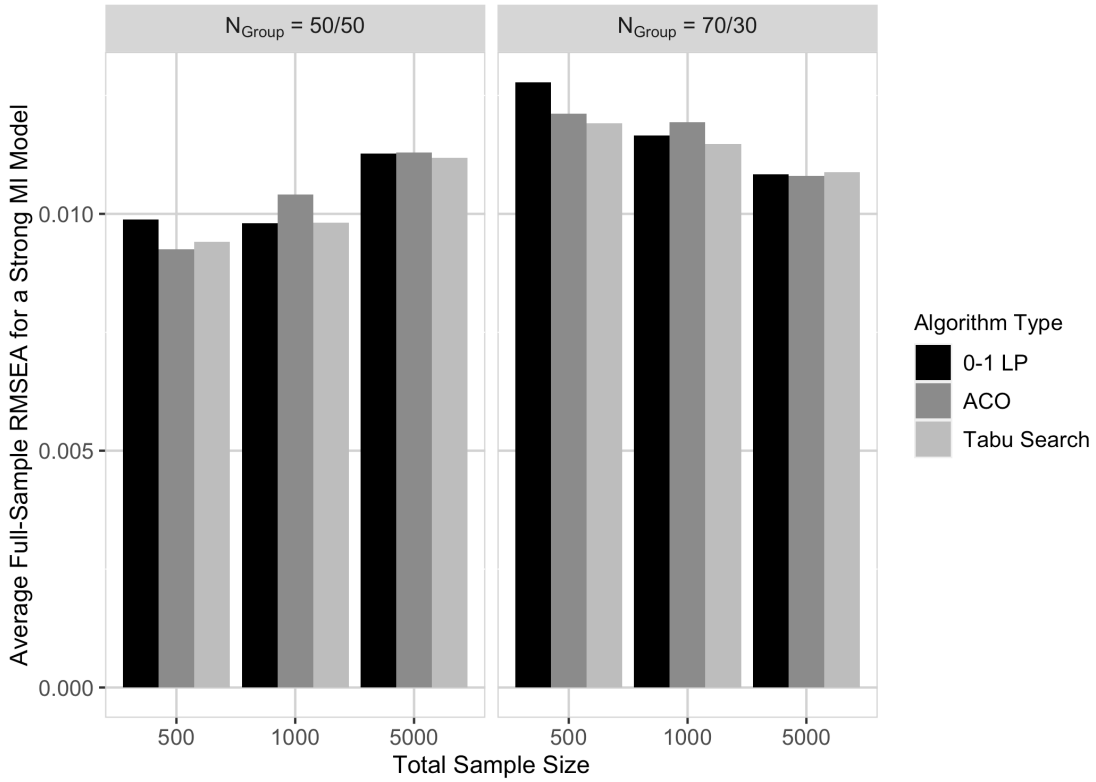
fit statistics indicated "good" fit based on thresholds for dichotomous IRT models, where RMSEA ≤ 0.05 and SRMSR ≤ 0.05 (Maydeu-Olivares, 2015; Maydeu-Olivares & Joe, 2014).

Using this procedure, test data most often displayed evidence of strong MI (indicating equivalent b and a parameters between groups). Specifically, the proportion of conditions that aligned with a strong MI level were 0.415, 0.392, and 0.421 for tests selected by 0-1 LP, ACO, and Tabu search, respectively. The corresponding proportions for weak MI (indicating equivalent b parameters) were 0.130, 0.126, and 0.127. Then, for configural MI, the proportions were 0.168, 0.194, and 0.162 for 0-1 LP, ACO, and Tabu search, respectively. Interestingly, across all simulation conditions, the average RMSEA values aligned with "good" fit for a strong MI model (i.e., RMSEA ≤ 0.05); model misfit was driven by higher group-level SRMSR values. For example, average SRMSR values ranged from 0.017 – 0.093, 0.014 – 0.074, and 0.013 – 0.061, when fitting the test data to strong, weak, and configural MI models, respectively. For simplicity, the present analysis focused on the model fit indices for the strong MI level (although corresponding results for weak and configural MI can be seen in Figures A2 – A5). The effect size tables indicated that DIF percentage, magnitude, and direction all influenced full-sample RMSEA and group-level SRMSR for the strong MI models. Algorithm type also affected RMSEA, whereas sample size balance played a role in SRMSR differences.

Looking first at full-sample RMSEA, the average fit statistics for conditions without simulated DIF were well below the threshold for "good" fit. Figure 12 shows that average RMSEA ($\overline{\text{RMSEA}}$) values ranged between approximately 0.009 and 0.013. Certain trends as a function of sample size balance and total sample size emerged in

Figure 12. For example, $\overline{\text{RMSEA}}$ and total sample size were positively related when the reference and focal groups had equal sample sizes (Column 1), but this trend reversed when 70% of the total sample was in the reference group (Column 2). Yet the differences in $\overline{\text{RMSEA}}$ among the conditions were small and did not affect the overall model fit.

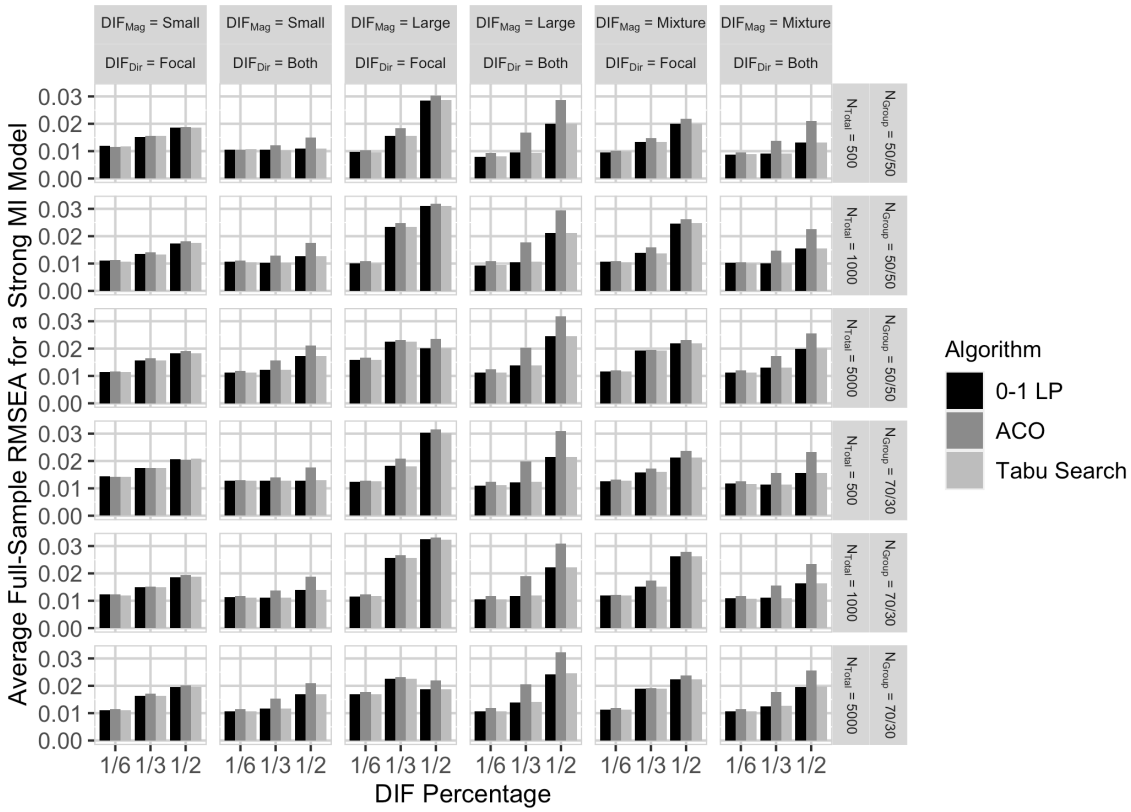
Figure 12. Average Full-Sample RMSEA for Strong MI Models in Conditions Without Simulated DIF



$\overline{\text{RMSEA}}$ was higher in conditions with simulated DIF (Figure 13) yet remained below 0.037. Here, $\overline{\text{RMSEA}}$ increased with higher DIF percentages, larger DIF magnitudes, and in the "focal group" direction. The largest $\overline{\text{RMSEA}}$ occurred when one-half of the item bank had large DIF in the "focal group" direction (Column 3). There was also some evidence that $\overline{\text{RMSEA}}$ was positively associated with total sample size, particularly with larger DIF percentages, but the differences in $\overline{\text{RMSEA}}$ were relatively small. Moreover, ACO sometimes produced tests with higher $\overline{\text{RMSEA}}$ than 0-1 LP or

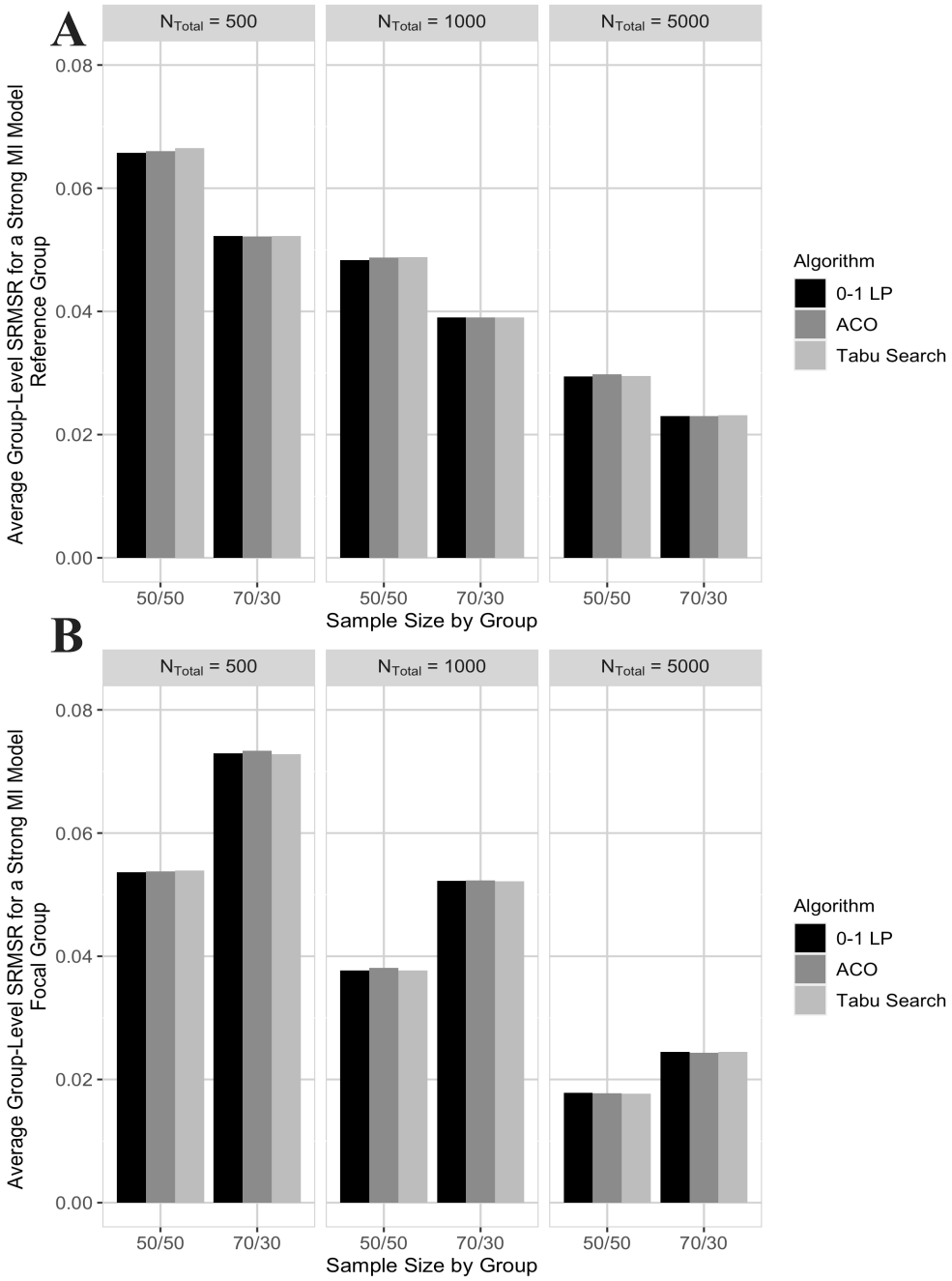
Tabu search. More noticeable differences in $\overline{\text{RMSEA}}$ among the algorithm types occurred in the "both groups" direction.

Figure 13. Average Full-Sample RMSEA for Strong MI Models in Conditions with Simulated DIF



Group-level SRMSR values were also computed for each ATA-selected test. Figure 14 shows the average SRMSRs for the reference group ($\overline{\text{SRMSR}}_{\text{R}}$; Panel A) and focal group ($\overline{\text{SRMSR}}_{\text{F}}$; Panel B) among simulation conditions without DIF. When the sample sizes were balanced, $\overline{\text{SRMSR}}_{\text{R}}$ was higher than $\overline{\text{SRMSR}}_{\text{F}}$. This trend reversed when the focal group comprised substantially fewer examinees than the reference group. For both groups, $\overline{\text{SRMSR}}$ decreased as sample size increased; in unbalanced samples, the group-level $\overline{\text{SRMSR}}$ became approximately equivalent when $N = 5000$ (Column 3).

Figure 14. Average Group-Level SRMSR for Strong MI Models in Conditions without Simulated DIF

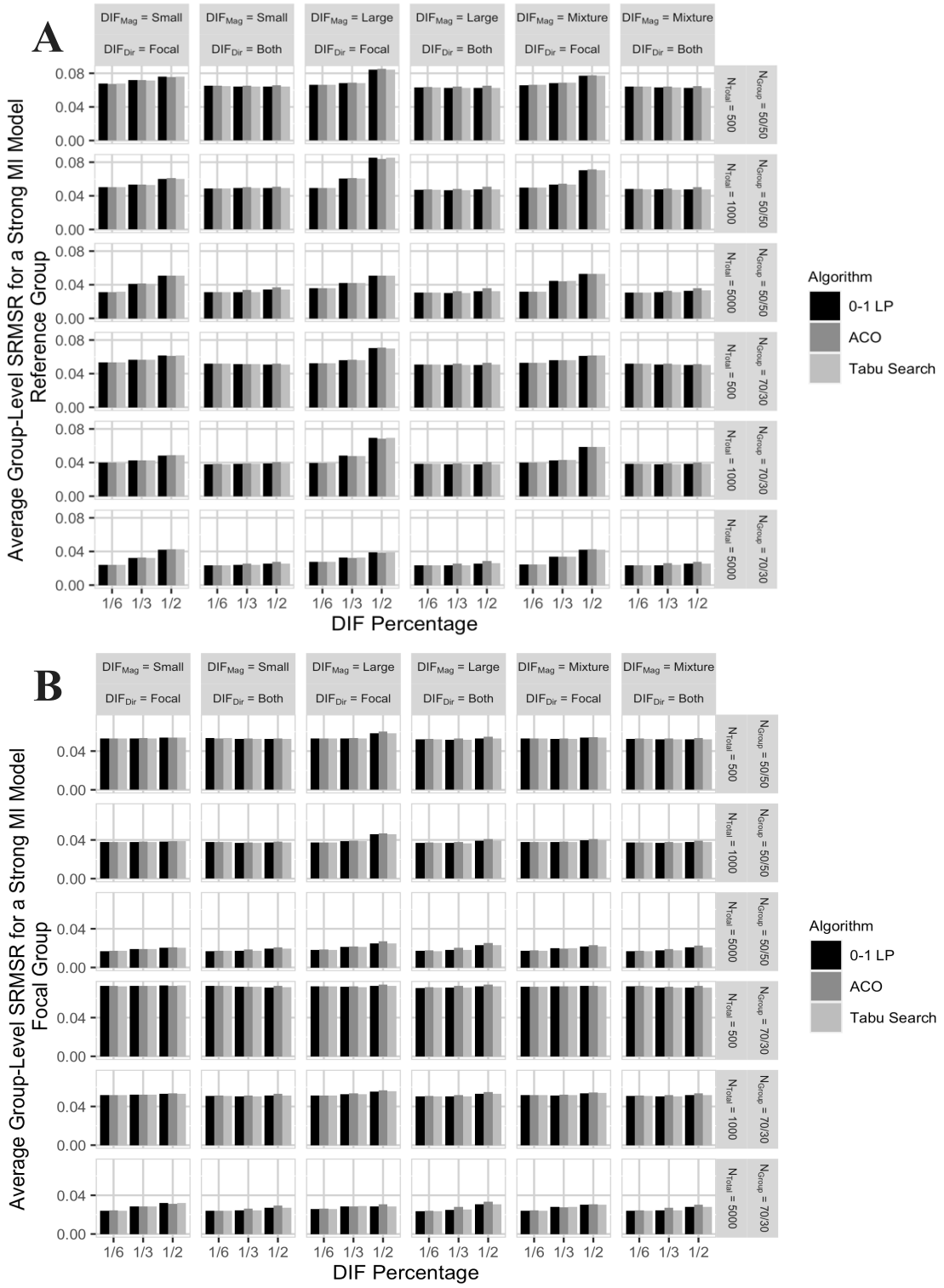


Moreover, $\overline{\text{SRMSR}}$ was more likely to remain below the threshold for "good" fit as total sample size increased.

The relationships among total sample size, sample size balance, and $\overline{\text{SRMSR}}$ generally replicated when DIF was simulated in the item banks. Figure 15 reveals that $\overline{\text{SRMSR}}$ uniformly increased as total sample size decreased. When $N \leq 1000$ and group-level sample sizes were unbalanced (Rows 4 – 5), $\overline{\text{SRMSR}}_{\text{F}}$ (Panel B) generally surpassed $\overline{\text{SRMSR}}_{\text{R}}$ (Panel A). Again, these differences between $\overline{\text{SRMSR}}_{\text{F}}$ and $\overline{\text{SRMSR}}_{\text{R}}$ dissipated as sample size increased for unbalanced samples. Yet when group-level sample sizes were balanced (Rows 1 – 3), $\overline{\text{SRMSR}}_{\text{R}}$ was generally higher than $\overline{\text{SRMSR}}_{\text{F}}$. Figure 15A also provides evidence of a small positive association between $\overline{\text{SRMSR}}_{\text{R}}$ and DIF percentage, especially in the "focal group" direction (Columns 1, 3, and 5).

Differences in $\overline{\text{SRMSR}}$ likely were a product of the group-level item parameter estimation. In scenarios where the ATA-selected test had more differentially functioning items, there were fewer anchor items when fitting the multiple-group IRT model. The differential parameter estimation appeared to produce larger differences between the observed and hypothesized measurement model for the reference group when sample sizes were balanced. Yet with unbalanced samples, smaller sample size for the focal group resulted in larger $\overline{\text{SRMSR}}$. It is important to emphasize that the variation in group-level $\overline{\text{SRMSR}}$ was relatively small across the examined conditions. Even when large DIF was simulated, many fitted models aligned with the strong MI level. As with the uDTF results, these model fit statistics suggest that evidence of measurement non-invariance at the item level was not strongly associated with evidence of measurement non-invariance at the test level.

Figure 15. Average Group-Level SRMSR for Strong MI Models in Conditions with Simulated DIF



As previously noted, the appendix includes $\overline{\text{RMSEA}}$ (Figures A2 – A3) and $\overline{\text{SRMSR}}$ (Figures A4 – A5) when fitting the test data to weak and configural MI models. Like when fitting a strong MI model, $\overline{\text{RMSEA}}$ values all aligned with "good" fit for the corresponding weak or configural models. Then, $\overline{\text{SRMSR}}$ values were more likely to be below the "good" fit threshold with less restrictive models (e.g., when fitting the test data to a configural versus weak MI model). Moreover, there was less variation in $\overline{\text{RMSEA}}$ and $\overline{\text{SRMSR}}$ across the other design factors (e.g., DIF percentage, algorithm type) when fitting the data to weak or configural MI models compared to a strong MI model. For the weak or configural MI models, total sample size most strongly influenced both $\overline{\text{RMSEA}}$ and $\overline{\text{SRMSR}}$, with average values decreasing as total sample size increased.

Test Score Precision. Test score precision for the ATA-selected tests was evaluated using the deviations between the selected test's TIF and a target TIF. Here, the target TIF was the information function for the full item bank. Smaller TIF deviations (Δ_{TIF}) were preferred, indicating closer alignment of the two functions. Figure 16 presents the average TIF deviations ($\bar{\Delta}_{\text{TIF}}$) for the simulation conditions without DIF. Considering regularized DIF's low FPRs in conditions without simulated DIF, the item parameter estimates were constrained to be equal in the subsequent multiple-group IRT models. These equivalent item parameters translated to equivalent TIFs between the reference and focal groups, and so Figure 16 presents the $\bar{\Delta}_{\text{TIF}}$ for both groups together. Notice in this figure that $\bar{\Delta}_{\text{TIF}}$ was uniformly higher for tests selected by ACO compared to 0-1 LP or Tabu search. Using estimated rather than true parameters also slightly increased $\bar{\Delta}_{\text{TIF}}$, although these differences never exceeded 0.14 in magnitude. There was

also a small increase in $\bar{\Delta}_{\text{TIF}}$ for $N = 500$ compared to $N \geq 1000$, particularly when using the ACO algorithm.

Figure 16. Average Test Information Function Deviations in Conditions without Simulated DIF

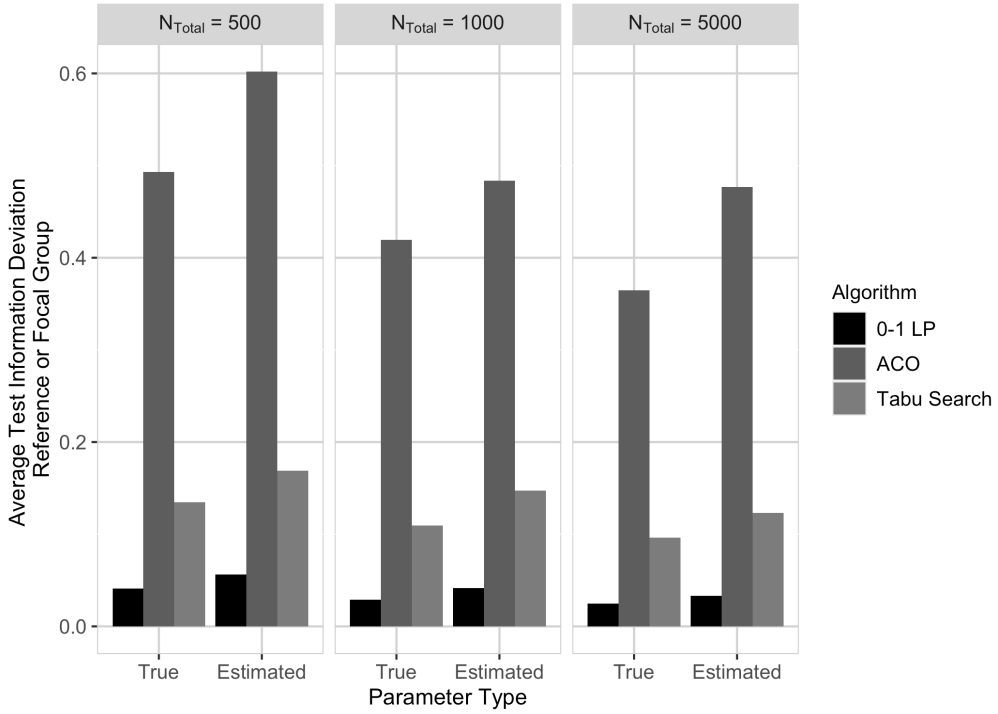
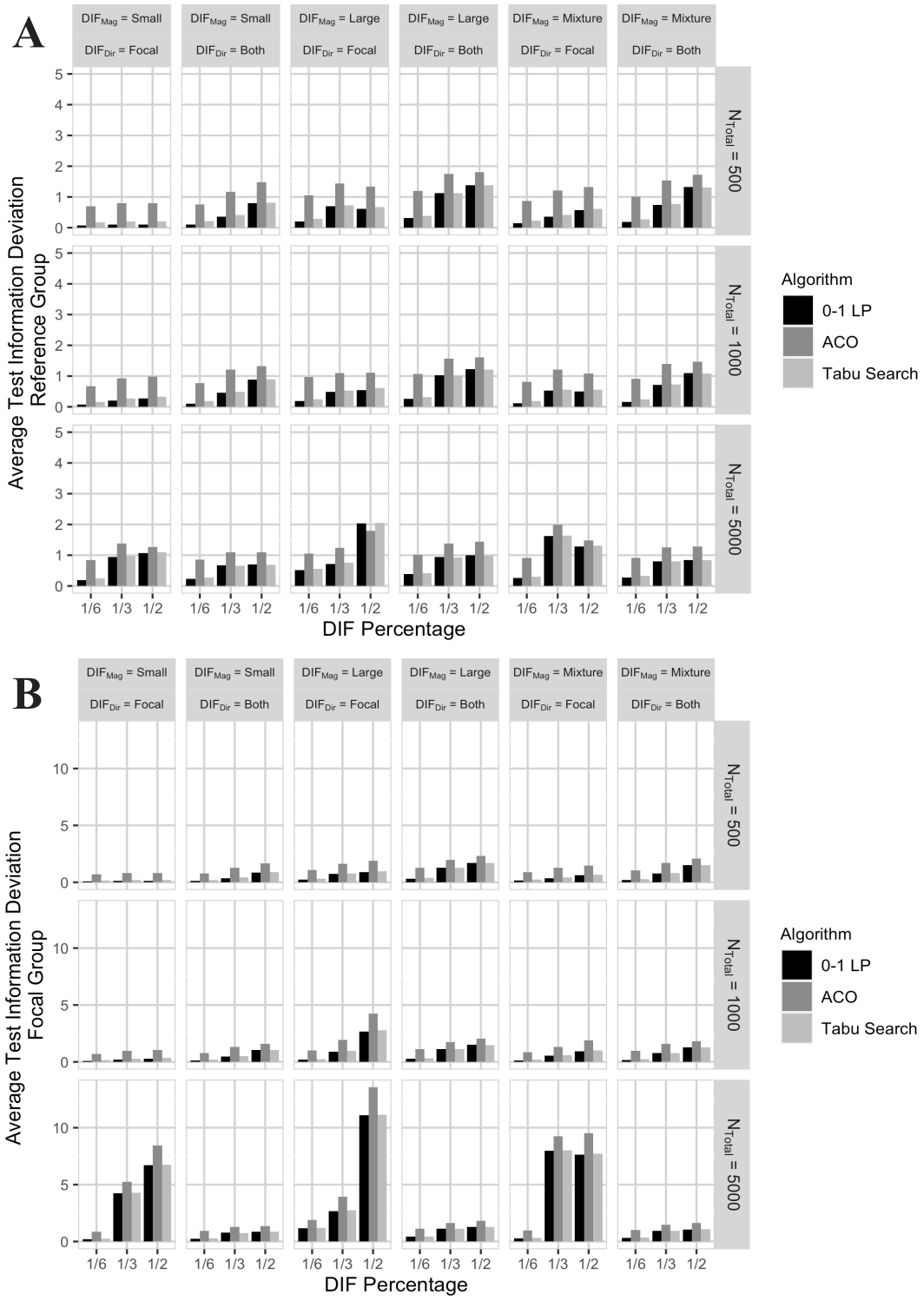


Figure 17 shows that $\bar{\Delta}_{\text{TIF}}$ was generally higher when the simulated item bank included differentially functioning items. For example, whereas $\bar{\Delta}_{\text{TIF}}$ never exceeded 0.63 in the "no DIF" conditions, the median $\bar{\Delta}_{\text{TIF}}$ for the reference and focal groups were 0.79 and 0.96, respectively, in the DIF conditions. There was also evidence in Figure 17 of small increases in $\bar{\Delta}_{\text{TIF}}$ with higher DIF percentages and magnitudes, and that ACO often selected tests with higher $\bar{\Delta}_{\text{TIF}}$ values than either 0-1 LP or Tabu Search. When looking between groups, $\bar{\Delta}_{\text{TIF}_{\text{F}}}$ was substantially higher than $\bar{\Delta}_{\text{TIF}_{\text{R}}}$ when $N = 5000$ and one-third or more of the bank had DIF in the "focal group" direction (Columns 1, 3, and 5, Row 3).

Figure 17. Average Test Information Function Deviations in Conditions with Simulated DIF



Comparing Figure 17 to Figure 8, higher $\bar{\Delta}_{\text{TIF}}$ was associated with more differentially functioning items in the ATA-selected tests, which in turn were associated with higher regularized DIF false and true positive rates. As previously noted, more differentially functioning items in the item bank translated to fewer anchor items for the multiple-group item parameter estimation. With fewer anchor items, the MML estimation resulted in different item parameter estimates between the groups. For the focal group, these estimated parameters appeared to produce either smaller or larger test information values at certain θ values and subsequently larger $\bar{\Delta}_{\text{TIF}}$.

The total TIF values (computed as the sum of the item information values at each θ value) were also evaluated for each ATA-selected test. As shown in Figure A6 in the Appendix, there were fewer noticeable trends in average TIF values ($\overline{\text{TIF}}$) across the various design factors. In cases of non-uniform DIF, $\overline{\text{TIF}}$ in the "both groups" direction was slightly higher than that in the "focal group" direction. In the "both groups" direction, there was also a small but discernable positive relationship between DIF percentage and $\overline{\text{TIF}}$. Still, the range of $\overline{\text{TIF}}$ across all examined conditions was relatively small (i.e., $57.05 \leq \overline{\text{TIF}}_{\text{R}} \leq 64.51$; $55.09 \leq \overline{\text{TIF}}_{\text{F}} \leq 64.39$). Interestingly, although the focal group demonstrated higher $\bar{\Delta}_{\text{TIF}}$ than the reference group, full test $\overline{\text{TIF}}$ was not noticeably different between groups. This result suggests that the item parameters were not necessarily uniformly more or less informative for the focal group, but rather more or less informative at different locations along the θ continuum.

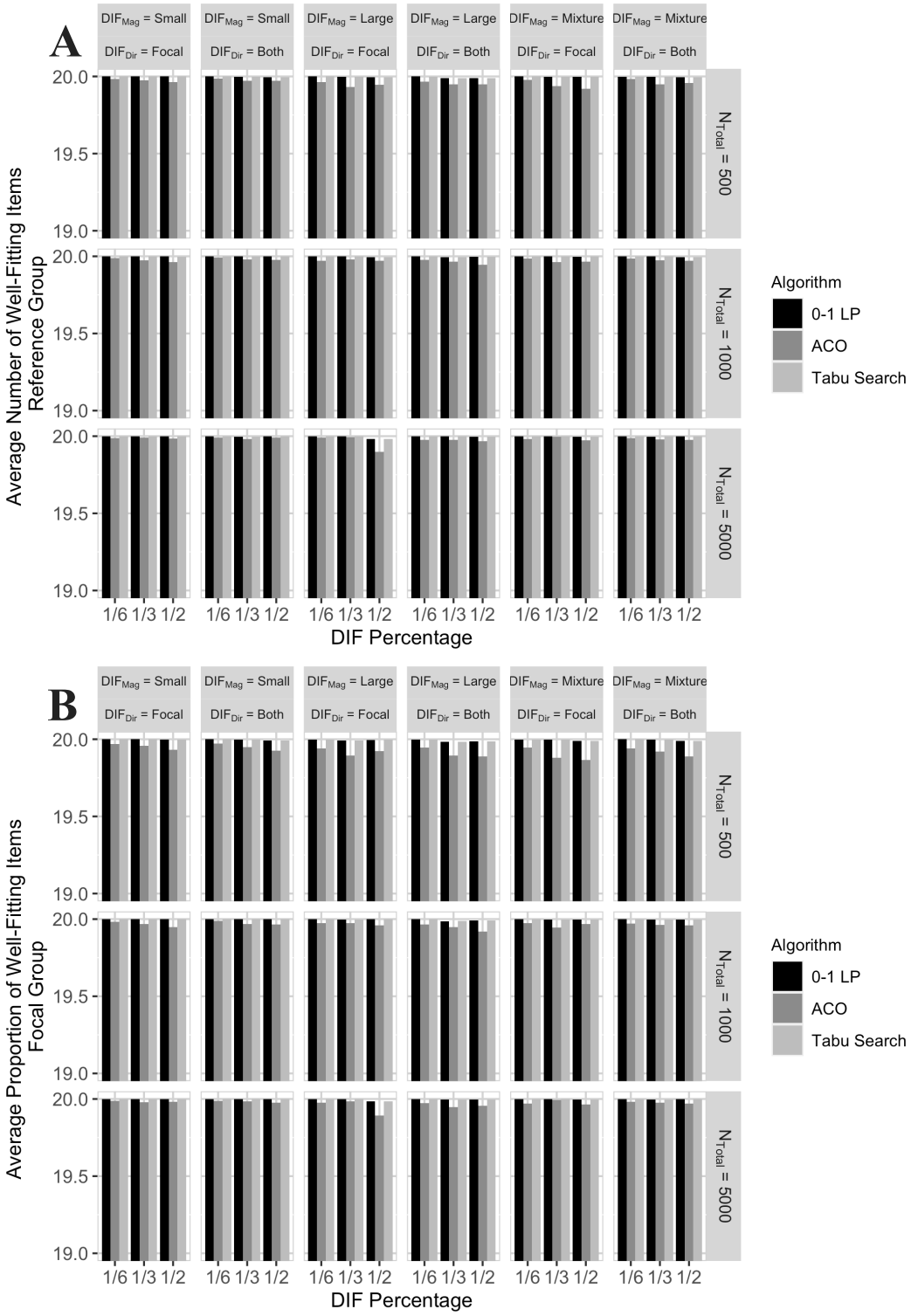
Figure A6 also reveals differences in test information when using the true versus estimated parameters for item response generation. Holding other factors constant, $\overline{\text{TIF}}_{\text{R}}$ was 58.59 and $\overline{\text{TIF}}_{\text{F}}$ was 58.44 when using the true parameters (Rows 1 – 3). Then $\overline{\text{TIF}}_{\text{R}}$

and $\overline{\text{TIF}_\text{F}}$ was 61.47 and 61.30, respectively, when using estimated parameters (Rows 4 – 6). Yet parameter type (true or estimated) did not noticeably affect $\bar{\Delta}_{\text{TIF}}$. Given that the initial parameter estimation produced more informative items in the bank (e.g., with overestimated discrimination parameters, see Figure 6), higher information increased both the TIF of the ATA-selected tests and the target TIF. With simultaneous increases in both TIFs, the $\bar{\Delta}_{\text{TIF}}$ was less noticeably affected.

Item Fit. In the Unbiased-ATA procedure, items were categorized as well-fitting or misfitting using the $S - \chi^2$ fit statistic (Orlando & Thissen, 2000) with $\alpha = 0.10$. Only sample size and algorithm type demonstrated notable effects on the number of well-fitting items in the ATA-selected tests (see Tables 3 – 5). Yet Figure 18 demonstrates largely negligible differences in the average numbers of well-fitting items across the design factors. The average number of well-fitting items was consistently high, ranging between 19.74 and 20.00 for the reference group and between 19.80 and 20.00 for the focal group. Tests tended to contain slightly more misfitting items when there was large DIF for 1/2 of the item bank in the "focal group" direction with $N = 5000$ (Column 3, Row 3); this condition had the highest regularized DIF FPRs. There was also greater variation in proportions among tests selected by ACO compared to tests selected by 0-1 LP or Tabu search. Regardless, the magnitude of differences between ACO and either 0-1 LP or Tabu search remained below 0.21.

External Validity. The final psychometric property used to evaluate the ATA-selected tests was the correlation between $\hat{\theta}$ from the "best" test and $\hat{\theta}$ from an external criterion measure. Few design factors demonstrated moderate or large η_p^2 for the external validity coefficients: whereas DIF direction influenced the correlations for the reference

Figure 18. Average Number of Well-Fitting Items within Selected Tests in Conditions with Simulated DIF

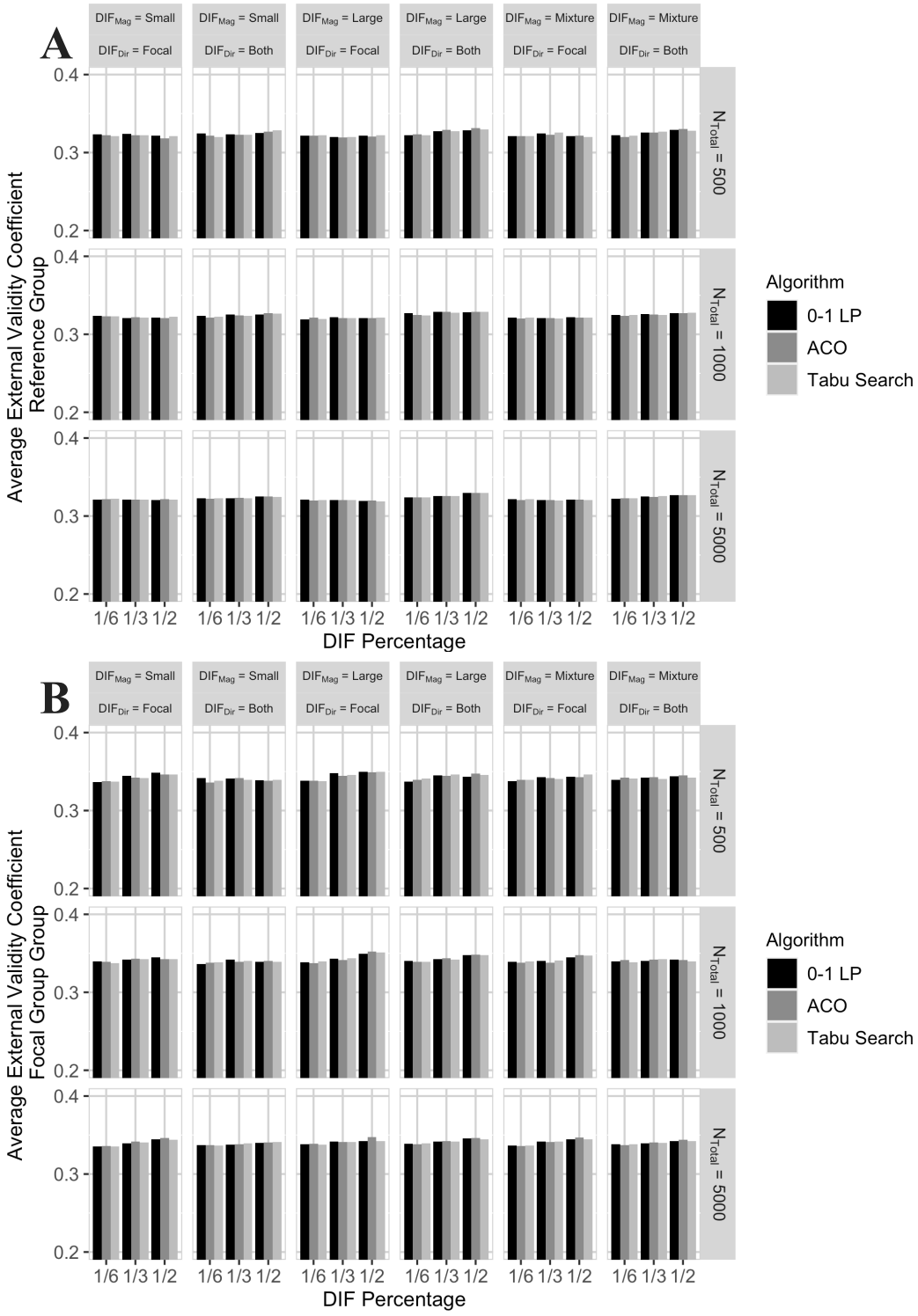


group, DIF percentage influenced the correlations for the focal group (see Tables 3 – 5, Columns 11 – 12). Note that these models had considerably low R^2 values, and so inferences regarding effect size values should be interpreted cautiously.

Figure 19 demonstrates few substantial differences in external validity coefficients across the examined testing scenarios. ATA-selected tests tended to have stronger correlations with the external criterion measure in cases with larger DIF magnitudes and percentages, as well as in the "focal group" direction. Average correlations were also often larger for the focal group (Panel B) compared to the reference group (Panel A). Differences in parameter estimates between groups (due to more differentially functioning items in the test) can differentially affect θ estimation, in turn affecting correlations with the external criterion $\hat{\theta}$. Figure 19 provides evidence that the focal group parameter estimates produced $\hat{\theta}$ with slightly stronger linear relationships with the external criterion $\hat{\theta}$ than those for the reference group. Importantly, when differences arose, the magnitude of differences among the average correlations were small, with ranges often less than 0.025.

Comparison to "Worst-Case Scenarios." The results thus far have shown that even when Unbiased-ATA selected tests with more differentially functioning items, the other psychometric properties of the selected tests were often relatively strong. For example, tests still demonstrated evidence of strong test-level MI and high proportions of well-fitting items. To contextualize the psychometric strength of the ATA-selected tests, it is helpful to compare these tests to how poorly an item combination might perform on the examined test characteristics. In other words, did Unbiased-ATA tend to select "good" item combinations when "worse" item combinations were possible? Or were item

Figure 19. Average Correlations between Selected Tests and an External Criterion in Conditions with Simulated DIF

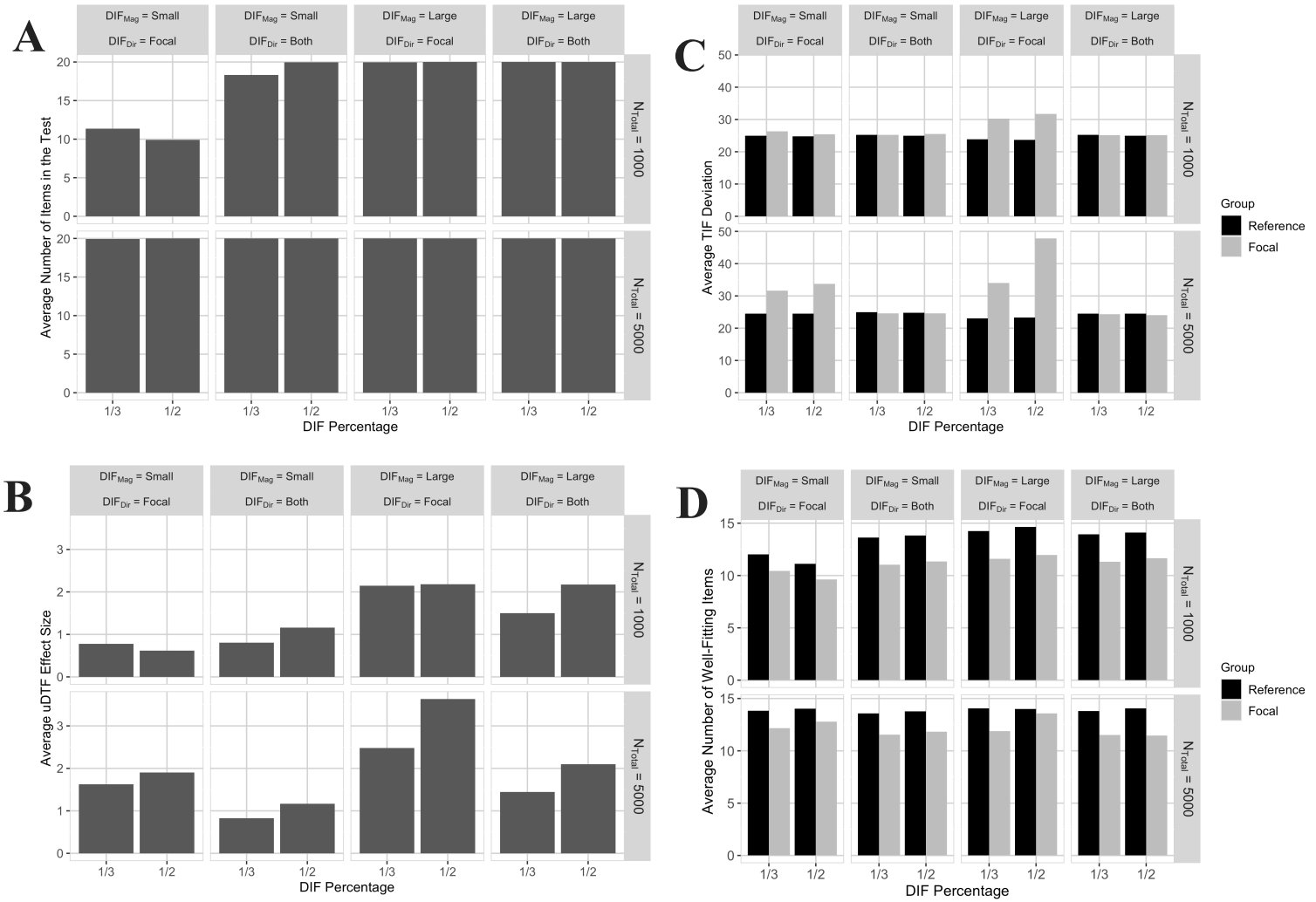


banks constructed such that most, if not all, item combinations would demonstrate similarly strong psychometric properties?

To gain insight into these questions, another proof-of-concept simulation was performed using the same 64 conditions as the previous simulation for DIF_{True} items. Again, each of the conditions used the 100 corresponding item banks and θ parameters that were generated in the original large-scale simulation. Four ATA algorithms were run based on the four criteria of the Unbiased-ATA objective function. These criteria were reversed such that each ATA algorithm was selecting item combinations that maximized either (a) the number of differentially functioning items, (b) the uDTF effect size, (c) the TIF deviations for the reference and focal groups, or (d) the number of misfitting items. Four separate ATAs were used to find the lower bound of the psychometric properties independent of any other criterion. The ATAs were again conducted using Tabu search with five random starts. Figure 20 presents the results of this follow-up simulation for the average number of differentially functioning items selected (Panel A), uDTF effect size (Panel B), TIF deviations (Panel C), and number of well-fitting items (Panel D).

Taken together, the results in Figure 20 indicate that item combinations with substantially weaker psychometric properties were possible in the examined item banks. For example, the average ATA-selected test comprised close to 20 differentially functioning items (as denoted by regularized DIF) in most conditions (see Panel A). Additionally, $\overline{\text{uDTF}}$ ranged from approximately 0.56 to 3.89 with an interquartile range of 1.05 to 2.17 (see Panel B). In comparison, $\overline{\text{uDTF}}$ from tests selected by Unbiased-ATA

Figure 20. Psychometric Characteristics for Tests Selected using Algorithms that Reverse the Unbiased-ATA Objective Function



were strongly right-skewed with a maximum of 1.36. Similar trends were evident for $\bar{\Delta}_{\mathrm{TIF}}$ and the proportion of well-fitting items, with average values substantially different from those for Unbiased-ATA tests. For instance, $\bar{\Delta}_{\mathrm{TIF}}$ in Figure 20C was consistently higher than the maximum $\bar{\Delta}_{\mathrm{TIF}}$ in a test selected by Unbiased-ATA ($\bar{\Delta}_{\mathrm{TIF}_F} = 15.27$). Overall, this follow-up simulation provided evidence that Unbiased-ATA can select item combinations with a balance of desirable psychometric properties from item banks that have the possibility for weaker test characteristics. In other words, it was not necessarily the case that the Study 1 item banks were constructed in such a way that most item combinations would show strong psychometric properties.

Unbiased-ATA Performance using IRT-LRT

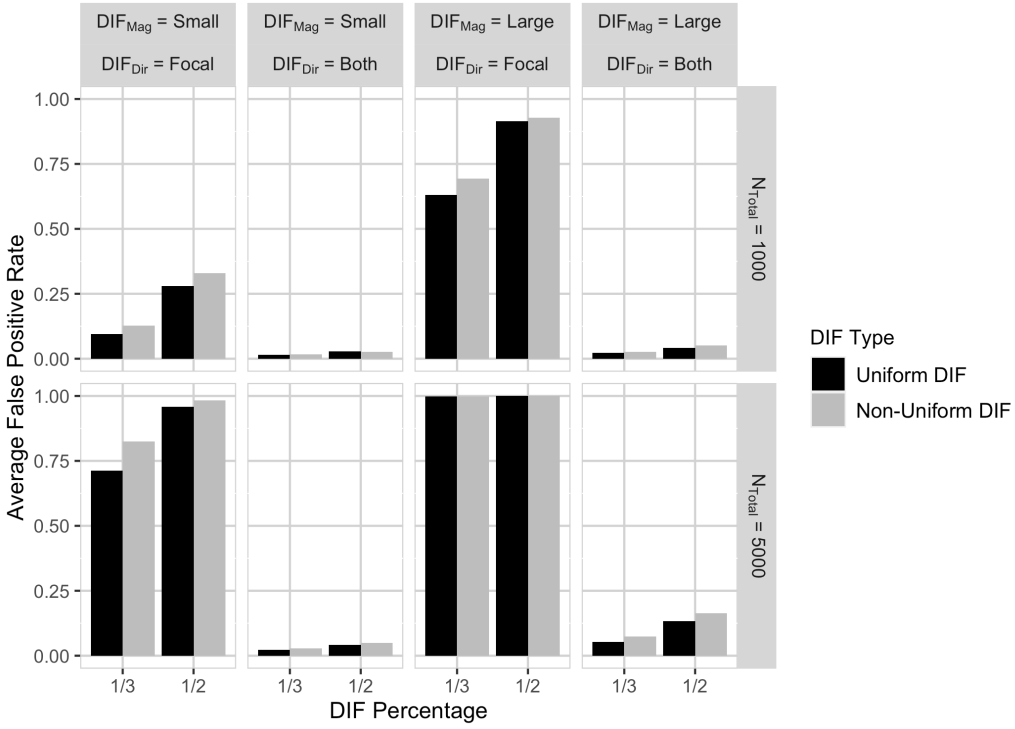
In numerous testing scenarios, regularized DIF demonstrated very high FPRs that would be considered unacceptable in practice. These FPRs were associated with ATA-selected tests with relatively weaker psychometric properties. To better understand the correspondence between the DIF detection method and the characteristics of the "best" test selected by Unbiased-ATA, a follow-up simulation replacing regularized DIF with the IRT-LRT method (e.g., Thissen et al., 1988, 1993) was designed. IRT-LRT is commonly used for DIF identification in psychological and educational measurement (see Teresi et al., 2021).

This simulation used the same 64 conditions as the two proof-of-concept simulations presented above. Recall that these conditions were selected because they revealed unreasonably high FPRs with regularized DIF. Here, all item and person parameters were randomly regenerated, and the procedure exactly followed the large-scale Study 1 simulation (including using all three algorithm types) albeit for the DIF

detection method. A backward IRT-LRT method was implemented. Specifically, for each item, a model with all item parameters constrained to be equal was compared to a model with only that item's parameters freely estimated. The models were then compared using a likelihood ratio test statistic. This procedure has also been termed an "all-other" approach (Bolt, 2002; S.-H. Kim & Cohen, 1998; as cited in Teresi et al., 2021). Following Teresi et al. (2021), the Benjamini-Hochberg adjustment was applied to control the false discovery rate for multiple-testing. Note that this application of the IRT-LRT method did not assume prior knowledge of anchor items (for a cogent discussion of anchor item selection in DIF methods, see Teresi et al., 2021).

Figure 21 presents the $\overline{\text{FPR}}$ values when categorizing differentially functioning items using IRT-LRT. DIF direction again played a notable role in $\overline{\text{FPR}}$. Specifically, IRT-LRT more often mistakenly identified items as differentially functioning in the "focal group" direction compared to the "both groups" direction. $\overline{\text{FPR}}$ often exceeded 0.50 in the "focal group" direction, particularly as sample size, DIF magnitude, and DIF percentage increased. Yet in the "both groups" direction, IRT-LRT maintained $\overline{\text{FPR}}$ values closer to the nominal rate ($\alpha = 0.05$) even for large samples and DIF magnitudes. Certain differences in performance were apparent between regularized DIF and IRT-LRT. For example, IRT-LRT's $\overline{\text{FPR}}$ were slightly smaller than regularized DIF in most "focal group" direction conditions (Columns 1 and 3) and demonstrably smaller in the "both groups" direction (Columns 2 and 4). However, the exceedingly large $\overline{\text{FPR}}$ in numerous conditions were consistent across DIF detection methods.

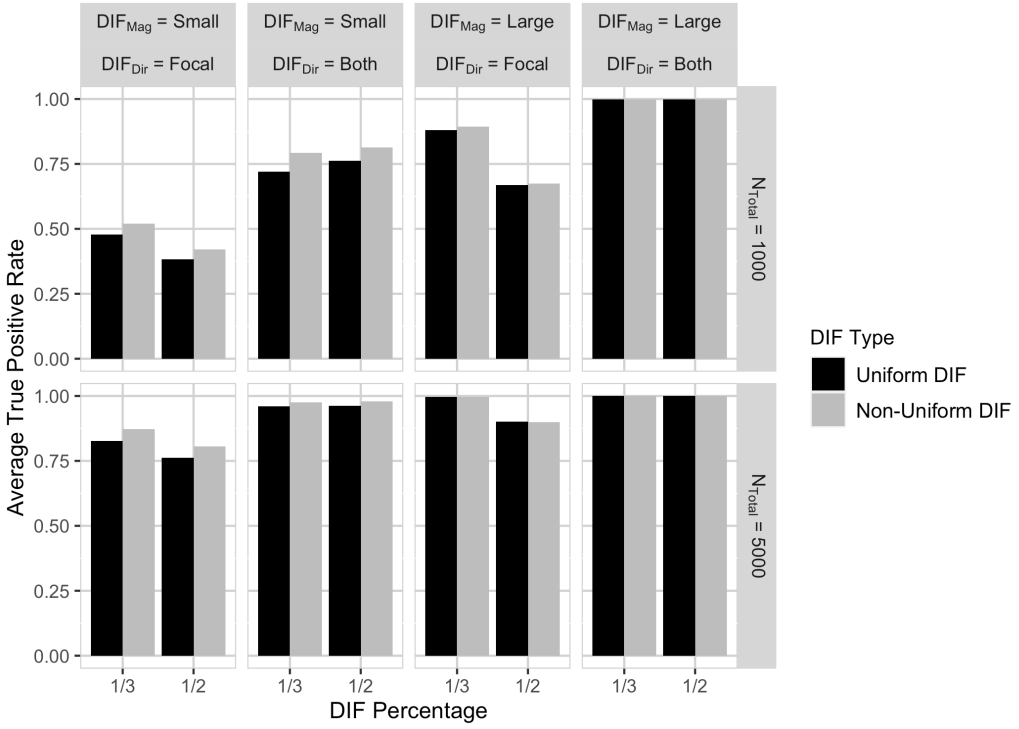
Figure 21. Average False Positive Rates for IRT Likelihood Ratio Test



The corresponding $\overline{\text{TPR}}$ values for IRT-LRT are shown in Figure 22. Notice that $\overline{\text{TPR}}$ increased with larger sample sizes and DIF magnitudes. When DIF modifications were applied to both groups (Columns 2 and 4), IRT-LRT had similar $\overline{\text{TPR}}$ to regularized DIF (see Figure 3) even with substantially smaller $\overline{\text{FPR}}$. However, when DIF modifications were applied only to the focal group (Columns 1 and 3), IRT-LRT demonstrated slightly smaller $\overline{\text{TPR}}$ than regularized DIF (the one exception being when small DIF was applied to the focal group with $N = 1000$).

In the appendix, Figures A7 – A12 show characteristics of the "best" tests selected by Unbiased-ATA when paired with IRT-LRT. Taken together, these figures reveal very similar trends in the ATA-selected tests' psychometric properties as compared to those with regularized DIF (as shown in Figures 7, 8, 11, 13, 15, and 17).

Figure 22. Average True Positive Rates for IRT Likelihood Ratio Test



Despite slight differences in average values whether using regularized DIF or IRT-LRT, the result patterns remained largely consistent such that weaker psychometric properties aligned with higher IRT-LRT false and true positive rates. For example, the ATA-selected tests comprised more differentially functioning items in the "focal group" DIF direction with $N = 5000$. There were also more differentially functioning items when large DIF modifications were made to one-half of the item bank with $N = 1000$. These conditions were then associated with higher strong MI model fit statistics.

Interestingly, $\overline{\text{uDTF}}$ (Figure A9) and $\bar{\Delta}_{\text{TIF}_\text{F}}$ (Figure A12) were substantially elevated only when there was large DIF for one-third of the items in the "focal group" direction with $N = 5000$. Other testing scenarios with high IRT-LRT false and true positive rates were not associated with inflated $\overline{\text{uDTF}}$ or $\bar{\Delta}_{\text{TIF}}$. Looking at the average estimated item parameters in the selected tests (Figure A7), distributions of group-level $\hat{b}$

were roughly equivalent in all but one scenario (the condition with larger $\overline{\text{uDTF}}$ and $\bar{\Delta}_{\text{TIF}_\text{F}}$). This result suggests that in many cases where IRT-LRT mistakenly identified invariant items as differentially functioning, Unbiased-ATA selected items with smaller average differences in group-level parameter estimates. In contrast, high FPRs and TPRs with regularized DIF more consistently corresponded to noticeable group differences in average $\hat{b}$ among the selected items.

Discussion

Unbiased-ATA uses an IRT framework to systematically evaluate and reduce test bias among subgroups during ATA. Study 1 was designed to identify test scenarios wherein Unbiased-ATA could construct 20-item tests with evidence of relatively stronger psychometric properties. Here, psychometric strength was operationalized by (a) fewer differentially functioning items, (b) stronger evidence of full-test MI, (c) smaller deviations between the test's information function and a target TIF, (d) more well-fitting items, and (e) higher external validity coefficients.

The Study 1 simulations revealed that the psychometric strength of the ATA-selected tests largely depended upon the item bank characteristics. Given the way the Unbiased-ATA method was defined, when the DIF detection method categorized more items in the bank as differentially functioning, there were fewer anchor items for estimating the multiple-group IRT model. Fewer anchors translated to more items with differing estimated parameters across groups. Given that Unbiased-ATA draws heavily on group-specific indices, differing parameters directly affected the uDTF index, deviations between the select test's TIF and a target TIF, and the number of well-fitting items. The IRT-LRT analyses highlighted that fewer anchor items did not always result in

larger group-level parameter differences. Thus, weaker psychometric characteristics of the ATA-selected tests occurred when the bank contained both more items categorized as differentially functioning, and there were larger differences in the $\hat{b}$ and $\hat{a}$ parameters between the reference and focal groups.

Differences in test characteristics across the design factors generally aligned with differences in the DIF detection method's false and true positive rates. Specifically, ATA-selected tests showed stronger psychometric properties in testing scenarios with (a) smaller DIF percentages, (b) smaller DIF magnitudes, and (c) when DIF modifications were made evenly to both groups. Moreover, there was evidence of better performance in terms of DIF identification, uDTF, and TIF deviations with smaller total sample sizes. Alternatively, larger sample sizes were associated with smaller strong MI model fit statistics. Algorithm type also played a role, such that ATA-selected tests demonstrated better properties pairing the Unbiased-ATA objective function with either 0-1 LP or Tabu search instead of ACO.

In the current study, regularized DIF produced substantially inflated FPRs. These FPRs were often paired with high TPRs, such that most (if not all) of the items in the bank were categorized as differentially functioning. The FPRs reported here exceeded those reported in Belzak and Bauer's (2020) original examination of regularized DIF. For instance, the highest FPR that Belzak and Bauer (2020) reported was 0.33, occurring when large DIF was simulated in six out of 12 items with $N = 500$. The stark differences in FPRs between this study and Belzak and Bauer (2020) are likely due to differences in simulation design. Specifically, the current study used substantially larger item banks and sample sizes, such as $N = \{500, 1000, 5000\}$ compared to $N = \{250, 500, 1000\}$ in

Belzak and Bauer (2020). The results presented here provide evidence that regularized DIF's ability to accurately identify invariant items declines in larger item banks with high DIF percentages, particularly when paired with total sample sizes greater than 1,000.

IRT-LRT also demonstrated high FPRs in many of the same conditions as regularized DIF. The replication of results across DIF detection methods suggests that the large FPRs were likely a product of the item bank characteristics rather than the detection method. It is plausible that the DIF scenarios simulated in Study 1 were too extreme (e.g., DIF percentages were too high) to expect good performance from any detection method. Empirical studies have identified item banks with percentages of differentially functioning items ranging from 0% (Marc et al., 2008) to over 65% (F. Y. Huang et al., 2006; Sheppard et al., 2006). Considering studies with item banks comprising 60 or more items, reported percentages include 3% (Pauwels et al., 2014), 7.1% (Van den Broeck et al., 2012), and 38% (C. D. Huang et al., 1997; Sheppard et al., 2006). The DIF percentages of 33% and 50% used in Study 1 were therefore on the higher end of the amount of differentially functioning items seen in practice. Studies 2 and 3 will examine whether the relationship between FPRs and Unbiased-ATA holds among item banks with fewer differentially functioning items.

Another notable trend that emerged for the DIF detection methods was the relationship between DIF direction and FPRs. Specifically, FPRs were substantially higher for both regularized DIF and IRT-LRT in the "focal group" direction compared to the "both group" direction. This result corroborates previous DIF research. For example, Kopf et al. (2015) summarized multiple studies that used the "all-other" anchor item approach (as was used with IRT-LRT in the current study), saying that "methods from the

all-other and the *equal-mean difficulty anchor class* displayed seriously inflated false alarm rates when the direction of DIF was unbalanced (i.e., the DIF effects did not cancel out between groups and one group was favored in the test)" (pp. 27 – 28). Furthermore, DeMars (2020) suggested that a "DIF item favoring [one group] to a large degree" is more easily identified as differentially functioning than "several items favoring [another group] to a smaller degree" (p. 68). Although Belzak and Bauer (2020) did not examine DIF direction, the Study 1 findings indicate that larger DIF magnitudes in one direction are positively associated with FPRs for regularized DIF.

Even though high numbers of differentially functioning items were identified in the bank and, in some conditions, selected by Unbiased-ATA, item-level violations of MI did not necessarily translate to large test-level violations of MI. For instance, average uDTF effect sizes were smaller than 2.0 (or a score difference of 10% between groups; Chalmers et al., 2016). Additionally, average RMSEA values for most ATA-selected tests aligned with the strong MI level (indicating evidence for equal b and a parameters between groups). There were also negligible differences in external validity coefficients and the number of well-fitting items across the conditions. High proportions of “DIF” items therefore did not always degrade other psychometric properties of the ATA-selected test. Rather, the effect of these “DIF” items on the test characteristics depended on the extent to which estimated item parameters differed between groups.

Follow-up simulations provided further evidence of Unbiased-ATA’s ability to select items with a desirable balance of psychometric properties. In cases where a high proportion of items in the ATA-selected test were categorized as differentially functioning by regularized DIF, a noticeably smaller proportion were truly non-invariant.

Unbiased-ATA therefore could select among the best-performing $\text{DIF}_{\text{RegDIF}}$ items using other indices of test score accuracy and precision; by incorporating other item characteristics, Unbiased-ATA might better differentiate between false and true positive $\text{DIF}_{\text{RegDIF}}$ items. Applying Unbiased-ATA in the test construction procedure can thus reduce the number of differentially functioning items that test developers need to review. The current study suggests that test developers might find upon their review that many of these items have small or negligible practical DIF magnitudes.

Furthermore, Unbiased-ATA selected tests with a balance of the predefined psychometric characteristics even when item combinations with weaker properties were plausible. For instance, tests with percent scoring differences of 15%, Δ_{TIF} greater than 25, or proportions of well-fitting items below 0.75 were possible within the examined item banks. Using a method like Unbiased-ATA highlights the advantages of combining different psychometric criteria within the objective function. As an example, consider an ATA objective function that exclusively minimizes Δ_{TIF}. Applying that ATA algorithm to these item banks could produce an item combination with minimal Δ_{TIF}, but high item-level MI or few well-fitting items. Study 3 will provide further insight into this supposition by directly comparing Unbiased-ATA to other objective functions.

The Study 1 results also revealed relatively small differences in the "best" test characteristics when pairing Unbiased-ATA with either 0-1 LP, Tabu search, or ACO. Algorithm type more strongly influenced $\bar{\Delta}_{\text{TIF}}$ and the number of invariant items, with ACO generally selecting weaker item combinations than either 0-1 LP or Tabu search. Both 0-1 LP and Tabu search performed similarly across Study 1 in terms of the ATA-selected test characteristics. However, 0-1 LP took substantially longer to find a solution,

and was not always able to find an optimal solution within the given time frame (i.e., five minutes). It is possible that 0-1 LP demonstrated performance difficulties due to the relatively small item bank size. Additionally, a longer time limit would likely increase the likelihood of selecting an optimal solution. Yet interestingly, the proportion of optimal solutions was noticeably smaller in conditions without simulated DIF. It is probable that in these conditions, the differences among the possible item combinations were small. Then, 0-1 LP had more difficulty comparing across the possible objective function values to identify the "best" solution.

To support this claim, one "no DIF" replication was re-run and the 0-1 LP solver's progress was observed. After five minutes, the solver (using branch-and-bound) still had over 50,000 nodes to compare with less than 0.1% differences in the objective function at each iteration. In other words, the solver was comparing among various item combinations with small differences in objective function criteria. Furthermore, there was a non-linear relationship between the proportion of "optimal" solutions and the variance of the selected objective function values across all conditions (i.e., both conditions with and without simulated DIF). Specifically, conditions with fewer "optimal" solutions had miniscule spread of objective function values. When the proportion of "optimal" solutions was large, the variance of the objective function values exponentially increased. Granted, the variance of the objective function values (computed across simulation repetitions) does not directly equate to having many item combinations within a single bank with nearly equivalent psychometric properties. However, there is evidence that 0-1 LP had more difficulty finding an optimal solution in conditions where the spread of selected objective function values was small.

In summary, Tabu search emerged as the most effective ATA algorithm in Study 1. This metaheuristic algorithm was able to select item combinations that were as strong as those selected by 0-1 LP (a method guaranteed to find an optimal solution if one exists). Yet Tabu search found a solution on average 4.90 minutes faster than 0-1 LP. Tabu search can be advantageous for empirical ATA applications: not only is it computationally inexpensive, but it also might be easier to implement for test developers who are less familiar with complex optimization techniques. Studies 2 and 3 thus exclusively used Tabu Search in conjunction with the Unbiased-ATA method.

Study 1 has provided an introductory understanding of Unbiased-ATA's performance with a variety of item banks and sample sizes. Namely, Unbiased-ATA appears promising for identifying an item combination with desirable psychometric properties in certain conditions. A related advantage of Unbiased-ATA is its role in efficiently reducing the number of differentially functioning items for test developers to review. However, numerous questions about Unbiased-ATA remain unanswered. For instance, would altering the weights of the objective function criteria help counteract the influence of miscategorized "DIF items?"

Chapter 4: Comparison of Weighting Schemes

As presented above, the Unbiased-ATA objective function was originally designed with equal weights for each criterion (i.e., $w_i = 1, i \in \{1,2,3,4\}$). This weighting scheme assumed that all psychometric properties considered in the objective function were equally important for the resulting test (Jia et al., 1998). In the context of Unbiased-ATA, equal weighting implies that having a test with evidence of item- and test-level invariance is just as important as having a test with high precision and good item fit. Equal weighting has been commonly used in ATA studies that incorporate multiple criteria (e.g., Olaru & Danner, 2021; Schultze & Eid, 2018), as well as when computing composite scores (e.g., in multivariate generalizability studies; G. A. Marcoulides, 1994).

It is plausible that alternative weighting schemes might improve upon the efficacy of Unbiased-ATA. For instance, when the criterion empirical distributions differ, each criterion can be weighted by the inverse of its variance (i.e., an index of precision). Less-precise criteria will then be given lesser weight. This inverse-variance weighting scheme was proposed for use in meta-analyses to compute weighted means with effect sizes from multiple studies (Hedges, 1983; Hedges & Vevea, 1998). However, inverse-variance weighting relies on independent measurements or criteria, an assumption that likely does not hold for the Unbiased-ATA objective function. The Study 1 results also demonstrated small to negligible variation in the objective function criteria: when computing condition-specific variances for each of the criteria, the maximum variances were 0.00143, 0.00142, 0.00173, and 0.00152 for γ_{DIF}, γ_{DTF}, $\gamma_{\text{Precision}}$, and γ_{IF}, respectively. With

minor differences in precision among the criteria, using inverse-variance weighting would likely not have a demonstrable effect on Unbiased-ATA's performance.

Alternatively, test developers might consider one or more criteria to be more important for the resulting test. In a recent study using ACO to develop invariant psychosocial short-form scales, Jankowsky and colleagues (2020) placed greater weight on the MI criterion within their ATA objective function. Specifically, the authors weighted the MI criterion by two. The other criteria in the objective function (representing overall model fit, test score reliability, and an equal number of positively and negatively worded items; see p. 485) were each given a weight of one, implying equal importance among these three criteria for the overall test.

Two other weighting schemes that place differential importance on the criteria are rank-sum (Einhorn & McCoach, 1977; Stillwell et al., 1981) and rank-order centroid (ROC) weights (Barron & Barrett, 1996). These approximate weighting schemes require test developers to rank the objective function from most to least important. Weights of decreasing size are then assigned based on these rankings. In applications where the true weight values are unknown, it can be easier for test developers to rank criteria in terms of importance rather than select quantitative values (Barron & Barrett, 1996). Previous research suggests that rank-sum and ROC weights can produce more accurate results (in terms of predictions and decision quality) in multiple-attribute decision making than uniform weighting (Barron & Barrett, 1996; Jia et al., 1998; Stillwell et al., 1981). In other studies, rank-based weighting schemes have performed similarly to equal weighting (Einhorn & McCoach, 1977). In certain cases, ROC weights have outperformed rank-sum

weights (Barron & Barrett, 1996; Jia et al., 1998), but these differences can be small in magnitude (Jia et al., 1998).

Selecting among rank-based weighting mechanisms largely relies on the data characteristics and test developers' preferences for the resulting psychometric properties. The differences in weight magnitudes for ROC weights become successively smaller as the criterion importance ranking decreases. On the other hand, rank-sum weights maintain uniform magnitude differences across the criteria. Therefore, "the greater the concentration of value in the first few attributes, the more attractive the ROC method" (Jia et al., 1998, p. 91). In Study 1, large proportions of items were often categorized as differentially functioning in the item bank. It is arguably advantageous in these scenarios for Unbiased-ATA to focus on reducing item and test bias, thus placing more importance on γ_{DIF} and γ_{DTF} than the other two criteria. Additionally, ranking γ_{DTF} above γ_{DIF} could aid Unbiased-ATA in selecting items that might be categorized as DIF but combine to show negligible differential functioning at the test level. Differential weighting with the ROC method thus merits further examination in the Unbiased-ATA context.

Study 2 extended the previous analyses to evaluate whether differential weighting of the criteria in the objective function affected Unbiased-ATA's performance. Specifically, equal weighting (as presented in Study 1) was compared to (a) double-weighting of the two MI criteria (Jankowsky et al., 2020) and (b) ROC weighting. The aim of this study was to provide additional guidance to practitioners seeking to implement the most effective variation of Unbiased-ATA in their test construction processes.

Simulation Design

Weighting Schemes

Study 2 compared three objective functions that differed by the value of the weights applied to each criterion. To explicitly account for these weights, Equation 20 can be rewritten as

$$f(\text{UATA}) = w_1\gamma_{\text{DIF}} + w_2\gamma_{\text{DTF}} + w_3\gamma_{\text{Precision}} + w_4\gamma_{\text{IF}}. \quad (25)$$

When objective functions incorporate multiple criteria, the associated weights are typically normalized (i.e., $\sum_{i=1}^{W} w_i = 1$ where W equals the total number of weights in the function). This normalization ensures equivalent scaling of the different criteria. Recall that in the proposed Unbiased-ATA objective function, the criteria were constructed to all be on the same [0,1] scale, precluding the necessity for weight normalization. In Study 2, normalized weights were used to better compare the three weighting schemes.

The first objective function used equal weighting to replicate the Study 1 objective function. With four criteria, the weights would then be $w_1 = w_2 = w_3 = w_4 = 1/4$. However, the $\gamma_{\text{Precision}}$ and γ_{IF} are each composed of two terms (one for the reference group and the other for the focal group). To obtain $w_3 = w_4 = 1/4$ when distributing across the terms in each criterion, each of these weights was divided by two (i.e., $w_3 = w_4 = 1/8$). This is equivalent to rewriting $\gamma_{\text{Precision}}$ and γ_{IF} such that the reference and focal group terms were each divided by two, and then using $w_3 = w_4 = 1/4$. Incorporating these weights into Equation 25, the equal weighting objective function is

$$f(\text{UATA}_{\text{EW}}) = (1/4)\gamma_{\text{DIF}} + (1/4)\gamma_{\text{DTF}} + (1/8)\gamma_{\text{Precision}} + (1/8)\gamma_{\text{IF}}, \quad (26)$$

where EW denotes "equal weighting." Then, the second objective function mirrored Jankowsky et al.'s (2020) study, wherein the MI criteria received a weight twice the magnitude as the other criteria. Incorporating the weight normalization with $w_1 = w_2 = 1/3$ and $w_3 = w_4 = 1/6$, the full objective function is

$$f(\mathrm{UATA_{DW}}) = (1/3)\gamma_{\mathrm{DIF}} + (1/3)\gamma_{\mathrm{DTF}} + (1/12)\gamma_{\mathrm{Precision}} + (1/12)\gamma_{\mathrm{IF}}, \quad (27)$$

where DW denotes "double weighting."

The third objective function in Study 2 used a ROC weighting scheme (Barron & Barrett, 1996). Generally, ROC weights are computed as (Barron & Barrett, 1996; Jia et al., 1998, Equation 4)

$$w_{(i)} = \frac{1}{W}\sum_{m=1}^{W}\frac{1}{m}, \quad (28)$$

where again W is the total number of weights, or criteria, in the objective function. Given four criteria, the weights are computed as $w_1 = \frac{1}{4}\left(1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4}\right) = 0.52$, $w_2 = \frac{1}{4}\left(\frac{1}{2} + \frac{1}{3} + \frac{1}{4}\right) = 0.27$, $w_3 = \frac{1}{4}\left(\frac{1}{3} + \frac{1}{4}\right) = 0.15$, and $w_4 = \frac{1}{4}\left(\frac{1}{4}\right) = 0.06$ (Jia et al., 1998, p. 90).

The largest weight was assigned to γ_{DTF} in Study 2. The previous study showed that ATA-selected tests could have relatively higher proportions of items categorized as DIF but small to negligible DTF. In scenarios with many bank items categorized as DIF, it is preferred that Unbiased-ATA selects an item combination with less evidence of DTF. The second largest weight was assigned to γ_{DIF} to encourage Unbiased-ATA to focus on items categorized as anchors in the bank. Then, $\gamma_{\mathrm{Precision}}$ and γ_{IF} were ranked based on the amount of variation seen in ATA-selected tests' test information and item fit proportions in Study 1. Specifically, greater importance was placed on $\gamma_{\mathrm{Precision}}$ because

of greater differences in TIF deviations in the selected tests across conditions. Combining these weights with the criteria, the objective function is

$$f(\mathrm{UATA}_{\mathrm{ROC}}) = 0.27\gamma_{\mathrm{DIF}} + 0.52\gamma_{\mathrm{DTF}} + (0.15/2)\gamma_{\mathrm{Precision}} + (0.06/2)\gamma_{\mathrm{IF}}. \quad (29)$$

It merits comment that the criterion ranking chosen for Study 2 is one of 24 possible rankings. The ranking of importance in Equation 29 ($\gamma_{\mathrm{DTF}} > \gamma_{\mathrm{DIF}} > \gamma_{\mathrm{Precision}} > \gamma_{\mathrm{IF}}$) was selected to both reflect Jankowsky et al.'s (2020) higher weighting of the MI criteria while also attempting to reduce the consequences of high false and true positive DIF items in the bank. Other testing contexts might necessitate different criterion rankings. Given that the ROC weights are relatively easy to compute, test developers can modify the rank-ordering and obtain alternative weights.

Simulation Design Factors

The design factors used in Study 2 were based on those that demonstrated relatively larger effects on Unbiased-ATA's performance in Study 1. A total of four design factors were manipulated. These factors included: (a) Total sample size, $N = \{500,1000,5000\}$, (b) direction of DIF, (c) magnitude of DIF, and (d) percentage of DIF in the item bank.

Study 2 only examined conditions with simulated DIF where the type of DIF was an even mixture of uniform and non-uniform parameter differences. DIF direction then indicated whether parameter modifications were made only to the focal group ("focal group" direction), or modifications were made evenly to both the reference and focal groups ("both groups" direction). The DIF magnitudes were either small, large, or a mixture of small and large. Here, the small and large DIF magnitudes referred to wABC values of 0.1 or 0.2, respectively (Edelen et al., 2015).

Recall that in Study 1, the percentages of differentially functioning items in the bank were either 1/6 (10 items), 1/3 (20 items), or 1/2 (30 items). The Study 1 results suggested that neither regularized DIF nor IRT-LRT maintained FPRs close to the nominal rate ($\alpha = 0.05$) with high DIF percentages. To examine Unbiased-ATA's performance with smaller amounts of DIF, the Study 2 percentages were instead set at 1/15 (4 items), 1/10 (6 items), or 1/5 (12 items). Note that these DIF percentages were chosen such that an even number of items had simulated DIF, which was necessary for the "both groups" direction.

Other design factors in the Study 2 simulation were held constant across conditions. First, a balanced sample size was used to maintain sufficient sample sizes for 2PLM item parameter estimation (Sahin & Anil, 2017). Additionally, Study 1 often demonstrated negligible differences in results whether using true or estimated item parameters. True item parameters were then used in Study 2 to provide more control over the introduction of simulated DIF, and thus to better understand Unbiased-ATA's performance in the new DIF conditions. Finally, each ATA algorithm was conducted using Tabu search, which performed similarly to 0-1 LP within a substantially shorter time frame in the previous study. In total, the simulation comprised 54 conditions.

Simulation Procedure

The Study 2 simulation procedure and the examined dependent variables mirrored those from Study 1. After generating the person parameter values and item responses according to the condition specifications, the bank-level analyses for Unbiased-ATA were conducted using regularized DIF to categorize anchor and differentially functioning items. Three ATA algorithms were then run with each item bank, corresponding to the

three weighting schemes: equal weighting (EW), double weighting of the MI criteria (DW), and rank-order centroid weighting (ROC). To account for the possibility of local minima, Tabu search was repeated five times for each algorithm and the item combination with the highest objective function value was selected as the "best" test. Each simulation condition was replicated $R = 100$ times. Again, all analyses were run in R statistical software, version 4.1.1 (R Core Team, 2021). The Study 2 simulation used the same R packages as Study 1.

Results

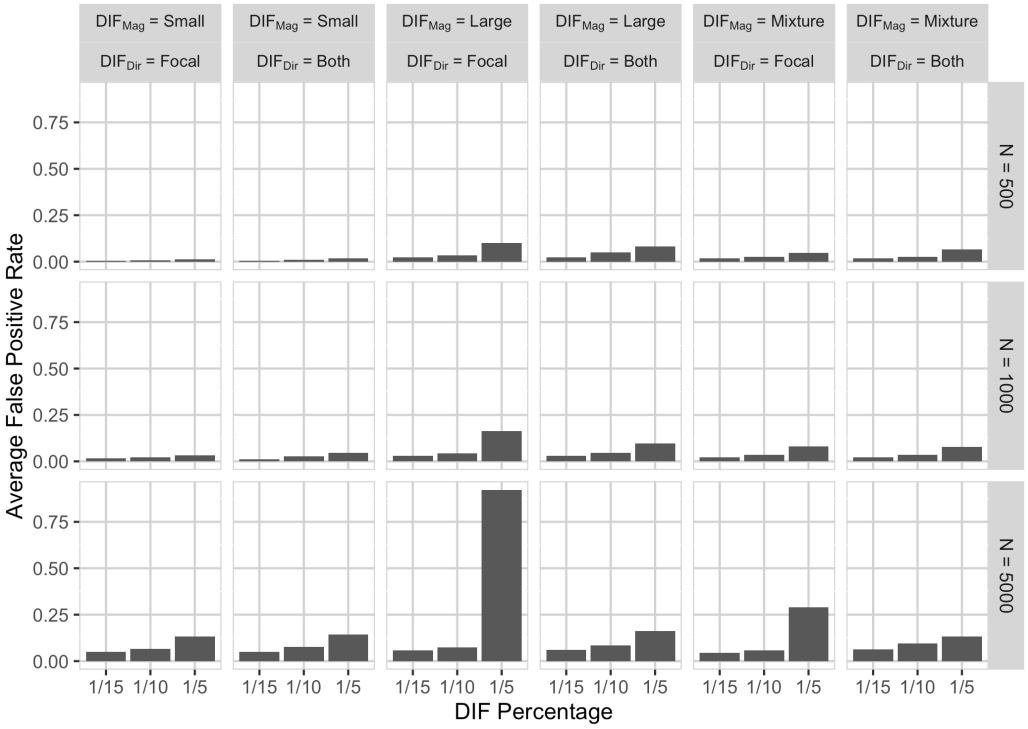
Regularized DIF Performance

Figure 23 presents the $\overline{\text{FPR}}$ results for the various testing scenarios in Study 2. When 1/5 or fewer of the items in the bank were differentially functioning, regularized DIF maintained $\overline{\text{FPR}}$ values closer to the nominal error rate ($\alpha = 0.05$). The median $\overline{\text{FPR}}$ across the examined conditions was 0.046 with an interquartile range of 0.024 to 0.077. For reference, with DIF percentages between 1/6 and 1/2 in Study 1, the $\overline{\text{FPR}}$ interquartile range spanned 0.077 to 0.303.

$\overline{\text{FPR}}$ values in Study 2 were positively associated with DIF percentage. Marginalizing across the other factors, $\overline{\text{FPR}}$ was 0.031, 0.045, and 0.145 for DIF percentages of 1/15, 1/10, and 1/5, respectively. Moreover, $\overline{\text{FPR}}$ increased as sample size increased from $N = 500$ to $N = 5000$. The largest $\overline{\text{FPR}}$ occurred when 1/5 of the item bank had large DIF in the "focal group" direction and $N = 5000$ (Column 3, Row 3); in this condition, $\overline{\text{FPR}}$ reached 0.923. $\overline{\text{FPR}}$ also exceeded 0.25 with a mixture of small and large DIF in the "focal group" direction and $N = 5000$ (Column 5, Row 3). The inflated

$\overline{\text{FPR}}$ with large DIF percentages, large sample sizes, and in the "focal group" direction align with the results seen in Study 1 (see Figure 2).

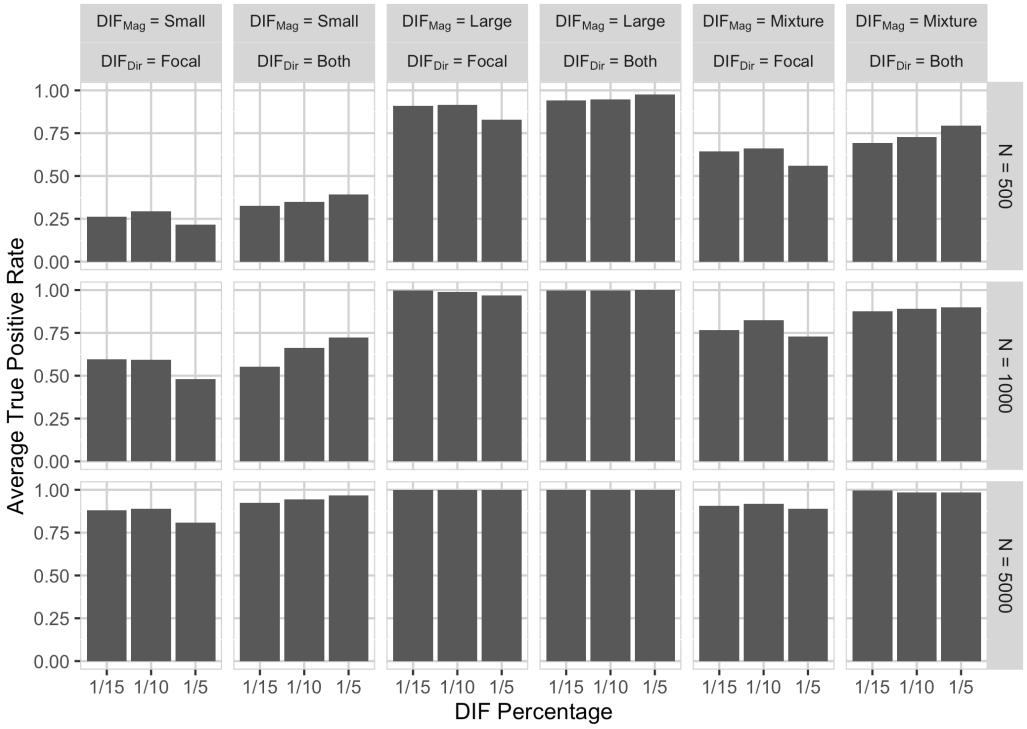
Figure 23. Average False Positive Rates for Regularized DIF with Smaller DIF Percentages



Even with smaller $\overline{\text{FPR}}$, regularized DIF still maintained moderate to high TPRs in most conditions. As shown in Figure 24, $\overline{\text{TPR}}$ exceeded 0.50 in 87% of the examined conditions and exceeded 0.80 in 63% of the conditions. The smallest $\overline{\text{TPR}}$ occurred with small DIF magnitudes and $N = 500$ ($0.215 \leq \overline{\text{TPR}} \leq 0.392$; Columns 1-2, Row 1). Alternatively, $\overline{\text{TPR}}$ ranged from 0.828 to 1.00 with large DIF magnitudes. The relationship between DIF percentage and $\overline{\text{TPR}}$ depended on the DIF direction. Specifically, for the "focal group" direction, $\overline{\text{TPR}}$ was slightly smaller when DIF percentage was 1/5 compared to when DIF percentage was 1/10 or 1/15. Yet for the "both groups" direction, $\overline{\text{TPR}}$ generally increased as DIF percentage increased. For either

direction, these relationships weakened as sample size increased (given that $\overline{\text{TPR}}$ neared the ceiling of 1.00).

Figure 24. Average True Positive Rates for Regularized DIF with Smaller DIF Percentages



Taken together, the $\overline{\text{FPR}}$ and $\overline{\text{TPR}}$ results indicate that within most item banks, regularized DIF correctly categorized items as either differentially functioning or invariant. The important exception was for the largest sample size ($N = 5000$), DIF percentage (1/5), and DIF magnitude paired with the "focal group" direction. In this condition, close to all items in the bank were categorized as differentially functioning (even though only 12 out of 60 items were truly non-invariant). Studies 1 and 2 together provide evidence that regularized DIF maintains a more acceptable balance between Type I error and power rates in 60-item banks when 1/5 or fewer of the items are differentially functioning.

Comparing Tests Across Weighting Schemes

Parameter Summaries. Figure A13 in the appendix reveals negligible differences in the average estimated $\hat{a}$ and $\hat{b}$ values among items selected by the three weighting schemes. There was preliminary evidence that tests selected by DW comprised items with slightly lower $\hat{a}$ (Figure A13A) and higher $\hat{b}$ (Figure A13B), on average, than EW or ROC with larger DIF magnitudes and percentages in the bank. However, these differences were less than 0.05 in magnitude and therefore should be interpreted with caution.

Moreover, Figure A13A suggests that regardless of weighting scheme, ATA-selected tests comprised items with smaller $\hat{a}$ on average as sample size increased. Figure A13B then shows that the average $\hat{b}$ slightly decreased as DIF percentages increased in the "both groups" direction (Columns 2, 4, and 6). The average estimated item parameters were also roughly equivalent between the reference and focal groups. Across the examined simulation conditions, average $\hat{a}$ ranged between 1.488 and 1.543 for both the reference and focal groups. The corresponding range for average $\hat{b}$ was 0.939 – 1.051.

Effect Sizes. Partial η^2 effect sizes were next computed to gauge whether using different weighting schemes influenced the characteristics of the ATA-selected tests. Because Study 2 only examined test conditions with differentially functioning items in the bank, a series of fully crossed, multiway ANOVAs was fit to the data. Each dependent variable was regressed on weighting scheme, sample size, DIF percentage, DIF magnitude, DIF direction, and the two- and three-way interactions between these factors. In this analytic approach, all design factors were conceptualized as fixed effects.

The values for the dependent variables were averaged across the 100 simulation repetitions in each condition.

Table 6 gives the η_p^2 values for the 11 different models. The weighting scheme most strongly influenced the TIF deviations and number of well-fitting items within the ATA-selected tests (Columns 6 – 9). For instance, weighting scheme had a large main effect on the number of well-fitting items for either the reference or focal group ($\eta_p^2 \geq 0.43$), as well as when paired with DIF percentage and total sample size ($0.18 \leq \eta_p^2 \leq 0.62$). Weighting had negligible effects on the item- and test-level MI (Columns 1 – 5). Moreover, sample size and DIF characteristics consistently demonstrated moderate to large effects on the MI characteristics, TIF deviations, and the number of well-fitting items. For example, $\eta_p^2 = 0.75$ for sample size when examining the full-sample RMSEAs for a strong MI model. Finally, η_p^2 values for the various design factors were relatively smaller when modeling the external validity coefficients.

Notice also in Table 6 that regressing the group-level SRMSRs on the dependent variables produced R^2 values of 1.00. Further examination of these models revealed that the high R^2 values were due to including sample size as a predictor. Indeed, coupled with small total sums of squares for SRMSR_R and SRMSR_F, sample size accounted for roughly all the variation in the models. Refitting the models without sample size reduced R^2 to less than 1.00 and all other η_p^2 values to less than 0.01. Therefore, the SRMSR effect sizes for other design factors in Table 6 should be interpreted with caution.

Table 6. Partial η^2 Effect Sizes When Regressing Test Properties on Weighting Scheme, Sample Size, and DIF Characteristics

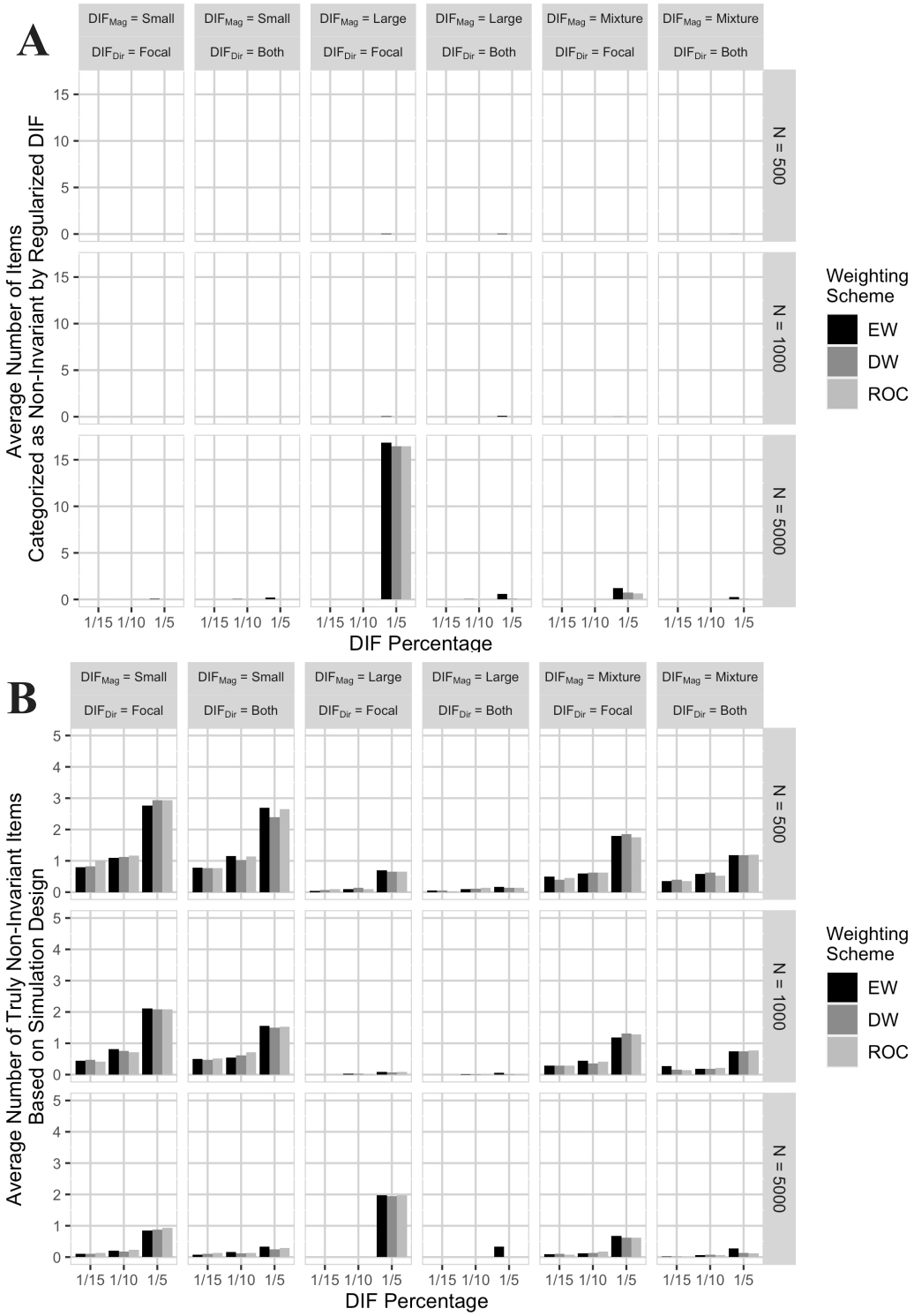
	Measurement Invariance					Information		Item Fit		Validity	
Design Factor	$n_{\text{No DIF}}$	uDTF	RMSEA	SRMSR_{R}	SRMSR_{F}	$\Delta_{\text{TIF}_{\text{R}}}$	$\Delta_{\text{TIF}_{\text{F}}}$	$n_{\text{Fitting}_{\text{R}}}$	$n_{\text{Fitting}_{\text{F}}}$	r_{R}	r_{F}
Weighting	0.00	0.01	0.00	0.03	0.00	**0.20**	**0.17**	**0.55**	**0.43**	0.07	0.02
Total Sample Size (N)	**0.24**	**0.26**	**0.75**	**1.00**	**1.00**	**0.76**	**0.80**	**0.61**	**0.53**	**0.19**	**0.18**
DIF Percentage (Perc)	**0.24**	**0.28**	**0.52**	**0.33**	0.05	**0.95**	**0.95**	**0.77**	**0.75**	0.08	0.05
DIF Magnitude (Mag)	**0.21**	**0.23**	**0.17**	**0.28**	0.09	**0.80**	**0.82**	**0.34**	**0.35**	**0.20**	**0.13**
DIF Direction (Dir)	0.12	0.11	**0.49**	**0.55**	**0.44**	**0.78**	**0.67**	0.01	0.02	**0.13**	**0.17**
Weighting x N	0.00	0.01	0.01	0.01	0.04	**0.16**	0.11	**0.39**	**0.18**	0.06	0.02
Weighting x Perc	0.00	0.02	0.02	0.01	0.03	**0.29**	**0.24**	**0.62**	**0.52**	0.08	0.07
Weighting x Mag	0.00	0.01	0.02	0.00	0.01	0.03	0.03	**0.14**	0.12	0.02	0.01
Weighting x Dir	0.00	0.00	0.00	0.01	0.00	0.05	0.03	0.07	0.04	0.02	0.03
N x Perc	**0.38**	**0.41**	**0.16**	**0.73**	**0.33**	**0.84**	**0.88**	**0.73**	**0.67**	**0.15**	**0.15**
N x Mag	**0.34**	**0.35**	**0.54**	**0.47**	**0.41**	**0.14**	0.07	**0.16**	**0.14**	**0.22**	**0.17**
N x Dir	**0.22**	**0.20**	0.01	0.02	0.09	**0.38**	0.09	0.06	**0.20**	0.08	0.02
Perc x Mag	**0.34**	**0.38**	0.11	**0.42**	**0.32**	**0.83**	**0.86**	**0.49**	**0.53**	**0.21**	**0.23**
Perc x Dir	**0.22**	**0.20**	**0.49**	**0.60**	**0.29**	**0.76**	**0.60**	0.07	0.05	0.12	**0.15**
Mag x Dir	**0.20**	**0.19**	0.04	**0.28**	0.11	**0.50**	**0.21**	0.10	0.06	0.03	0.09
Weighting x N x Perc	0.00	0.03	0.02	0.06	0.03	**0.26**	**0.18**	**0.50**	**0.26**	**0.17**	0.03
Weighting x N x Mag	0.00	0.01	0.03	0.06	0.05	0.08	0.04	0.11	0.03	**0.13**	0.03
Weighting x N x Dir	0.00	0.00	0.02	0.02	0.04	0.05	0.03	0.04	0.01	0.10	0.03
Weighting x Perc x Mag	0.00	0.02	0.04	0.02	0.05	0.05	0.06	**0.20**	**0.22**	0.11	0.06
Weighting x Perc x Dir	0.00	0.00	0.01	0.02	0.03	0.07	0.05	0.07	0.06	0.07	0.03
Weighting x Mag x Dir	0.00	0.00	0.02	0.05	0.01	0.02	0.02	0.03	0.04	0.02	0.02
N x Perc x Mag	**0.51**	**0.52**	**0.56**	**0.50**	**0.47**	**0.27**	**0.29**	**0.29**	**0.31**	**0.54**	**0.15**
N x Perc x Dir	**0.36**	**0.34**	**0.15**	**0.50**	**0.24**	**0.29**	0.02	**0.22**	**0.37**	**0.25**	**0.28**
N x Mag x Dir	**0.33**	**0.33**	**0.37**	**0.42**	**0.35**	**0.23**	**0.21**	**0.16**	**0.22**	**0.37**	0.09
Perc x Mag x Dir	**0.33**	**0.31**	**0.22**	**0.24**	**0.15**	**0.47**	0.10	**0.15**	0.11	**0.31**	0.01
R^2	0.859	0.865	0.911	1.000	1.000	0.980	0.979	0.940	0.927	0.838	0.725
Adjusted R^2	0.645	0.660	0.776	0.999	0.999	0.950	0.948	0.850	0.817	0.591	0.308

Note. Effect sizes greater than or equal to 0.13 are bolded (Cohen, 1992).

Item- and Test-Level MI. Item-level MI was operationalized as the number of differentially functioning items in the ATA-selected tests. Figure 25 presents the average number of items in each test that were (a) categorized as differentially functioning by regularized DIF ($\text{DIF}_{\text{RegDIF}}$; Panel A) or (b) truly non-invariant items (DIF_{True}; Panel B). As shown in Figure 25A, the ATA-selected tests comprised the most $\text{DIF}_{\text{RegDIF}}$ items (over 15 items on average) when there was large DIF for 1/5 of the items in the "focal group" direction with $N = 5000$ (Column 3, Row 3). Regularized DIF also demonstrated the highest $\overline{\text{FPR}}$ in this condition. Excluding this condition, the average number of $\text{DIF}_{\text{RegDIF}}$ items was very small and ranged from 0.00 to 1.21. Non-zero average numbers of $\text{DIF}_{\text{RegDIF}}$ items more often occurred with $N = 5000$ and DIF percentage of 1/5. Importantly, there were negligible differences across the weighting schemes. When marginalizing across the other factors, the average number of $\text{DIF}_{\text{RegDIF}}$ items was 0.361, 0.320, and 0.316 for EW, DW, and ROC, respectively.

Figure 25B demonstrates that the average number of DIF_{True} items did not exceed 3, or 15% of the 20-item test, even in conditions with high $\overline{\text{FPR}}$ and $\overline{\text{TPR}}$. Weighting scheme was not associated with differences in the average number of DIF_{True} items; the maximum difference between any pairwise combination of weighting schemes was only 0.33. Irrespective of the weighting scheme, Unbiased-ATA tended to select more DIF_{True} items with smaller sample sizes and when 1/5 of the items in the bank were differentially functioning. Comparing Figure 25B to Figure 24, tests comprised more DIF_{True} items when $\overline{\text{TPR}}$ were lower. For example, the highest average number of DIF_{True} items occurred for small DIF magnitudes with 1/5 differentially functioning items and $N = 500$ (Columns 1-2, Row 1). Regularized DIF also had the lowest power in these

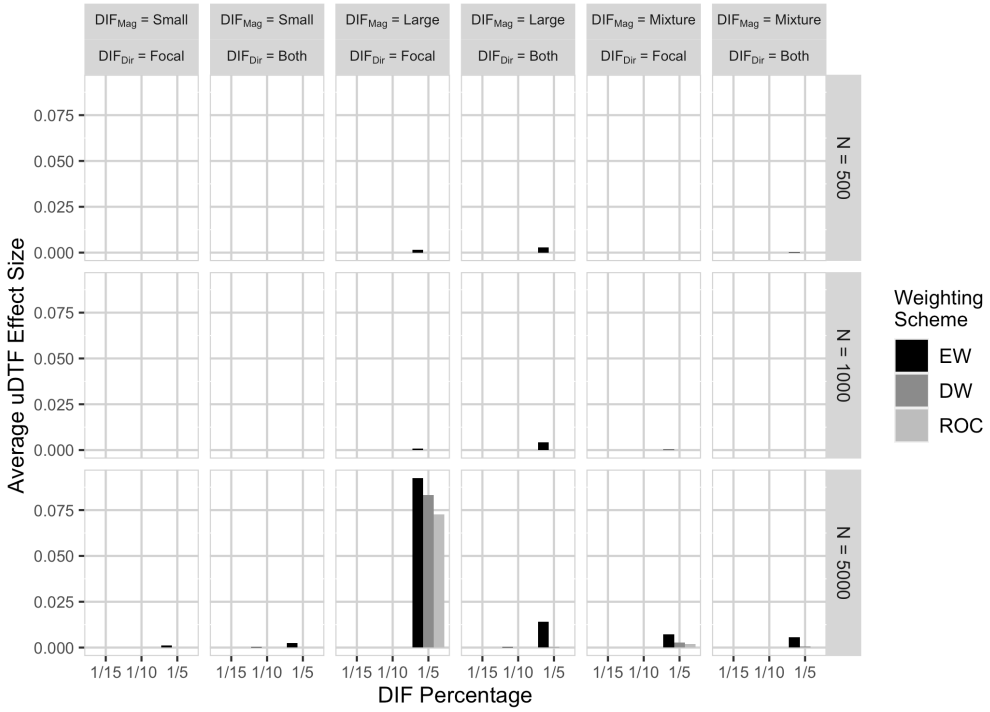
Figure 25. Average Number of Differentially Functioning Items in the Selected Tests Across Weighting Schemes



conditions, with $\overline{\text{TPR}} < 0.40$ (see Figure 24). Lower power translated to more DIF_{True} items in the bank categorized as anchor items (i.e., invariant). Then, Unbiased-ATA was more likely to preference these miscategorized items during item selection.

Test-level MI was evaluated using the uDTF effect size (Figure 26). The highest $\overline{\text{uDTF}}$ was 0.092, occurring in the same condition with the highest regularized DIF $\overline{\text{FPR}}$ and the highest average number of $\text{DIF}_{\text{RegDIF}}$ items. Excluding this condition, $\overline{\text{uDTF}}$ ranged from 0.000 to 0.014, translating to a maximum percent scoring difference between groups of only 0.07% (Chalmers et al., 2016). Non-zero $\overline{\text{uDTF}}$ values more often occurred when $N = 5000$ and the DIF percentage was 1/5. Furthermore, differences among the weighting schemes were small in magnitude: whereas Figure 26 indicates that EW sometimes produced tests with higher $\overline{\text{uDTF}}$ than DW or ROC, the magnitude of those differences was at most 0.02.

Figure 26. Average uDTF Effect Size for the Selected Tests Across Weighting Schemes



Test-level MI was also evaluated by fitting the data from each ATA-selected test to a series of successively more restrictive models representing configural, weak, or strong MI. The full-sample RMSEA and group-level SRMSR fit statistics were then calculated for each model and averaged across the simulation replications in each condition. Figures A14 and A17 in the appendix display the average full-sample RMSEA and group-level SRMSR values, respectively, when fitting the ATA-selected test data to a strong MI model.

To summarize these figures, there was little evidence of differences in model fit statistics as a function of the weighting scheme. Any differences in $\overline{\text{RMSEA}}$ did not exceed 0.002, nor did differences in $\overline{\text{SRMSR}}_{\text{R}}$ or $\overline{\text{SRMSR}}_{\text{F}}$ exceed 0.001. Across all testing scenarios, $\overline{\text{RMSEA}}$ remained below the threshold of "good" fit ($\text{RMSEA} = 0.05$) and there was little variation in $\overline{\text{RMSEA}}$ for different sample sizes and DIF characteristics. $\overline{\text{RMSEA}}$ was highest ($\overline{\text{RMSEA}} = 0.020$) when large DIF occurred for 1/5 of the items in the "focal group" direction with $N = 5000$ (Figure A14, Column 3, Row 3). $\overline{\text{SRMSR}}_{\text{R}}$ and $\overline{\text{SRMSR}}_{\text{F}}$ were also not noticeably different across DIF characteristics and generally remained below the corresponding "good" fit threshold ($\text{SRMSR} = 0.05$) when $N \geq 1000$. Comparing the average fit statistics between groups, $\overline{\text{SRMSR}}_{\text{F}}$ was uniformly lower than $\overline{\text{SRMSR}}_{\text{R}}$. These model fit statistics were also influenced by the total sample size. For example, marginalizing across other factors, $\overline{\text{SRMSR}}_{\text{R}}$ decreased from 0.065 to 0.031 and $\overline{\text{SRMSR}}_{\text{F}}$ decreased from 0.053 to 0.017 as N increased from 500 to 5,000.

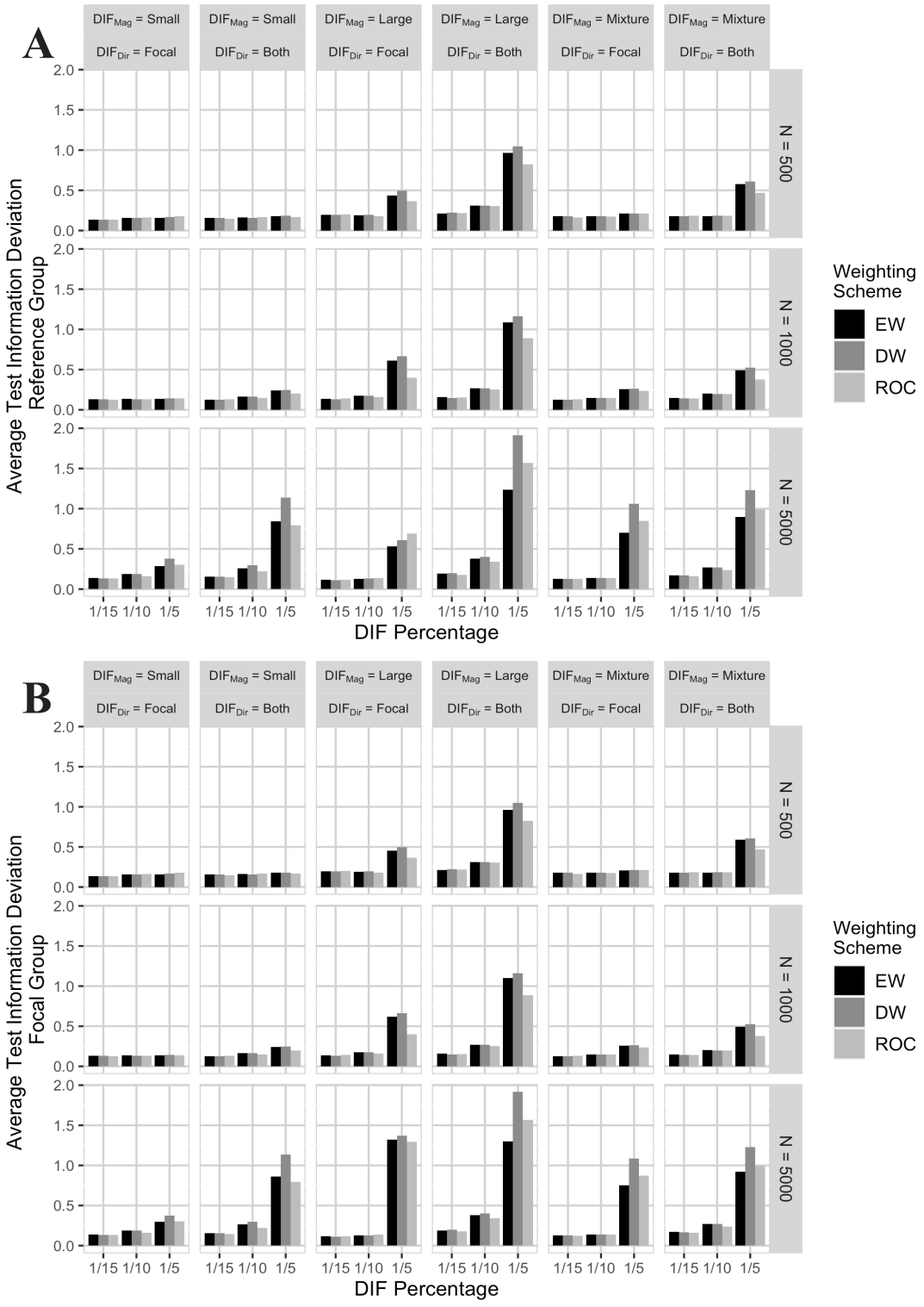
Figures A15 – A16 and Figures A18 – A19 show corresponding plots for $\overline{\text{RMSEA}}$ and $\overline{\text{SRMSR}}$, respectively, when fitting the test data to either weak or configural MI models. As with the strong MI results, there were negligible differences in the average

model fit statistics among the three weighting schemes. Rather, both $\overline{\text{RMSEA}}$ and $\overline{\text{SRMSR}}$ decreased both as total sample size increased and when fitting successively less restrictive models. Still, when $N \geq 1000$, both $\overline{\text{RMSEA}}$ and $\overline{\text{SRMSR}}$ aligned with "good" fit for a strong MI model.

Test Information. Figure 27 shows small but noticeable differences in $\bar{\Delta}_{\text{TIF}}$ among the weighting schemes when 1/5 of the items in the bank were differentially functioning (reflecting moderate to large η_p^2 in Table 6). Specifically, DW selected tests with slightly higher $\bar{\Delta}_{\text{TIF}}$ for both the reference (Panel A) and focal groups (Panel B). Differences in $\bar{\Delta}_{\text{TIF}}$ between DW and either EW or ROC were larger as sample size increased. When marginalizing across the other DIF characteristics, $\bar{\Delta}_{\text{TIF}_{\text{R}}}$ values with DIF percentage of 1/5 and $N = 500$ were 0.453, 0.420, and 0.371 for DW, EW, and ROC, respectively. When $N = 5000$, these averages were 1.054, 0.750, and 0.867. With $N \leq 1000$, large DIF magnitudes, and 1/5 differentially functioning items (e.g., Columns 3 – 4, Rows 1 – 2), ROC produced smaller $\bar{\Delta}_{\text{TIF}}$ than EW. This trend reversed at $N = 5000$ (Row 3), with EW producing the smallest relative $\bar{\Delta}_{\text{TIF}}$ among the weighting schemes. Across the testing scenarios, however, differences in $\bar{\Delta}_{\text{TIF}}$ among weighting schemes did not exceed 0.677. Figure A20 in the appendix also reveals negligible differences in the full TIF values as a function of weighting scheme.

Additional trends in $\bar{\Delta}_{\text{TIF}}$ were apparent in Figure 27 irrespective of weighting scheme, although differences in $\bar{\Delta}_{\text{TIF}}$ were again small and should be interpreted with caution. For example, $\bar{\Delta}_{\text{TIF}}$ was noticeably higher with DIF percentages of 1/5 compared to percentages of 1/15 or 1/10. This increase in $\bar{\Delta}_{\text{TIF}}$ was exacerbated with larger DIF

Figure 27. Average Test Information Function Deviations for Selected Tests Across Weighting Schemes

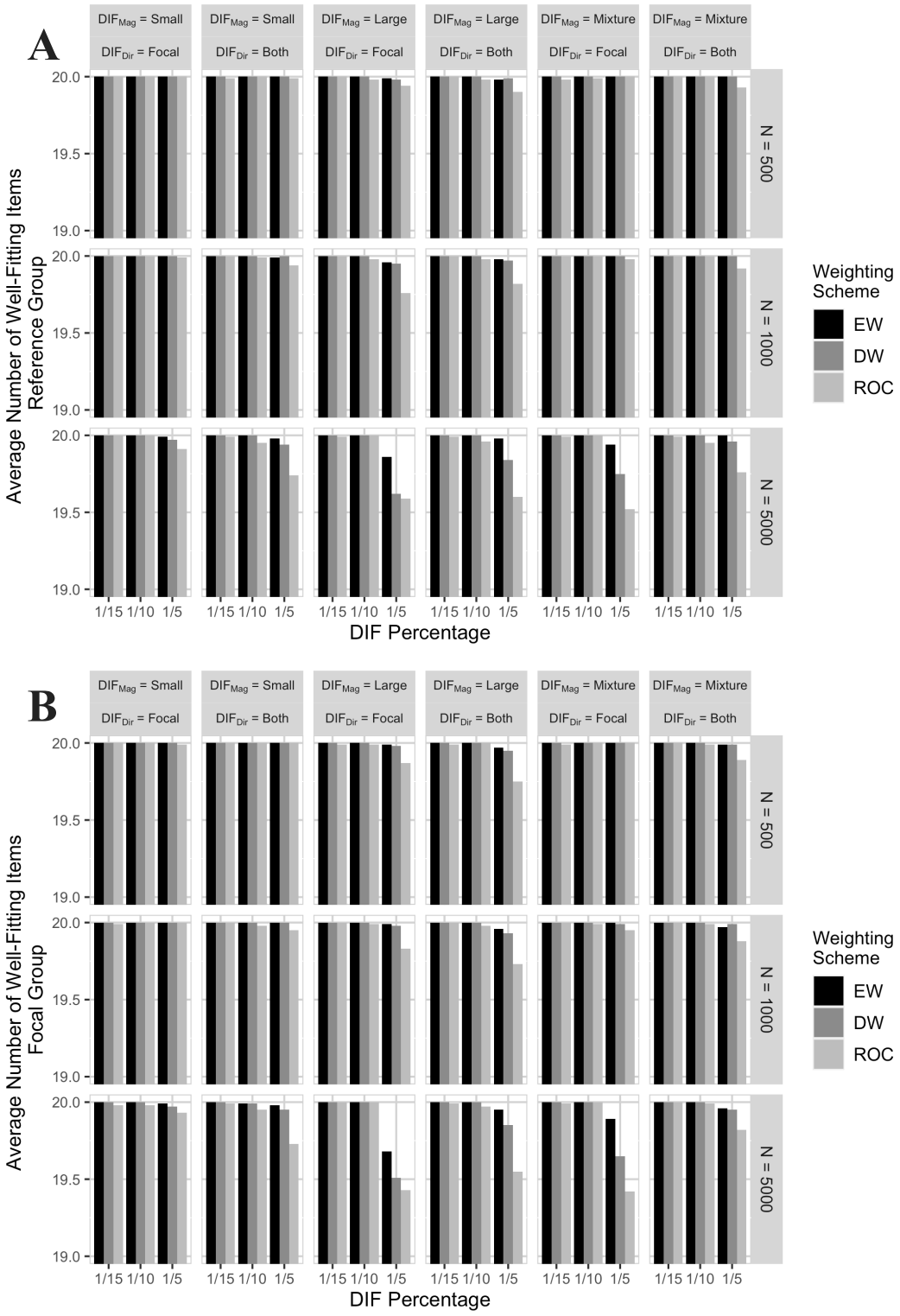


magnitudes, larger sample sizes, and in the "both groups" direction. Moreover, smaller DIF percentages in Study 2 were associated with minimal differences in $\bar{\Delta}_{\mathrm{TIF}}$ between the reference and focal groups. Recall from Study 1 that $\bar{\Delta}_{\mathrm{TIF_F}}$ neared or exceeded 10.00 in item banks with high regularized DIF FPRs (Figure 17B, Row 3). Given the lower FPRs in Study 2, more anchor items in the bank meant that more estimated item parameters were constrained to be equivalent between groups. More similar group-level parameter estimates thus produced more similar $\bar{\Delta}_{\mathrm{TIF_R}}$ and $\bar{\Delta}_{\mathrm{TIF_F}}$ values.

Item Fit. Next, the average number of well-fitting items by group, as categorized by the $S - \chi^2$ index (Orlando & Thissen, 2000), was calculated for each simulation condition. Factors incorporating the weighting scheme had moderate to large effects ($0.14 \leq \eta_p^2 \leq 0.62$) when examining the number of well-fitting items. However, Figure 28 reveals that these large effect sizes translated to minimal differences in practice. Notice in this figure that the average number of well-fitting items in ATA-selected tests ranged between 19.42 and 20.00 items. Among the weighting schemes, ROC selected the fewest well-fitting items on average for either the reference or focal group as sample size, DIF magnitude, and DIF percentage increased. EW consistently selected tests with the most well-fitting items on average. Yet the maximum difference in the average number of well-fitting items between ROC and EW tests was 0.47. Hence, the variation in Figure 28 as a function of weighting scheme could be attributed to only a handful of items at most.

External Validity. The final psychometric property examined in Study 2 was the correlation between the ATA-selected tests and an external criterion measure (hypothesized to measure a related latent trait). The average correlations ($\bar{r}$) ranged from 0.305 to 0.352 across the various testing scenarios (Figure A21). The correlations were

Figure 28. Average Number of Well-Fitting Items for Selected Tests Across Weighting Schemes



slightly higher for the focal group ($\bar{r}_F = 0.337$) compared to the reference group ($\bar{r}_R = 0.321$). Although there was larger variation in $\bar{r}$ among smaller sample sizes, no other trends across the design characteristics were readily apparent. Indeed, differences in $\bar{r}$ as a function of weighting scheme were small in magnitude. Each weighting scheme selected tests with higher $\bar{r}$ in assorted conditions, possibly the result of chance variation. In summary, the three weighting schemes selected tests with relatively similar linear relationships with an external criterion measure.

Sensitivity Analysis

The results above suggest more noticeable differences in $\overline{\text{uDTF}}$, $\bar{\Delta}_{\text{TIF}}$, and item fit among the weighting schemes as DIF percentage increased. To further evaluate these trends, the Study 2 simulation was extended to conditions where 1/3 or 1/4 of the items in the bank were differentially functioning. Figure A22 in the appendix shows the $\overline{\text{FPR}}$ for regularized DIF with DIF percentages between 1/15 and 1/4. For each simulation condition, the differences in uDTF, Δ_{TIF}, and the number of well-fitting items when either subtracting EW from DW ("DW – EW") or when subtracting EW from ROC ("ROC – EW") were calculated. These differences were then averaged across the 100 replications of each simulation condition. Figures A23 through A25 then plot these average differences; here, a negative value indicates that the test characteristic value was higher on average for EW.

For all three test characteristics, there was evidence of growing differences between EW and the other two weighting schemes with larger DIF percentages. Specifically, as DIF percentage increased from 1/15 to 1/4 of the bank, on average (a) EW selected tests with increasingly higher uDTF values than DW or ROC, (b) DW

selected tests with increasingly larger Δ_{TIF} than EW (and ROC to a smaller extent), and (c) ROC selected tests with increasingly fewer well-fitting items than EW (and DW to a smaller extent). Larger differences in uDTF more often appeared in conditions with higher regularized DIF $\overline{\text{FPR}}$ (see Figure A22). For example, correlations between $\overline{\text{FPR}}$ and average differences in uDTF values were -0.609 for DW – EW, and -0.859 for ROC – EW. For Δ_{TIF} and the number of well-fitting items, the pattern of differences also generally aligned with inflated $\overline{\text{FPR}}$ when $N \leq 1000$. When $N = 5000$, differences between EW and the other weighting schemes were often largest in the "both groups" direction (which had relatively smaller $\overline{\text{FPR}}$ than the "focal groups" direction; Figure A22, Row 3). However, it is important to emphasize that the magnitudes of differences among the weighting schemes were relatively small. Therefore, even when up to one-quarter of the bank contained differentially functioning items and $\overline{\text{FPR}}$ neared 1.00, using DW or ROC did not translate to tests with substantially dissimilar test characteristics on average than EW.

Discussion

Study 2 aimed to compare the efficacy of three weighting schemes—EW, DW, and ROC—when paired with the Unbiased-ATA objective function. In this study, efficacy was operationalized using the psychometric properties of the ATA-selected tests, including evidence for item- and test-level MI, the alignment of the TIF with a target function, and the number of well-fitting items. Differential weighting of the objective function criteria can be advantageous when test developers place greater importance on a particular test property (e.g., allowing larger differences in TIFs to prioritize reductions in test bias).

When 1/5 or fewer of the items in the bank were differentially functioning, the three weighting schemes performed similarly across the examined test characteristics. There was preliminary evidence that each weighting scheme was associated with reductions in a particular psychometric property. Specifically, EW selected items with higher $\overline{\text{uDTF}}$, DW selected items that produced higher $\bar{\Delta}_{\text{TIF}}$, and ROC selected higher numbers of misfitting items. The effects of weighting scheme on these psychometric properties extended to item banks where 1/3 to 1/4 of the items were differentially functioning. However, a recurring theme was that the magnitude of differences in average test characteristics as a function of weighting scheme were often minimal. There was thus little evidence that differential weighting of the objective function criteria resulted in tests with substantially stronger psychometric properties than EW.

Weighting scheme had a small effect on Unbiased-ATA's performance because of the positive correlations among the objective function criteria. As an example, consider the correlation matrix among γ_{DIF}, γ_{DTF}, $\gamma_{\text{Precision}}$, and γ_{IF} across the 54 conditions x 100 trials = 5400 objective function values when using EW. The sample correlations ranged from 0.072 ($r_{\text{DTF,IF}}$) to 0.465 ($r_{\text{DTF,Precision}}$). When criteria are positively correlated, then EW is expected to perform equivalently to differential weighting schemes (Newman, 1977; Wilks, 1938). Differential weighting should have a larger impact on the psychometric characteristics of the selected tests if alternative objective function criteria are used with Unbiased-ATA that introduce lower (or negative) correlations (von Winterfeldt & Edwards, 1986, Chapter 11). Thus, it is recommended that test developers examine the correlations among the objective function criteria (e.g.,

using a preliminary simulation or based on previous literature) prior to implementing Unbiased-ATA with differential weighting.

The low DIF percentages in the Study 2 item banks likely also affected the minimal differences between EW and DW or ROC. Both DW and ROC were operationalized to select items with less evidence of item- or test-level bias at the expense of higher Δ_{TIF} and item misfit. DW or ROC then prioritize anchor items in the bank that might also have slightly lower information or a significant $S - \chi^2$ statistic. Specifically, because DW equally weights the precision and item fit criteria, this weighting scheme will allow higher Δ_{TIF} to select more well-fitting items. ROC instead places higher weight on item precision and will select items to obtain lower Δ_{TIF} at the expense of item misfit. Given that there were lower DIF percentages in the banks, regularized DIF was less likely to incorrectly categorize invariant items as differentially functioning. Higher proportions of anchor items resulted in more equivalent item parameters between the groups. Then, differences among weighting schemes during item selection might have been based on only one or two items, translating to trivial differences when comparing the "best" test characteristics.

Consider next the case wherein regularized DIF mistakenly categorized more invariant items as differentially functioning. There are then fewer anchor items in the bank and thus more item parameter estimates will differ between the reference and focal groups. DW and ROC then have fewer anchor items to prioritize in item selection and will instead search among the $\text{DIF}_{\text{RegDIF}}$ items to minimize Δ_{TIF} and item misfit. DW and ROC still search for items that combine to have smaller uDTF values, but again the differences with EW stem from a handful of items at best.

When applying Unbiased-ATA in practice, the choice of weighting scheme largely depends on the test developers' goals. Certain test types or intended test goals might necessitate that higher importance is placed upon different objective function criteria. Prior knowledge of the item bank characteristics will likely also play a role in how the criteria are weighted. The results presented here suggest that in item banks with little evidence of differentially functioning items, placing higher weight upon the MI criteria (as with the DW and ROC weighting schemes) does not produce tests with substantially stronger psychometric properties than EW. Therefore, differential weighting of MI criteria might not be required if an item bank has been reviewed for DIF prior to the implementation of ATA. Higher weighting of the MI criteria might instead be beneficial if test developers are concerned about large item- and test-level bias within an item bank, especially during initial item bank reviews.

Determining the optimal set of weights for objective functions with multiple criteria is a robust research field (for a review, see Jia et al., 1998). Numerous alternative weighting schemes are possible, including rank-sum weights (Einhorn & McCoach, 1977; Stillwell et al., 1981). Moreover, different weighting schemes can outperform others in various testing scenarios (Y.-H. Chang & Yeh, 2001; Jia et al., 1998). Future research should thus explore Unbiased-ATA's performance with other weighting schemes and item bank characteristics before assuming that the negligible effects found in Study 2 extend to other conditions.

In addition to comparisons among weighting schemes, Study 2 provided more information regarding regularized DIF's performance. With DIF percentages less than or equal to 1/5, the method generally maintained Type I error rates below 0.10 while

simultaneously demonstrating power levels greater than 0.75. Only in a handful of conditions did $\overline{\text{FPR}}$ exceed 0.25, with $\overline{\text{FPR}}$ over 0.90 when there was large DIF in the "focal group" direction for 1/5 of the items and $N = 5000$. Combined with the Study 1 results, the current research demonstrates that regularized DIF is more likely to incorrectly categorize invariant items as differentially functioning when there are more truly differentially functioning items in the bank and $N > 1000$.

In summary, Study 2 further corroborates the supposition that Unbiased-ATA can select item combinations that demonstrate a reasonable balance of item- and test-level MI, TIF alignment, and item fit. Importantly, Unbiased-ATA's performance continues to depend on the item bank characteristics, such that inaccurate DIF categorizations and item parameter estimation can weaken the ATA-selected test properties. Study 2 also showed few practical advantages of using a differential weighting scheme with Unbiased-ATA. Based on the similar performances among the weighting schemes in these item banks, the Study 3 simulation used EW with Unbiased-ATA.

Chapter 5: Comparison to Alternative Algorithms

Thus far, Studies 1 and 2 have exclusively focused on Unbiased-ATA's performance in relation to (a) varying testing scenarios or (b) methods for assigning differential importance to the objective function criteria. Yet it remains to be seen how Unbiased-ATA works in relation to other objective functions. Numerous other ATA algorithms have been proposed in both the IRT and SEM literatures (e.g., Adema et al., 1991; G. A. Marcoulides & Drezner, 2004; K. M. Marcoulides, 2020; Raborn et al., 2020; Schultze & Eid, 2018). For example, ATA algorithms in IRT test construction often maximize the full TIF, or minimize the differences between the selected test's TIF and a target information function (e.g., Armstrong et al., 1998; van der Linden & Adema, 1998). Moreover, SEM researchers have combined indices of test score precision, model fit, and test-level MI with personality items (e.g., Jankowsky et al., 2020; Olaru & Danner, 2021). Indeed, Unbiased-ATA was constructed by combining and building upon elements of these previous algorithms.

Study 3 extends the previous simulation studies to compare Unbiased-ATA to previously established ATA algorithms. This comparison was designed to aid researchers' and practitioners' understanding of Unbiased-ATA's performance in the context of the extant literature. Specifically, this study addresses whether Unbiased-ATA improves upon other methods in quantifiable and meaningful ways.

Simulation Design

Objective Functions

Study 3 compared the Unbiased-ATA method (using equal weighting, see Equation 26) to two IRT-based objective functions. The first incorporated one criterion

measuring test score precision, operationalized as the deviation between the estimated and target TIF for the full sample. This type of algorithm has been commonly used in ATA with IRT test construction (e.g., Armstrong et al., 1998; van der Linden & Adema, 1998). Using the notation defined previously, the "TIF-only" objective function is

$$f(\text{TIF}) = 1 - \left(\sum_{k=1}^{K} \left| \left\{ \sum_{j=1}^{n_{\text{test}}} I_j(\theta_k) x_j \right\} - \text{TIF}_{\text{T}}(\theta_k) \right| \right). \tag{30}$$

Recall that the test score precision criterion in $f(\text{UATA})$ summed the TIF deviations across the two groups (Equation 14). To mirror objective functions used in previous IRT research, $f(\text{TIF})$ did not separate by group. In this case, a single-group IRT model was fit to the item bank data prior to algorithm implementation.

Unbiased-ATA was also compared to a modified version of the objective functions used in previous research integrating ATA with MI analyses in an SEM framework (Jankowsky et al., 2020; Olaru et al., 2018). In those studies, the researchers used an objective function evaluating (a) test-level MI based on changes in the CFI between MGCFA models, (b) test score precision with McDonald's ω (McDonald, 2013), and (c) structural validity with the RMSEA and SRMSR fit statistics. Study 3 transformed this objective function for use within an IRT perspective, drawing on the three related criteria from $f(\text{UATA})$ that respectively measured test-level MI, test score precision, and approximate model-data fit. Specifically, the transformed objective function without a criterion for item-level MI is

$$f(\text{DTF}) = (1/3)\gamma_{\text{DTF}} + (1/6)\gamma_{\text{Precision}} + (1/6)\gamma_{\text{IF}}, \tag{31}$$

where the criteria are as defined in Equations 11, 14, and 19. The weights assigned here were used to scale the full objective function value to the $[0,1]$ interval. Since $\gamma_{\text{Precision}}$ and γ_{IF} each consist of two terms, $1/3$ was replaced with $1/6$.

It merits comment that $f(\text{DTF})$ was not a direct replication of the objective functions used in previous research (e.g., Jankowsky et al., 2020; Olaru et al., 2018). For example, previous studies used model rather than item fit indices. However, Unbiased-ATA was compared to the objective function in Equation 31 to address whether incorporating a criterion for item-level MI improved the resulting test characteristics above and beyond using only a criterion for test-level MI.

Simulation Design and Procedure

The Study 3 simulation procedure mirrored that from Study 2, again using the design factors with the largest relative influences in Study 1. These design factors included (a) total sample size, (b) DIF magnitude, (c) DIF percentage, and (d) DIF percentage. Following item response and θ generation, the bank-level analyses (e.g., regularized DIF, IRT parameter estimation) were conducted on each item bank. Then, three ATA algorithms were run with Tabu search for each of the three objective functions—$f(\text{UATA})$, $f(\text{TIF})$, and $f(\text{DTF})$—as shown in Equations 26, 30, and 31. To account for the possibility of local minima solutions, Tabu search was repeated five times from random starting points for each algorithm. The dependent variables and analytic plan did not change between Study 2 and Study 3. Again, each of the 54 simulation conditions in Study 3 was repeated $R = 100$ times. All simulations and analyses were conducted in R statistical software, version 4.1.1 (R Core Team, 2021). The same R packages were used in Study 3 as in the previous studies.

Results

Regularized DIF Performance

The average FPR and TPR results for regularized DIF replicated from Study 2 to Study 3, which was expected given that the two simulations used the same item bank characteristics. $\overline{\text{FPR}}$ in Study 3 (see Figure A26 in the Appendix) increased as a function of DIF magnitude and sample size. When large DIF occurred for 1/5 of the items in the "focal group" direction with $N = 5000$, $\overline{\text{FPR}}$ reached 0.927. $\overline{\text{FPR}}$ was also notably high ($\overline{\text{FPR}} = 0.378$) for the same DIF percentage, DIF direction, and sample size but with a mixture of small and large DIF. Excluding these two conditions, $\overline{\text{FPR}}$ ranged from 0.005 to 0.178, with a median value of 0.040 and an interquartile range of 0.023 to 0.072. Furthermore, regularized DIF maintained moderate to high power with the Study 3 item banks (Figure A27), with a median $\overline{\text{TPR}}$ of 0.888 (interquartile range: 0.650 – 0.973). $\overline{\text{TPR}}$ were again smallest with small DIF magnitudes and $N = 500$ ($0.210 \leq \overline{\text{TPR}} \leq 0.397$). Taken together, regularized DIF maintained a desirable balance of Type I error and power in most testing scenarios with up to 12 differentially functioning items in the 60-item bank.

Comparing Tests Across ATA Objective Functions

Parameter Summaries. Figure A28 in the appendix presents the average $\hat{a}$ (Panel A) and $\hat{b}$ (Panel B) values among the items in the ATA-selected tests. $f(\text{UATA})$ and $f(\text{DTF})$ tended to select items with similar average $\hat{b}_{\text{R}}$ and $\hat{b}_{\text{F}}$ across the testing conditions. In the "both groups" direction, $f(\text{TIF})$ also produced tests with average $\hat{b}_{\text{R}}$ that were like those produced by the other two objective functions. Yet in the "focal group" direction (Columns 1, 3, and 5), $f(\text{TIF})$ selected items with lower average $\hat{b}_{\text{F}}$ than

$f(\text{UATA})$ or $f(\text{DTF})$. For example, in the "focal group" direction with $N = 5000$, the average $\hat{b}_{\text{R}}$ across DIF percentages and magnitudes were 1.030, 1.029, and 1.030 for $f(\text{UATA})$, $f(\text{TIF})$, and $f(\text{DTF})$, respectively. For the same conditions, the average $\hat{b}_{\text{F}}$ were 1.028, 0.971, and 1.027. Furthermore, there were few differences in average $\hat{a}$ among the objective function types. At times, $f(\text{UATA})$ selected item combinations with slightly lower $\hat{a}$ than $f(\text{TIF})$ or $f(\text{DTF})$, but there was considerable variation across the simulation conditions. Any pairwise difference in average $\hat{a}$ between two objective function types was at most 0.122.

Irrespective of objective function type, Figure A28 also reveals some variation in average $\hat{a}$ and $\hat{b}$ as a function of DIF characteristics and total sample size. For instance, average $\hat{a}$ generally decreased as sample size increased. Relatedly, when DIF modifications were made evenly to the reference and focal groups, there was a small inverse relationship between DIF percentage and average $\hat{b}$ among the selected items. However, differences in average $\hat{b}$ and $\hat{a}$ were small in magnitude, with average $\hat{b}$ ranging between 0.917 and 1.079 across conditions and average $\hat{a}$ ranging between 1.493 and 1.560.

Effect Sizes. Table 7 presents the η_p^2 values for the 11 fully crossed, multiway ANOVAs regressing the average test characteristics (e.g., number of differentially functioning items, uDTF effect size,) on the ATA objective function type, sample size, and the various DIF characteristics. Scanning the columns of Table 7, the objective function type—$f(\text{UATA})$, $f(\text{TIF})$, or $f(\text{DTF})$—accounted for substantial variation in many psychometric properties of the ATA-selected tests. Specifically, objective function type had medium to large effect sizes as a main effect in 10 out of the 11 models (Row

Table 7. Partial η^2 Effect Sizes When Regressing Test Properties on Objective Function, Sample Size, and DIF Characteristics

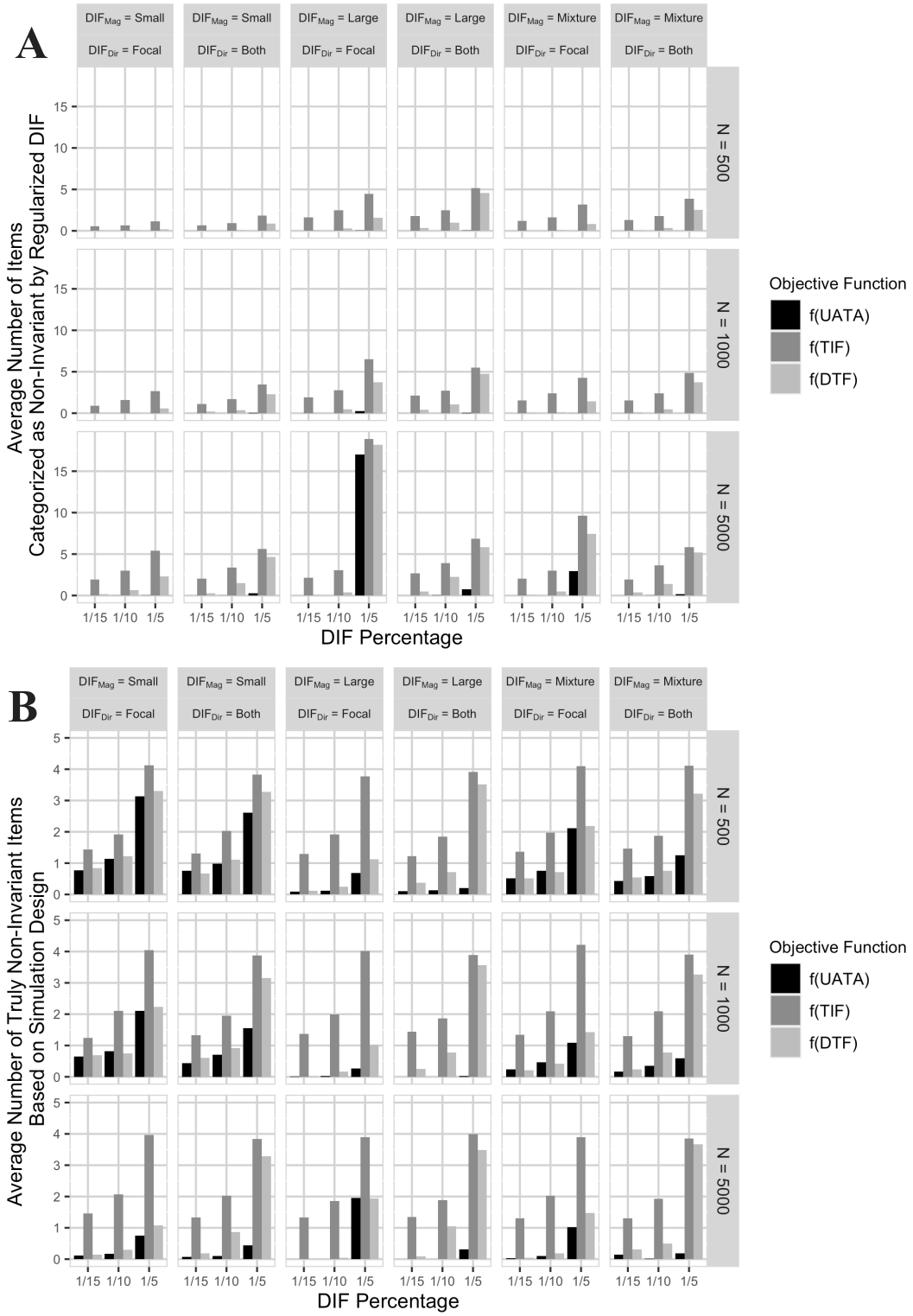
	Measurement Invariance					Information		Item Fit		Validity	
Design Factor	$n_{\text{No DIF}}$	uDTF	RMSEA	SRMSR_{R}	SRMSR_{F}	$\Delta_{\text{TIF}_{\text{R}}}$	$\Delta_{\text{TIF}_{\text{F}}}$	$n_{\text{Fitting}_{\text{R}}}$	$n_{\text{Fitting}_{\text{F}}}$	r_{R}	r_{F}
ATA Objective Function (ATA)	**0.71**	**0.99**	**0.97**	**0.69**	**0.94**	**0.98**	**0.98**	**1.00**	**1.00**	0.00	**0.16**
Total Sample Size (N)	**0.59**	**0.24**	**0.60**	**0.99**	**1.00**	0.11	0.04	0.08	**0.28**	**0.16**	0.04
DIF Percentage (Perc)	**0.76**	**0.87**	**0.92**	**0.54**	**0.81**	**0.86**	**0.87**	0.03	0.03	0.05	0.10
DIF Magnitude (Mag)	**0.46**	**0.86**	**0.82**	0.11	**0.71**	**0.92**	**0.85**	**0.13**	0.04	0.06	0.05
DIF Direction (Dir)	0.04	**0.67**	0.06	**0.61**	0.02	**0.58**	**0.51**	0.10	0.00	**0.18**	0.04
ATA x N	0.11	0.10	**0.16**	0.03	**0.56**	**0.45**	**0.14**	**0.16**	**0.47**	0.02	0.07
ATA x Perc	**0.35**	**0.86**	**0.78**	0.03	**0.72**	**0.57**	**0.77**	0.12	0.12	0.04	0.04
ATA x Mag	0.07	**0.88**	**0.79**	0.01	**0.70**	**0.70**	**0.84**	**0.26**	0.11	0.10	0.10
ATA x Dir	0.06	**0.81**	**0.62**	**0.23**	**0.42**	**0.16**	**0.82**	**0.16**	0.00	0.04	0.03
N x Perc	**0.59**	**0.16**	0.10	**0.39**	**0.64**	**0.16**	**0.21**	0.03	0.11	0.11	0.10
N x Mag	**0.22**	**0.14**	**0.15**	**0.19**	**0.51**	**0.22**	0.09	0.02	0.05	0.09	**0.30**
N x Dir	**0.25**	**0.30**	0.02	0.01	0.02	0.05	**0.25**	0.03	0.06	0.03	0.00
Perc x Mag	**0.49**	**0.52**	**0.64**	**0.17**	**0.65**	**0.41**	**0.56**	0.07	0.04	0.19	0.12
Perc x Dir	**0.18**	**0.52**	**0.16**	**0.57**	0.01	0.11	**0.35**	0.04	0.04	0.01	0.06
Mag x Dir	**0.16**	**0.47**	0.11	0.11	0.01	**0.16**	**0.34**	0.11	0.03	0.08	0.02
ATA x N x Perc	0.04	0.10	0.12	0.10	**0.25**	**0.31**	0.10	0.05	**0.22**	0.03	0.02
ATA x N x Mag	0.05	**0.39**	**0.13**	0.12	0.12	**0.30**	**0.19**	0.03	0.08	**0.29**	0.09
ATA x N x Dir	0.01	**0.22**	0.06	0.03	**0.23**	**0.18**	**0.34**	0.08	0.10	0.01	0.03
ATA x Perc x Mag	0.03	**0.28**	**0.36**	0.03	**0.37**	**0.31**	**0.27**	**0.18**	0.09	**0.14**	0.07
ATA x Perc x Dir	0.04	**0.57**	**0.54**	0.03	**0.52**	**0.35**	**0.66**	0.06	0.09	0.08	0.11
ATA x Mag x Dir	0.01	**0.50**	**0.29**	0.02	**0.21**	0.07	**0.59**	**0.22**	0.05	0.06	0.06
N x Perc x Mag	**0.43**	**0.20**	**0.23**	**0.41**	**0.52**	**0.15**	**0.13**	0.08	0.12	0.15	0.12
N x Perc x Dir	**0.46**	**0.26**	0.07	**0.16**	0.12	0.03	**0.16**	0.01	0.01	**0.13**	0.12
N x Mag x Dir	**0.28**	0.09	0.09	**0.29**	0.02	0.01	0.05	0.11	0.10	**0.23**	**0.15**
Perc x Mag x Dir	**0.33**	**0.35**	0.03	**0.42**	0.01	0.08	**0.23**	0.02	0.03	**0.20**	0.04
R^2	0.937	0.992	0.984	0.999	1.000	0.988	0.990	0.997	0.998	0.747	0.701
Adjusted R^2	0.840	0.981	0.960	0.997	0.999	0.969	0.975	0.993	0.995	0.363	0.249

Note. Effect sizes greater than or equal to 0.13 are bolded (Cohen, 1992).

1). Indeed, the η_p^2 values even reached 1.00 for the number of well-fitting items (indicating that the objective function type accounted for 100% of the variation when controlling for the other design factors). The relationship between the objective function type and test-level MI (Columns 3 – 6) or TIF deviations (Columns 7 – 8) depended on the DIF characteristics in the bank. However, there were fewer notable interactions between objective function type and DIF characteristics for item fit (Columns 9 – 10). Notice also in Table 7 that the effect sizes for objective function type were relatively smaller for the external validity coefficients (Columns 11 – 12). These ANOVA models also had the smallest relative R^2 values, indicating that the design factors accounted for less variation in the validity coefficients as compared to the dependent variables.

Item- and Test-Level MI. Figure 29A displays the average number of items in the ATA-selected tests that were categorized as differentially functioning (i.e., non-invariant) by regularized DIF ($\text{DIF}_{\text{RegDIF}}$). $f(\text{TIF})$ selected more $\text{DIF}_{\text{RegDIF}}$ items on average than either $f(\text{UATA})$ or $f(\text{DTF})$, and $f(\text{DTF})$ generally selected more $\text{DIF}_{\text{RegDIF}}$ items than $f(\text{UATA})$. For example, $f(\text{TIF})$ tests comprised approximately three more $\text{DIF}_{\text{RegDIF}}$ items on average than $f(\text{UATA})$ tests, whereas tests from $f(\text{DTF})$ and $f(\text{UATA})$ differed by approximately one $\text{DIF}_{\text{RegDIF}}$ item on average. Both $f(\text{TIF})$ and $f(\text{DTF})$ selected more $\text{DIF}_{\text{RegDIF}}$ items with higher DIF percentages, larger DIF magnitudes, and larger total sample sizes. On the other hand, $f(\text{UATA})$ consistently selected the fewest $\text{DIF}_{\text{RegDIF}}$ items; only in conditions with regularized DIF $\overline{\text{FPR}}$ greater than 0.30 did the average number of $\text{DIF}_{\text{RegDIF}}$ items for $f(\text{UATA})$ surpass 1.00. Fewer $\text{DIF}_{\text{RegDIF}}$ items for $f(\text{UATA})$ is understandable given that it was the only objective function to explicitly penalize item combinations based on regularized DIF

Figure 29. Average Number of Differentially Functioning Items in the Selected Tests Across Objective Function Types



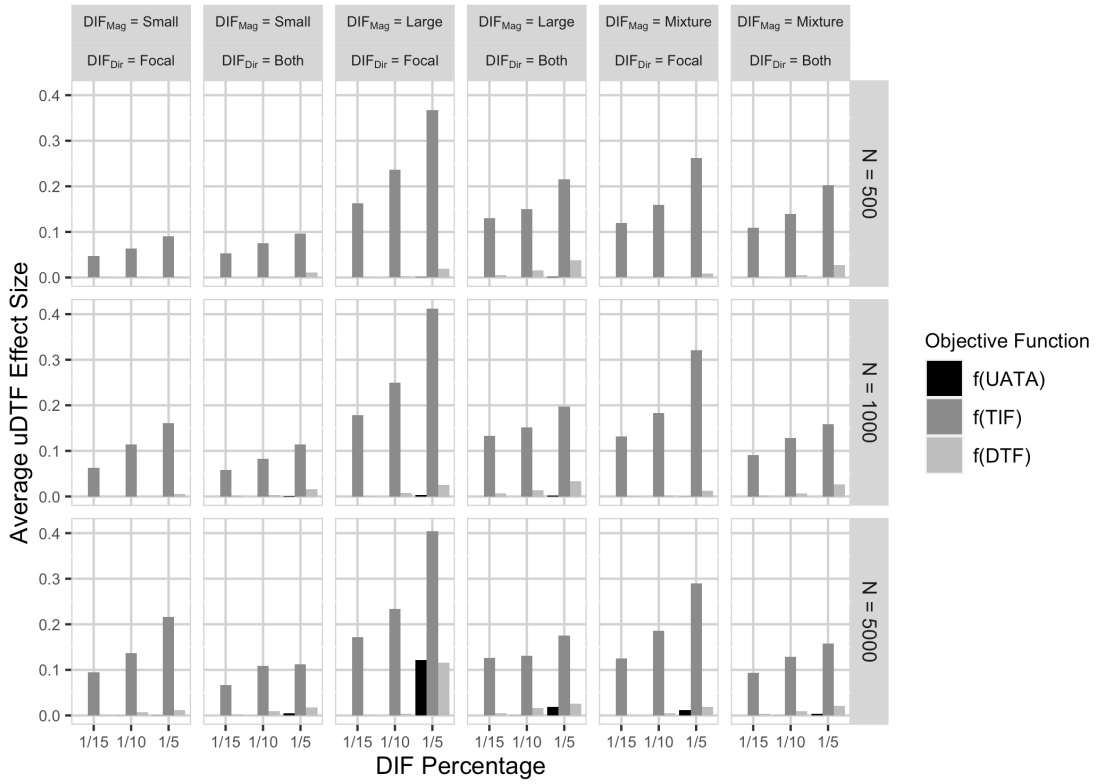
categorizations. Comparing the panels in Row 3, Column 3 of Figures 29A and A26, all three objective function types selected tests with the most $\mathrm{DIF}_{\mathrm{RegDIF}}$ items on average when $\overline{\mathrm{FPR}}$ was highly inflated.

The average number of truly non-invariant items ($\mathrm{DIF}_{\mathrm{True}}$) in each test is presented in the bottom panel of Figure 29. $f(\mathrm{TIF})$ tests consistently comprised the most $\mathrm{DIF}_{\mathrm{True}}$ items among the objective function types; for $f(\mathrm{TIF})$, the average number of $\mathrm{DIF}_{\mathrm{True}}$ items increased only as a function of DIF percentage. Comparing $f(\mathrm{TIF})$ to $f(\mathrm{UATA})$, differences in the average number of selected $\mathrm{DIF}_{\mathrm{True}}$ items ranged from 0.550 to 3.860, with larger differences occurring with large DIF in the "both groups" direction. Furthermore, $f(\mathrm{DTF})$ selected more $\mathrm{DIF}_{\mathrm{True}}$ items than $f(\mathrm{UATA})$ in conditions with higher DIF percentages, large DIF magnitudes, or the "both groups" direction; otherwise, $f(\mathrm{DTF})$ and $f(\mathrm{UATA})$ selected similar numbers of $\mathrm{DIF}_{\mathrm{True}}$ items, on average. Differences between $f(\mathrm{UATA})$ and either of the other objective function types did not exceed 3.86 items on average. Across all conditions, the maximum average numbers of $\mathrm{DIF}_{\mathrm{True}}$ items were 3.14, 4.21, and 3.67 for $f(\mathrm{UATA})$, $f(\mathrm{TIF})$, and $f(\mathrm{DTF})$, respectively.

As in Study 2, $f(\mathrm{UATA})$ selected relatively more $\mathrm{DIF}_{\mathrm{True}}$ items when DIF magnitudes and sample sizes were smaller (e.g., Columns 1 – 2, Row 1). These conditions paralleled those with lower regularized DIF $\overline{\mathrm{TPR}}$. Lower power translated to higher proportions of $\mathrm{DIF}_{\mathrm{True}}$ items that were categorized as anchor items. With a larger pool of anchor items, $f(\mathrm{UATA})$ might select more $\mathrm{DIF}_{\mathrm{True}}$ items that simultaneously minimize the other objective function properties (e.g., TIF deviations). The average number of $\mathrm{DIF}_{\mathrm{True}}$ items for $f(\mathrm{UATA})$ also increased as a function of DIF percentage, although that relationship was stronger with smaller DIF magnitudes and sample sizes.

In tests selected by $f(\text{TIF})$, greater evidence of item bias translated to larger uDTF effect sizes. Figure 30 shows that $\overline{\text{uDTF}}$ for $f(\text{TIF})$ tests ranged from 0.048 to 0.413 and were positively associated with DIF percentage. For example, $\overline{\text{uDTF}}$ for $f(\text{TIF})$ were 0.109, 0.148, and 0.219 for DIF percentages of 1/15, 1/10, and 1/5, respectively. $f(\text{TIF})$ tests also demonstrated higher $\overline{\text{uDTF}}$ with larger DIF magnitudes and in the "focal group" direction compared to the "both groups" direction. In contrast, there were smaller differences in $\overline{\text{uDTF}}$ between tests selected by either $f(\text{DTF})$ or $f(\text{UATA})$. $f(\text{DTF})$ selected tests with slightly higher $\overline{\text{uDTF}}$ when there was large DIF for 1/5 of the items in the bank, but differences in $\overline{\text{uDTF}}$ between $f(\text{DTF})$ and $f(\text{UATA})$ did not exceed 0.036. Comparing Figure 30 to Figure 29, greater differences in the average

Figure 30. Average uDTF Effect Size for Selected Tests Across Objective Function Types



number of non-invariant items between $f(\text{TIF})$ and $f(\text{UATA})$ were associated with larger differences in $\overline{\text{uDTF}}$. Yet differences between $f(\text{DTF})$ and $f(\text{UATA})$ in the average number of non-invariant items did not translate to demonstrable differences in $\overline{\text{uDTF}}$.

Test-level MI was also evaluated using the full-sample RMSEA and group-level SRMSR statistics when fitting the test data to a strong MI model. $\overline{\text{RMSEA}}$ was below Hu and Bentler's (1999) threshold for "good" fit across all examined conditions (Figure A29). $f(\text{TIF})$ selected tests with higher $\overline{\text{RMSEA}}$ than $f(\text{DTF})$ or $f(\text{UATA})$, particularly with larger DIF magnitudes. $\overline{\text{RMSEA}}$ was generally similar between $f(\text{DTF})$ and $f(\text{UATA})$, although $f(\text{DTF})$ selected tests with slightly higher fit statistics with large DIF in the "both groups" direction. $f(\text{DTF})$ also selected more differentially functioning items in these conditions. However, the magnitude of differences among the objective function types was small, never exceeded 0.018 between any two algorithms. Such differences are unlikely to impact categorical decisions of "good" or "bad" model fit.

There were even smaller differences in $\overline{\text{SRMSR}}_{\text{R}}$ and $\overline{\text{SRMSR}}_{\text{F}}$ among the three objective function types. Figure A32 demonstrates that $f(\text{TIF})$ at times selected items with slightly higher $\overline{\text{SRMSR}}_{\text{R}}$ and $\overline{\text{SRMSR}}_{\text{F}}$ than $f(\text{DTF})$ or $f(\text{UATA})$, but differences never exceeded 0.007 in magnitude. When $N \geq 1000$, most $\overline{\text{SRMSR}}$ values represented "good" fit for a strong MI model irrespective of objective function type; the exceptions were for $f(\text{TIF})$ when 1/5 of the items were differentially functioning. As in the previous studies, $\overline{\text{SRMSR}}_{\text{F}}$ were uniformly smaller than $\overline{\text{SRMSR}}_{\text{F}}$. Additionally, both $\overline{\text{SRMSR}}_{\text{R}}$ and $\overline{\text{SRMSR}}_{\text{F}}$ again decreased as sample size increased.

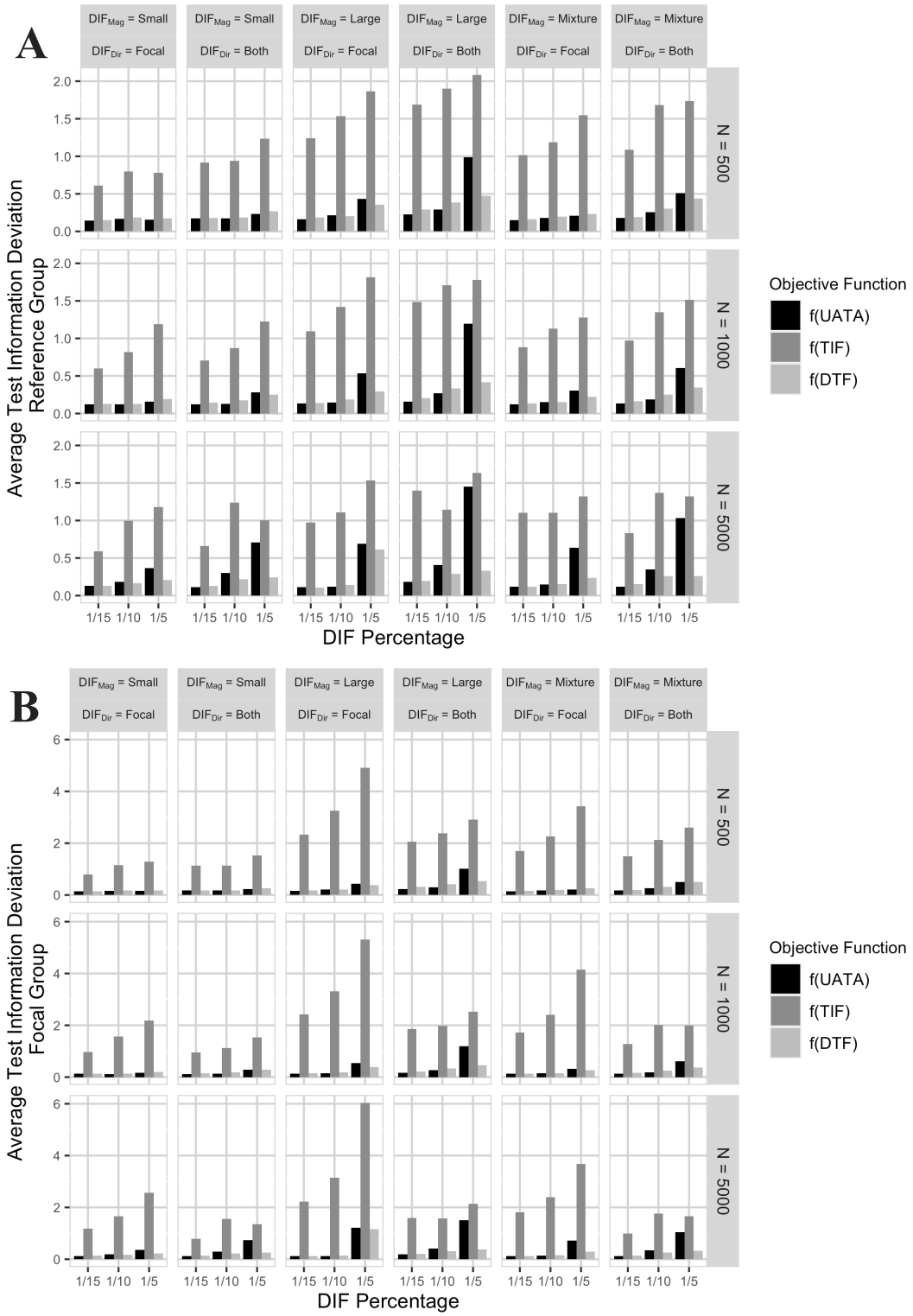
For comparison, Figures A30 – A31 and A33 – A34 show the $\overline{\text{RMSEA}}$ and $\overline{\text{SRMSR}}$ values, respectively, when fitting the test data to weak and configural MI models.

Across these figures, there were negligible differences in the average fit statistics among the three objective function types. As in the previous studies, total sample size most strongly influenced $\overline{\text{RMSEA}}$ and $\overline{\text{SRMSR}}$ for both weak and configural MI models, with average model fit statistics more likely to align with the "good" fit categorization for the given model as total sample size increased.

Test Information. Although $f(\text{TIF})$ was designed to minimize Δ_{TIF} for the full sample, this objective function consistently selected tests with the largest group-level $\bar{\Delta}_{\text{TIF}}$. Figure 31 reveals that $f(\text{TIF})$ selected tests with higher $\bar{\Delta}_{\text{TIF}}$ as a function of DIF percentage and magnitude. $f(\text{TIF})$ tests also demonstrated higher $\bar{\Delta}_{\text{TIF}_\text{R}}$ and $\bar{\Delta}_{\text{TIF}_\text{F}}$ than tests selected by $f(\text{UATA})$: marginalizing across other design factors, differences between $f(\text{TIF})$ and $f(\text{UATA})$ ranged from 0.187 to 1.608 for $\bar{\Delta}_{\text{TIF}_\text{R}}$ and from 0.620 to 4.822 for $\bar{\Delta}_{\text{TIF}_\text{F}}$. Moreover, $f(\text{UATA})$ tests sometimes produced higher $\bar{\Delta}_{\text{TIF}}$ than $f(\text{DTF})$ tests, especially when there was more DIF in the "both groups" direction with larger sample sizes. In these conditions, $f(\text{UATA})$ also selected fewer differentially functioning items, prioritizing reductions in item bias at the expense of higher Δ_{TIF}.

Notice also in Figure 31 that $f(\text{TIF})$ tended to select tests with higher $\bar{\Delta}_{\text{TIF}_\text{F}}$ than $\bar{\Delta}_{\text{TIF}_\text{R}}$. The differences between $\bar{\Delta}_{\text{TIF}_\text{R}}$ and $\bar{\Delta}_{\text{TIF}_\text{F}}$ were smaller for tests selected by $f(\text{UATA})$ or $f(\text{DTF})$. Looking back at the parameter summaries for the ATA-selected tests, $f(\text{TIF})$ tests often comprised items with lower average $\hat{b}_F$ than $\hat{b}_R$. Group-level differences in $\hat{b}$ might have produced a test where the focal group TIF was substantially more different from the target function. Finally, there were minimal differences in full TIF values across the objective function types (Figure A35).

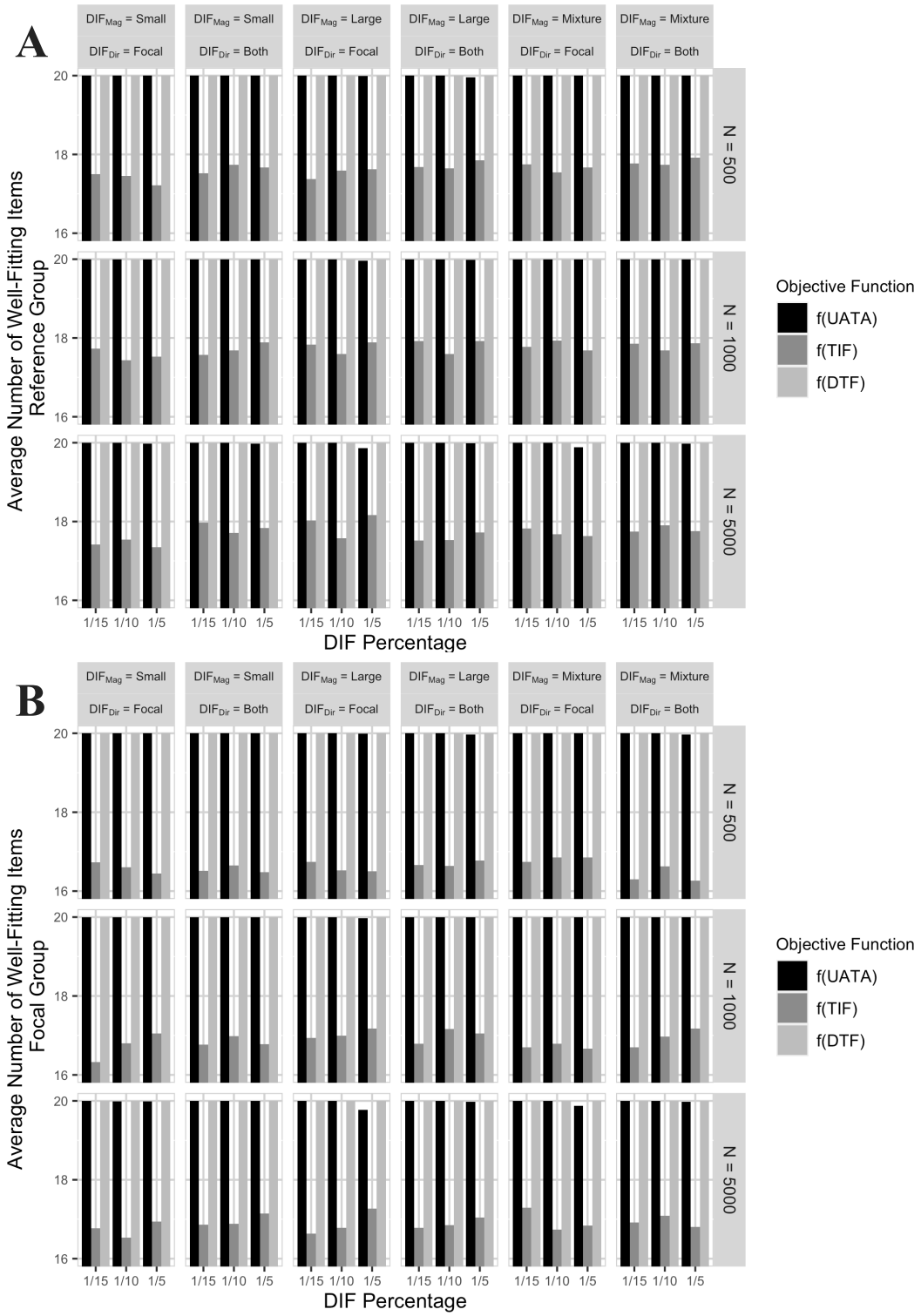
Figure 31. Average Test Information Function Deviations for Selected Tests Across Objective Function Types



Item Fit. Figure 32 indicates noticeable differences in the number of well-fitting items between $f(\text{TIF})$ and either $f(\text{UATA})$ or $f(\text{DTF})$. On average, $f(\text{TIF})$ selected tests with 16.26 to 18.16 well-fitting items. $f(\text{UATA})$ tests comprised at least 19.77 well-fitting items on average, whereas $f(\text{DTF})$ consistently selected tests wherein all 20 items demonstrated good fit. There were negligible differences in item fit between $f(\text{UATA})$ and $f(\text{DTF})$ across the examined testing scenarios; these objective functions also selected tests with similar average numbers of well-fitting items between groups. On the contrary, $f(\text{TIF})$ generally selected fewer well-fitting items for the focal group ($\overline{\text{Fit}}_{\text{F}} = 16.79$, Range: 16.26 – 17.29) than the reference group ($\overline{\text{Fit}}_{\text{R}} = 17.70$, Range: 17.22 – 18.16). Given that $f(\text{TIF})$ selected more differentially functioning items than the other objective functions, a higher proportion of items had differing parameter estimates between groups. Lower average $\hat{b}_{\text{F}}$ among the selected items seemingly translated to worse item fit for the focal group.

External Validity Coefficients. Finally, the three objective functions were compared on the correlations between the selected tests and an external criterion measure (Figure A36). There were few notable trends in $\bar{r}$ among the objective functions. Across groups, the ranges in $\bar{r}$ were 0.311 – 0.349, 0.309 – 0.350, and 0.311 – 0.351 for $f(\text{UATA})$, $f(\text{TIF})$, and $f(\text{DTF})$, respectively. Differences in $\bar{r}$ among the objective functions did not exceed 0.026, and each objective function selected tests with higher $\bar{r}$ in various conditions. Regardless of objective function type, the external validity coefficients were higher on average for the focal group ($\bar{r}_{\text{F}} = 0.339$) compared to the reference group ($\bar{r}_{\text{R}} = 0.321$).

Figure 32. Average Number of Well-Fitting Items for Selected Tests Across Objective Function Types



Discussion

Study 3 directly compared Unbiased-ATA to two objective functions based on the extant literature. The study goal was to gauge whether Unbiased-ATA selected 20-item tests with substantially different psychometric properties than ATA algorithms that either focused solely on full-sample TIF deviations or did not account for item bias. In other words, did Unbiased-ATA improve the psychometric strength of the resulting tests above and beyond these other objective functions?

When there were differences among the three objective function types, $f(\text{TIF})$ consistently selected tests with relatively weaker psychometric properties than $f(\text{UATA})$ and $f(\text{DTF})$. On average, tests with $f(\text{TIF})$ had (a) more differentially functioning items, (b) greater differential test functioning as evidenced by the uDTF effect size, and (c) more items that did not fit the IRT model. These trends were expected given that the $f(\text{TIF})$ objective function did not explicitly account for these properties. Interestingly, tests with $f(\text{TIF})$ also demonstrated the highest group-level $\bar{\Delta}_{\text{TIF}}$ values. $f(\text{TIF})$ will select more differentially functioning items that align well with the full-sample TIF when the items are assumed to be invariant. However, these items might have substantially different information functions at the group level. If items in the bank are differentially functioning, selecting items to minimize the full-sample TIF deviation can thus mask important information for the different groups.

$f(\text{UATA})$ and $f(\text{DTF})$ generally selected tests with more similar psychometric properties, although some interesting differences emerged. For one, $f(\text{UATA})$ tests comprised fewer differentially functioning items on average (reflecting the inclusion of γ_{DIF} in the objective function). However, the $f(\text{UATA})$ and $f(\text{DTF})$ tests did not largely

differ on the uDTF and strong MI model fit statistics. More evidence of item level MI for $f(\text{UATA})$ therefore did not translate to substantially more evidence of test level MI. Furthermore, $\bar{\Delta}_{\text{TIF}}$ values were slightly higher for $f(\text{UATA})$ tests compared to $f(\text{DTF})$. When including γ_{DIF}, it is likely that $f(\text{UATA})$ more often prioritized selecting anchor items at the expense of TIF deviations.

Although the objective function type affected the psychometric properties among the ATA-selected tests, it remains to be seen whether these results translate to meaningful differences in application. The Study 3 simulation provides evidence that $f(\text{UATA})$ generates tests with a greater balance of test score accuracy and precision than $f(\text{TIF})$. Such balance can be advantageous in many testing scenarios, especially when test developers are wary of differing measurement models among examinee subgroups. Yet the simulation results preclude a robust recommendation for using $f(\text{UATA})$ over $f(\text{DTF})$, or vice versa. It appears that incorporating a criterion for item-level bias in the objective function slightly improves item-level MI but does not definitively improve test-level MI; relatedly, the reduction in item-level bias can come at the expense of other psychometric properties. Further research is necessary to compare $f(\text{UATA})$ and $f(\text{DTF})$ in item banks with other characteristics (e.g., additional percentages and magnitudes of DIF, alternative item types, etc.).

While $f(\text{UATA})$ selected fewer differentially functioning items than $f(\text{TIF})$ and $f(\text{DTF})$ in most conditions, there were smaller differences among the objective function types for differential test functioning. Indeed, all three objective functions consistently selected items that together demonstrated good fit (Hu & Bentler, 1999) for a model with equivalent b and a parameters between groups. The item bank characteristics in Study 3

likely engendered these discrepancies between item- and test-level MI. Specifically, with relatively low proportions of DIF in the item bank, incorporating a few more differentially functioning items (particularly those with minimal differences in group-level item parameter estimates) might not noticeably affect differential test functioning. Larger differences in test-level MI might have occurred with higher proportions or magnitudes of DIF in the item bank. The relationship between differential item and test functioning with Unbiased-ATA is discussed further in the next chapter.

Chapter 6: General Discussion and Conclusions

Summary

ATA is an efficient, flexible method for constructing a psychosocial test based on one or more predefined psychometric properties. Numerous ATA algorithms have been proposed in the extant literature (e.g., Adema et al., 1991; Diao & van der Linden, 2011; van der Linden, 1998; van der Linden & Adema, 1998), with objective functions formulated to maximize indices of test score precision (e.g., test information, reliability coefficient estimates) and test score accuracy (e.g., model fit indices, external validity coefficients). However, fewer ATA algorithms have sought to explicitly reduce evidence of item- and test-level bias among the selected items (e.g., Jankowsky et al., 2020; Olaru et al., 2018, 2019; Olaru & Jankowsky, 2021). The current study evaluated a novel ATA objective function that applies an IRT framework to select items with a balance of (a) item-level MI, (b) test-level MI, (c) test score precision, and (d) item fit.

Study 1 found that Unbiased-ATA functioned as intended by constructing tests with evidence of both test score accuracy and precision, even when tests with weaker psychometric properties were possible. Unbiased-ATA's performance depended on both the accuracy of the DIF detection method and the similarities of the estimated item parameters between the reference and focal groups. If the DIF method both incorrectly categorizes invariant items (i.e., high FPRs) and correctly categorizes non-invariant items (i.e., high TPRs), Unbiased-ATA selects from a smaller pool of anchor items. The resulting test then comprises more differentially functioning items. When the estimated item parameters for the reference and focal group are markedly different, the resulting test will have weaker evidence of test-level MI and other group-level indices. In the

current study, regularized DIF and IRT-LRT often produced roughly similar FPRs and TPRs. Yet certain psychometric properties (e.g., item fit and TIF deviations) were stronger when Unbiased-ATA was paired with IRT-LRT because the selected items had more similar parameter estimates between groups. Granted, these test comparisons rely on the ways that the current study operationalized "psychometric strength" of ATA-selected tests. Specifically, group-level indices were commonly used to characterize "good" measurement characteristics. The trends might look different, or altogether disappear, if other psychometric properties (e.g., full-sample TIF deviations) are used.

The Study 1 results highlight the importance of using an accurate DIF detection method in conjunction with Unbiased-ATA. Ideally, the detection method should demonstrate an appropriate balance of false and true positives so that most items in the bank are correctly categorized as either anchors or differentially functioning. Yet in practice, test developers do not know the true status of an item and cannot compute indices like false positive rates. Previous empirical research (e.g., how often similar items have been identified as differentially functioning in other test administrations) and simulation studies can build support for or against using a particular detection method. An iterative DIF review process might also be warranted. Although regularized DIF was only applied once in the bank-level analyses, test developers might apply the method multiple times to screen out potential false positives or practically insignificant DIF. Indeed, combining a categorical DIF detection method with a related effect size (e.g., the wABC; Edelen et al., 2015) can aid test developers in gauging DIF's practical impact.

Study 2 revealed that differential weighting did not strongly influence Unbiased-ATA within the examined item banks. There was some evidence that (a) EW was

associated with higher uDTF values, (b) DW was associated with higher group-level TIF deviations, and (c) ROC weighting was associated with more item misfit. These differences were slightly exacerbated as percentage of DIF increased, but overall were small in magnitude and should not be over-interpreted. In fact, even when a quarter of the items in the bank were differentially functioning, no one weighting scheme produced tests with substantially stronger psychometric characteristics.

In ATA, the weighting scheme largely depends upon either the criterion characteristics, the test developers' goals, or a combination of these factors. The item bank characteristics also impact both the decision to use a particular weighting scheme and the properties of the corresponding test. For example, differential weighting that places greater importance on MI criteria can be advantageous when test developers hypothesize that an item bank contains many differentially functioning items. Relatedly, overweighting one or more MI criteria might be considered when high FPRs from the DIF detection method are of concern (although note that the Study 2 results did not find strong evidence in favor of DW or ROC in such conditions). The EW, DW, and ROC weighting schemes (and others, like rank-sum weights; Einhorn & McCoach, 1977; Stillwell et al., 1981), should be examined within item banks comprising different types and proportions of DIF, as well as when comparing tests on other psychometric properties.

The final study compared Unbiased-ATA to alternative objective functions that have been proposed in the extant literature. Compared to an objective function that minimized full-sample Δ_{TIF}, Unbiased-ATA generally selected tests with stronger evidence of test score accuracy and precision (as operationalized by the characteristics in

the current study). Certain trends here were expected given how the two objective functions were defined. For example, $f(\text{TIF})$ did not account for item- or test-level MI, and therefore selected more differentially functioning items on average than Unbiased-ATA. More differentially functioning items with $f(\text{TIF})$ led to higher group-level Δ_{TIF} as well. Yet there were negligible differences in the full TIF values between the two objective functions, either at the group level or for the full sample. Hence, focusing solely on full-sample test information can mask important group-level differences in the measurement models. For instance, substantial differences in Δ_{TIF} among groups suggest that the test is not providing similar information along the θ continuum for all examinees. Test developers should be aware of these deviations if the test goal is to provide sufficient information for certain θ ranges.

Unbiased-ATA was also compared to an objective function that excluded the item-level bias criterion (γ_{DIF}). On average, tests compiled by either Unbiased-ATA or $f(\text{DTF})$ had very similar psychometric properties. There was evidence that Unbiased-ATA selected items with slightly higher Δ_{TIF}, presumably to minimize the number of differentially functioning items. Still, these differences between Unbiased-ATA and $f(\text{DTF})$ were small in magnitude and did not translate to demonstrably dissimilar tests. The similarities between Unbiased-ATA and $f(\text{DTF})$ suggest that including a criterion for item-level MI does not substantially improve ATA performance above and beyond a criterion for test-level MI. Further research with other item banks and evaluative psychometric characteristics is necessary to better parse apart the differences between Unbiased-ATA and $f(\text{DTF})$.

Across the three studies, evidence of item-level bias did not always translate to substantial evidence of test-level bias. Specifically, the test data often fit a strong MI model even when most items were categorized as differentially functioning. In these cases, variation in uDTF values was also small; for instance, test scores between the reference and focal group differed by no more than 6.819% on average in Study 1. Item parameter estimation likely played a role, such that there were often small group-level differences in $\hat{b}$ and $\hat{a}$ in tests with many differentially functioning items. Indeed, even when large DIF was simulated, most estimated wABC values aligned with a "small DIF" categorization. Study 1 also showed that relatively fewer truly invariant items were selected with ATA, leading to more similar parameter estimates even if these items were categorized by regularized DIF as differentially functioning. Test-level bias might have been further minimized since item-level differences in alternating directions can "cancel out" when congregated at the test level (Chalmers et al., 2016).

Even when tests produced higher uDTF values, such as $f(\text{TIF})$ in Study 3, the test data still consistently fit a strong MI model. Note that the uDTF takes a different approach to identifying test-level bias than global fit statistics like the RMSEA or SRMSR. Specifically, whereas the uDTF sums item score functions across a span of θ values, the RMSEA and SRMSR are limited information statistics computed using interitem correlations (Maydeu-Olivares, 2015; Maydeu-Olivares & Joe, 2014). As shown in this research, relying solely on correlations can obscure differences in score functions at both the item and test levels. It also merits comment that global fit statistics like RMSEA and SRMSR are better used as guidelines rather than definitive markers of "best" fit (Lai & Green, 2016; Marsh et al., 2005). Indeed, the Study 1 results

demonstrated that these model fit statistics could contradict one another; whereas average RMSEA values aligned with "good" fit for strong MI models, the corresponding average SRMSR values often indicated "poor" fit. Therefore, multiple indices of test-level bias should be incorporated in the test review process to best understand the extent to which measurement models differ among groups. More research into appropriate fit thresholds for dichotomous IRT models might also be warranted.

Taken together, the results from the three studies suggest that Unbiased-ATA is a useful tool for automated latent trait test construction. Paired with a reliable DIF detection method, Unbiased-ATA appears particularly advantageous in the early stages of test development with item banks where DIF is hypothesized to occur. Specifically, Unbiased-ATA affords an efficient item review: by selecting a subset of items with the "best" psychometric properties among the bank, test developers reduce the number of potentially biased items that they need to evaluate. Scenarios where evidence of item bias is not reflected at the test level also suggest that Unbiased-ATA can help identify items with non-practically significant DIF (Edelen et al., 2015). Test developers can then decide how to revise (or potentially remove) those items for future use. Regardless of how Unbiased-ATA is used, test properties should be validated with other samples of examinees to ensure generalizability (Goetz et al., 2013).

The current studies also provided new insight into the performance of regularized DIF (Belzak & Bauer, 2020). The method demonstrated an acceptable balance of Type I error and power rates when 1/5 or fewer of the 60 items in the bank were differentially functioning and $N \leq 1000$, particularly when DIF was balanced (i.e., parameter modifications were made evenly to both groups). However, in other contexts regularized

DIF produced wildly inflated FPRs (i.e., FPR > 0.75). Researchers should thus be cautious about applying regularized DIF to test data (a) with sample sizes greater than 1,000 or (b) when large proportions of DIF are anticipated. Estimates from previous empirical research can help test developers identify whether regularized DIF would be appropriately applied to a given item bank. Given that IRT-LRT also demonstrated high FPRs in similar contexts, certain item bank characteristics (e.g., when more than 1/3 of the items are differentially functioning) might be associated with inaccurate item categorizations irrespective of the DIF detection method. An iterative process wherein the DIF detection method is applied, items are reviewed, and the detection method is re-applied can help test developers better screen for false positive results.

Regularized DIF is a nascent method in the DIF literature, and additional research with alternative item bank characteristics (e.g., polytomous items, different item bank sizes) is necessary. Relatedly, recall that test sum scores were used as proxies for θ to reduce computational burden (by essentially skipping the Expectation step of the EM algorithm). Future work should examine whether the results of the current research replicate when θ estimation is incorporated into the regularized DIF procedure. Additionally, although the BIC was used for model selection with regularized DIF, alternative criteria are possible, including the weighted average information criterion (WIC; Wu & Sepulveda, 1998). Moreover, researchers have recently proposed other variants of regularization for DIF detection (S. M. Chen et al., 2021; Y. Chen et al., 2021; C. Wang et al., 2022). These methods, and others that do not require initial anchor item selection (e.g., Yuan et al., 2021), warrant consideration for future research with Unbiased-ATA.

Limitations

Although the three simulation studies were designed to reflect realistic testing scenarios, the simulated data cannot represent all item and examinee characteristics seen in psychosocial measurement. Numerous avenues for future research remain, including combining Unbiased-ATA with (a) larger item banks, (b) polytomous item data, (c) more than two groups, (d) multidimensional IRT models, and (e) alternative DIF characteristics (e.g., other DIF percentages or magnitudes). Namely, other item banks should be examined. Whereas the current study used an item bank with increased information around $\theta = 1$, the parameters were limited in that $b \geq 0$. The current results would likely change with other item and person parameter distributions, as well as with a larger number of items. Moreover, the current studies operationalized DIF as group differences in item parameters while maintaining unidimensional θ. Researchers might instead incorporate DIF by simulating differences in θ dimensionality among groups. The Unbiased-ATA objective function can also be modified to encompass alternative psychometric properties. An advantage of ATA algorithms is that they afford substantial flexibility in defining the criteria used to select item combinations.

More work is also necessary to better understand differences among the algorithm types. Study 1 only examined three algorithms—0-1 LP, ACO, and Tabu search—based on their promising performance in previous research. Yet other algorithms are possible, including metaheuristic algorithms like simulated annealing (Cerny, 1985; Kirkpatrick et al., 1983) and genetic algorithms (Fraser, 1957; D. E. Goldberg, 1989; Yarkoni, 2010). Machine learning methods might also be applicable; for example, Glockner et al. (2020) applied neural networks to construct short-form personality tests with high predictive

validity. Furthermore, each of the examined algorithms in this study were implemented with various controls (e.g., number of iterations, number of ants in ACO) based on previous ATA literature. Future research should examine the reproducibility of the current results when modifying these algorithm controls.

Importantly, Study 1 demonstrated that Tabu Search performed relatively well compared to 0-1 LP and ACO. A potential limitation of this method, however, is that the algorithm can identify different "best" item combinations depending on its starting point. To account for the possibility of local minima solutions, Tabu search was run five times from random starting points (Jankowsky et al., 2020) in each application. Still, more work is warranted to better understand Tabu search's performance in relation to local minima.

Numerous limitations of and future directions for Unbiased-ATA within the context of test fairness also require exploration. In the current research, item bias was operationalized as DIF, meaning that individuals from different groups with the same θ have different probabilities of endorsing an item (Thissen et al., 1988, 1993). Yet DIF is arguably an imperfect form of measuring bias since items are "unreliable [measures] of the construct of interest" (Dorans, 2017, p. 223; as cited in Poe & Elliot, 2019). Therefore, relying solely on DIF categorizations as evidence for or against test bias is a limited approach. It is incumbent upon test developers to thoroughly examine why an item is differentially functioning, and how these discrepancies relate to the operationalization of the latent trait. Importantly, evidence of DIF should not be a catalyst for removing what is deemed "construct-irrelevant variance" (Bennett, 2022; Elliot, 2016; Randall, 2021). Wherever possible, test items should represent the full spectrum of

sociocultural perspectives relevant to the trait (Mislevy, 2018), rather than chiseling the trait and representative items to embody only the most privileged cultural perspectives (Randall, 2021).

It bears repeating that test bias is one facet of the larger conceptualization of test fairness. As it was defined here, Unbiased-ATA cannot directly speak to the usage and consequences of test scores, nor the accessibility of the test to examinees from different backgrounds. Additionally, because Unbiased-ATA is applied after items are written, this method cannot on its own change how the test developers conceptualized the trait of interest (a process that can implicitly reflect personal biases; Elder, 1997; Randall, 2021). Across these studies, ATA-selected tests demonstrated relatively weaker psychometric properties when there were both high regularized DIF FPRs and larger differences in group-level item parameter estimates. ATA thus cannot fully compensate for evidence of measurement non-invariance in the item bank. Even when incorporating MI criteria in the objective function, ATA might still construct tests with evidence of DIF or DTF if there are too many differentially functioning items in the bank.

Methods like Unbiased-ATA are therefore advantageous in that they can efficiently highlight items in need of review and revision. Yet the process of constructing unbiased tests begins not with item analysis, but many steps beforehand with trait conceptualization and item writing. Indeed, Unbiased-ATA (and similar test construction methods) should be considered one component of a broader validity study that prioritizes equity and diversity at each stage of the test construction process (e.g., Randall et al., 2022; Slomp, 2016). It is also imperative that alternative conceptualizations of test fairness are considered. As Poe and Elliot (2019) write, “Because culturally diverse

philosophical views of assessment fairness express concerns beyond deficit views—that is, beyond the assumption that once construct-irrelevant variance is removed that all will be well—it will be increasingly important to understand how non-Western cultures construct, interpret, resist, and transform evidence of fairness" (p. 15).

Conclusion

Test fairness is an integral component of latent trait test construction, equally important as the traditional teachings of test score validity and precision. Unbiased-ATA is one step toward explicitly incorporating indices of test bias into the test assembly process. Still, considerably more work is necessary. Acknowledging psychometrics' troubled history, often connected to racist movements like eugenics (Dixon-Román, 2020; Michell, 2021), the field must continue to reframe and improve test construction procedures to ensure equitable and inclusive testing practices.

References

Adema, J. J., Boekkooi-Timminga, E., & van der Linden, W. J. (1991). Achievement test construction using 0–1 linear programming. *European Journal of Operational Research*, *55*(1), 103–111. https://doi.org/10.1016/0377-2217(91)90195-2

Ali, U. S., & van Rijn, P. W. (2016). An evaluation of different statistical targets for assembling parallel forms in item response theory. *Applied Psychological Measurement*, *40*(3), 163–179. https://doi.org/10.1177/0146621615613308

American Psychological Association, American Educational Research Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, & Psychological Testing. (2014). *Standards for educational & psychological tests*. AERA.

Ames, A. J., & Penfield, R. D. (2015). An NCME instructional module on item-fit statistics for item response theory models. *Educational Measurement: Issues and Practice*, *34*(3), 39–48. https://doi.org/10.1111/emip.12067

Armstrong, R. D., Jones, D. H., & Kunce, C. S. (1998). IRT test assembly using network-flow programming. *Applied Psychological Measurement*, *22*(3), 237–247. https://doi.org/10.1177/01466216980223004

Attell, B. K., Cappelli, C., Manteuffel, B., & Li, H. (2020). Measuring functional impairment in children and adolescents: Psychometric properties of the Columbia Impairment Scale (CIS). *Evaluation & the Health Professions*, *43*(1), 3–15. https://doi.org/10.1177/0163278718775797

Baharloo, A. (2013). Test fairness in traditional and dynamic assessment. *Theory and Practice in Language Studies*, *3*(10), 1930–1938. https://doi.org/10.4304/tpls.3.10.1930-1938

Barron, F. H., & Barrett, B. E. (1996). Decision quality using ranked attribute weights. *Management Science*, *42*(11), 1515–1523. https://doi.org/10.1287/mnsc.42.11.1515

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bauer, D. J., Belzak, W. C., & Cole, V. T. (2020). Simplifying the assessment of measurement invariance over multiple background variables: Using regularized moderated nonlinear factor analysis to detect differential item functioning. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(1), 43–55. https://doi.org/10.1080/10705511.2019.1642754

Belzak, W. C. M. (2020). Testing differential item functioning in small samples. *Multivariate Behavioral Research*, *55*(5), 722–747. https://doi.org/10.1080/00273171.2019.1671162

Belzak, W. C. M. (2021). *RegDIF: Regularized differential item functioning* (0.1.0) [Computer software].

Belzak, W. C. M., & Bauer, D. J. (2020). Improving the assessment of measurement invariance: Using regularization to select anchor items and identify differential item functioning. *Psychological Methods*, *25*(6), 673–690. https://doi.org/10.1037/met0000253

Bennett, R. E. (2022). The good side of COVID-19. *Educational Measurement: Issues and Practice*, emip.12496. https://doi.org/10.1111/emip.12496

Ben-Shachar, M., Lüdecke, D., & Makowski, D. (2020). effectsize: Estimation of effect size indices and standardized parameters. *Journal of Open Source Software*, *5*(56), 2815. https://doi.org/10.21105/joss.02815

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238–246. https://doi.org/10.1037/0033-2909.107.2.238

Birnbaum, A. (1986). Some latent trait models and their use in inferring an examinee's ability. In Lord F. M. & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Addison-Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459. https://doi.org/10.1007/BF02293801

Boekkooi-Timminga, E. (1990). The construction of parallel tests from IRT-based item banks. *Journal of Educational Statistics*, *15*(2), 129–145. https://doi.org/10.3102/10769986015002129

Bolt, D. M. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education*, *15*, 113–141. https://doi.org/10.1207/S15324818AME1502_01

Borsboom, D., Cramer, A. O., Kievit, R. A., Scholten, A. Z., & Franić, S. (2009). The end of construct validity. In *The concept of validity: Revisions, new directions and applications*. IAP Information Age Publishing.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

Borsboom, D., Romeijn, J.-W., & Wicherts, J. M. (2008). Measurement invariance versus selection invariance: Is fair selection possible? *Psychological Methods*, *13*(2), 75–98. https://doi.org/10.1037/1082-989X.13.2.75

Boyd, S. P., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press.

Bradlow, E. T. (1996). Teacher's Corner: Negative information and the three-parameter logistic model. *Journal of Educational and Behavioral Statistics*, *21*(2), 179–185. https://doi.org/10.3102/10769986021002179

Browne, M., Rockloff, M., & Rawat, V. (2018). An SEM algorithm for scale reduction incorporating evaluation of multiple psychometric criteria. *Sociological Methods & Research*, *47*(4), 812–836. https://doi.org/10.1177/0049124116661580

Camilli, G. (2013). Ongoing issues in test fairness. *Educational Research and Evaluation*, *19*(2–3), 104–120. https://doi.org/10.1080/13803611.2013.767602

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105. https://doi.org/10.1037/h0046016

Cerny, V. (1985). Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. *Journal of Optimization Theory and Applications*, *45*(1), 41–51. https://doi.org/10.1007/BF00940812

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6). https://doi.org/10.18637/jss.v048.i06

Chalmers, R. P., Counsell, A., & Flora, D. B. (2016). It might not make a big DIF: Improved differential test functioning statistics that account for sampling variability. *Educational and Psychological Measurement*, *76*(1), 114–140. https://doi.org/10.1177/0013164415584576

Chang, Y.-H., & Yeh, C.-H. (2001). Evaluating airline competitiveness using multiattribute decision making. *Omega*, *29*(5), 405–415. https://doi.org/10.1016/S0305-0483(01)00032-9

Chang, Y.-W., Huang, W.-K., & Tsai, R.-C. (2015). DIF detection using multiple-group categorical CFA with minimum free baseline approach. *Journal of Educational Measurement*, *52*(2), 181–199. https://doi.org/10.1111/jedm.12073

Chen, P.-H. (2016). Three-element item selection procedures for multiple forms assembly: An item matching approach. *Applied Psychological Measurement*, *40*(2), 114–127. https://doi.org/10.1177/0146621615605307

Chen, P.-H. (2017). Should we stop developing heuristics and only rely on mixed integer programming solvers in automated test assembly? A rejoinder to van der Linden and Li (2016). *Applied Psychological Measurement*, *41*(3), 227–240. https://doi.org/10.1177/0146621617695523

Chen, P.-H., Chang, H.-H., & Wu, H. (2012). Item selection for the development of parallel forms from an IRT-based seed test using a sampling and classification approach. *Educational and Psychological Measurement*, *72*(6), 933–953. https://doi.org/10.1177/0013164412443688

Chen, S. M., Bauer, D. J., Belzak, W. M., & Brandt, H. (2021). Advantages of spike and slab priors for detecting differential item functioning relative to other Bayesian regularizing priors and frequentist lasso. *Structural Equation Modeling: A Multidisciplinary Journal*, 1–18. https://doi.org/10.1080/10705511.2021.1948335

Chen, Y., Li, C., & Xu, G. (2021). DIF statistical inference and detection without knowing anchoring items. *ArXiv:2110.11112 [Stat]*.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Cohen, J. (1973). Eta-squared and partial eta-squared in fixed factor ANOVA designs. *Educational and Psychological Measurement*, *33*, 107–112.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Cole, N. S., & Zieky, M. J. (2001). The new faces of fairness. *Journal of Educational Measurement*, *38*(4), 369–382. https://doi.org/10.1111/j.1745-3984.2001.tb01132.x

Colorni, A., Dorigo, M., & Maniezzo, V. (1991). Distributed optimization by ant colonies. In F. Varela & P. Bourgine (Eds.), *Proceedings of ECAL 91: First European Conference on Artificial Life*. Elsevier.

Cor, K., Alves, C., Gierl, M., & others. (2009). Three applications of automated test assembly within a user-friendly modeling environment. *Practical Assessment, Research, and Evaluation*, *14*(14), 1–23. https://doi.org/10.7275/02m6-1268

Cortina, J. M., Sheng, Z., Keener, S. K., Keeler, K. R., Grubb, L. K., Schmitt, N., Tonidandel, S., Summerville, K. M., Heggestad, E. D., & Banks, G. C. (2020). From alpha to omega and beyond! A look at the past, present, and (possible)

future of psychometric soundness in the Journal of Applied Psychology. *Journal of Applied Psychology*, *105*(12), 1351–1381. https://doi.org/10.1037/apl0000815

Costa, P. T., & McCrae, R. R. (2008). The Revised NEO Personality Inventory (NEO-PI-R). In *The SAGE handbook of personality theory and assessment: Volume 2—Personality neasurement and testing* (pp. 179–198). SAGE Publications Ltd. https://doi.org/10.4135/9781849200479.n9

Costa, P. T., McCrae, R. R., & Dye, D. A. (1991). Facet scales for Agreeableness and Conscientiousness: A revision of the NEO Personality Inventory. *Personality and Individual Differences*, *12*(9), 887–898. https://doi.org/10.1016/0191-8869(91)90177-D

Crocker, L., & Algina, J. (2008). *Introduction to classical & modern test theory*. Cengage Learning.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

Davies, A. (2010). Test fairness: A response. *Language Testing*, *27*(2), 171–176. https://doi.org/10.1177/0265532209349466

Davis, J. P. (2009). *A comparative study of item-level fit indices in item response theory* [Doctoral dissertation]. University of Minnesota.

Dawande, M., Kalagnanam, J., Keskinocak, P., Salman, F. S., & Ravi, R. (2000). Approximation algorithms for the multiple knapsack problem with assignment restrictions. *Journal of Combinatorial Optimization*, *4*(2), 171–186. https://doi.org/10.1023/A:1009894503716

De Champlain, A. F. (2010). A primer on classical test theory and item response theory for assessments in medical education: Classical test theory and item response theory. *Medical Education*, *44*(1), 109–117. https://doi.org/10.1111/j.1365-2923.2009.03425.x

DeGroot, M. H., & Schervish, M. J. (2012). *Probability and statistics* (4th ed). Addison-Wesley.

DeMars, C. E. (2020). Alignment as an alternative to anchor purification in DIF analyses. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(1), 56–72. https://doi.org/10.1080/10705511.2019.1617151

Diao, Q., & van der Linden, W. J. (2011). Automated test assembly using lp_Solve version 5.5 in R. *Applied Psychological Measurement*, *35*(5), 398–409. https://doi.org/10.1177/0146621610392211

Dixon-Román, E. (2020). A haunting logic of psychometrics: Toward the speculative and indeterminacy of Blackness in measurement. *Educational Measurement: Issues and Practice*, *39*(3), 94–96. https://doi.org/10.1111/emip.12375

Dodd, B. G., Koch, W. R., & De Ayala, R. J. (1993). Computerized adaptive testing using the partial credit model: Effects of item pool characteristics and different stopping rules. *Educational and Psychological Measurement*, *53*(1), 61–77. https://doi.org/10.1177/0013164493053001005

Dorans, N. J. (2011). Holland's advice for the fourth generation of test theory: Blood tests can be contests. In N. J. Dorans & S. Sinharay (Eds.), *Looking back* (Vol.

202, pp. 259–272). Springer New York. https://doi.org/10.1007/978-1-4419-9389-2_14
Dorans, N. J. (2013). ETS contributions to the quantitative assessment of item, test, and score fairness. In *ETS Research Report Series* (No. 2; pp. 1–38). Wiley Online Library.
Dorans, N. J. (2017). Contributions to the quantitative assessment of item test, and score fairness. In R. E. Bennett & M. von Davier (Eds.), *Advancing human assessment: The methodological, psychological and policy contributions of ETS* (pp. 201–230). Springer.
Dorigo, M., & Stützle, T. (2004). *Ant colony optimization*. MIT Press.
Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two-parameter logistic model. *Applied Psychological Measurement*, *13*(1), 77–90. https://doi.org/10.1177/014662168901300108
Drezner, Z., Marcoulides, G. A., & Salhi, S. (1999). Tabu search model selection in multiple regression analysis. *Communications in Statistics-Simulation and Computation*, *28*(2), 349–367. https://doi.org/10.1080/03610919908813553
Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, *105*(3), 399–412. https://doi.org/10.1111/bjop.12046
Edelen, M. O., Stucky, B. D., & Chandra, A. (2015). Quantifying 'problematic' DIF within an IRT framework: Application to a cancer stigma index. *Quality of Life Research*, *24*(1), 95–103. https://doi.org/10.1007/s11136-013-0540-4
Einhorn, H. J., & McCoach, W. (1977). A simple multiattribute utility procedure for evaluation. *Behavioral Science*, *22*(4), 270–282. https://doi.org/10.1002/bs.3830220405
Elder, C. (1997). What does test bias have to do with fairness? *Language Testing*, *14*(3), 261–277. https://doi.org/10.1177/026553229701400304
Elliot, N. (2016). A theory of ethics for writing assessment. *The Journal of Writing Assessment*, *9*(1).
Embretson, S. E., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement*, *38*(4), 343–368. https://doi.org/10.1111/j.1745-3984.2001.tb01131.x
Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.
Feuerstahler, L. M. (2018). Sources of error in IRT trait estimation. *Applied Psychological Measurement*, *42*(5), 359–375. https://doi.org/10.1177/0146621617733955
Finch, W. H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement*, *29*(4), 278–295. https://doi.org/10.1177/0146621605275728
Finch, W. H. (2016). Detection of differential item functioning for more than two groups: A Monte Carlo comparison of methods. *Applied Measurement in Education*, *29*(1), 30–45. https://doi.org/10.1080/08957347.2015.1102916
Finkelman, M. D., Kim, W., Roussos, L., & Verschoor, A. (2010). A binary programming approach to automated test assembly for cognitive diagnosis

models. *Applied Psychological Measurement*, *34*(5), 310–326. https://doi.org/10.1177/0146621609344846
Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, *8*(4), 370–378. https://doi.org/10.1177/1948550617693063
Fraser, A. S. (1957). Simulation of genetic systems by automatic digital computers. *Australian Journal of Biological Sciences*, *10*(4), 484–491. https://doi.org/10.1071/BI9570484
Fu, A., Narasimhan, B., & Boyd, S. (2020). CVXR: An R package for disciplined convex optimization. *Journal of Statistical Software*, *94*(14), 1–34. https://doi.org/10.18637/jss.v094.i14
Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2020). Datasheets for datasets. *ArXiv:1803.09010 [Cs]*.
Gierl, M. J., Lai, H., & Li, J. (2013). Identifying differential item functioning in multi-stage computer adaptive testing. *Educational Research and Evaluation*, *19*(2–3), 188–203. https://doi.org/10.1080/13803611.2013.767622
Gleason, K. A., Jensen-Campbell, L. A., & South Richardson, D. (2004). Agreeableness as a predictor of aggression in adolescence. *Aggressive Behavior*, *30*(1), 43–61. https://doi.org/10.1002/ab.20002
Glöckner, A., Michels, M., & Giersch, D. (2020). *Predicting personality test scores with machine learning methodology: Investigation of a new approach to psychological assessment* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/ysd3f
Glover, F. (1986). Future paths for integer programming and links to artificial intelligence. *Computers & Operations Research*, *13*(5), 533–549. https://doi.org/10.1016/0305-0548(86)90048-1
Goetz, C., Coste, J., Lemetayer, F., Rat, A.-C., Montel, S., Recchia, S., Debouverie, M., Pouchot, J., Spitz, E., & Guillemin, F. (2013). Item reduction based on rigorous methodological guidelines is necessary to maintain validity when shortening composite measurement scales. *Journal of Clinical Epidemiology*, *66*(7), 710–718. https://doi.org/10.1016/j.jclinepi.2012.12.015
Goldberg, D. E. (1989). *Genetic algorithms in search, optimization and machine learning*. Addison-Wesley.
Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist*, *48*(1), 26–34. https://doi.org/10.1037/0003-066X.48.1.26
Goldberg, L. R. (1999). A broad-bandwidth, public domain personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe* (Vol. 7). Tillburg University Press.
Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality*, *40*(1), 84–96. https://doi.org/10.1016/j.jrp.2005.08.007
Gonzalez, O. (2020). Psychometric and machine learning approaches to reduce the length of scales. *Multivariate Behavioral Research*, 1–17. https://doi.org/10.1080/00273171.2020.1781585

Gorin, J. S. (2007). Reconsidering issues in validity theory. *Educational Researcher*, *36*(8), 456–462. https://doi.org/10.3102/0013189X07311607

Green, D. E., Walkey, F. H., McCormick, I. A., & Taylor, A. J. W. (1988). Development and evaluation of a 21-item version of the Hopkins symptom checklist with New Zealand and United States respondents. *Australian Journal of Psychology*, *40*(1), 61–70. https://doi.org/10.1080/00049538808259070

Haladyna, T. M., & Downing, S. M. (2005). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, *23*(1), 17–27. https://doi.org/10.1111/j.1745-3992.2004.tb00149.x

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 23–34. https://doi.org/10.20982/tqmp.08.1.p023

Hambleton, R. K., & Jones, R. W. (2005). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12*(3), 38–47. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x

Haney, C., & Hurtado, A. (1994). The jurisprudence of race and meritocracy. *Law and Human Behavior*, *18*(3), 223–248.

Hansen, M., Cai, L., Stucky, B. D., Tucker, J. S., Shadel, W. G., & Edelen, M. O. (2014). Methodology for developing and evaluating the PROMIS(R) Smoking item banks. *Nicotine & Tobacco Research*, *16*(Suppl 3), S175–S189. https://doi.org/10.1093/ntr/ntt123

Harel, D., & Baron, M. (2019). Methods for shortening patient-reported outcome measures. *Statistical Methods in Medical Research*, *28*(10–11), 2992–3011. https://doi.org/10.1177/0962280218795187

Hattori, M., Zhang, G., & Preacher, K. J. (2017). Multiple local solutions and geomin rotation. *Multivariate Behavioral Research*, *52*(6), 720–731. https://doi.org/10.1080/00273171.2017.1361312

Hedges, L. V. (1983). A random effects model for effect sizes. *Psychological Bulletin*, *93*(2), 388–395. https://doi.org/10.1037/0033-2909.93.2.388

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, *3*(4), 486–504. https://doi.org/10.1037/1082-989X.3.4.486

Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedures. In *Test validity* (pp. 129–145). Lawrence Erlbaum.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Lawrence Erlbaum Associates, Inc.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Huang, C. D., Church, A. T., & Katigbak, M. S. (1997). Identifying cultural differences in items and traits: Differential item functioning in the NEO Personality Inventory. *Journal of Cross-Cultural Psychology*, *28*(2), 192–218. https://doi.org/10.1177/0022022197282004

Huang, F. Y., Chung, H., Kroenke, K., Delucchi, K. L., & Spitzer, R. L. (2006). Using the Patient Health Questionnaire-9 to measure depression among racially and ethnically diverse primary care patients. *Journal of General Internal Medicine*, *21*(6), 547–552. https://doi.org/10.1111/j.1525-1497.2006.00409.x

Huitzing, H. A., Veldkamp, B. P., & Verschoor, A. J. (2005). Infeasibility in automated test assembly models: A comparison study of different methods. *Journal of Educational Measurement*, *42*(3), 223–243. https://doi.org/10.1111/j.1745-3984.2005.00012.x

Hutchinson, B., & Mitchell, M. (2019). 50 years of test (un) fairness: Lessons for machine learning. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 49–58.

Ishihara, M., Harel, D., Levis, B., Levis, A. W., Riehm, K. E., Saadat, N., Azar, M., Rice, D. B., Sanchez, T. A., Chiovitti, M. J., Cuijpers, P., Gilbody, S., Ioannidis, J. P. A., Kloda, L. A., McMillan, D., Patten, S. B., Shrier, I., Arroll, B., Bombardier, C. H., … Thombs, B. D. (2019). Shortening self-report mental health symptom measures through optimal test assembly methods: Development and validation of the Patient Health Questionnaire-Depression-4. *Depression and Anxiety*, *36*(1), 82–92. https://doi.org/10.1002/da.22841

Jankowsky, K., Olaru, G., & Schroeders, U. (2020). Compiling measurement invariant short scales in cross–cultural personality assessment using ant colony optimization. *European Journal of Personality*, *34*(3), 470–485. https://doi.org/10.1002/per.2260

Jia, J., Fischer, G. W., & Dyer, J. S. (1998). Attribute weighting methods and decision quality in the presence of response error: A simulation study. *Journal of Behavioral Decision Making*, *11*(2), 85–105. https://doi.org/10.1002/(SICI)1099-0771(199806)11:2<85::AID-BDM282>3.0.CO;2-K

Jones, A. T., Kopp, J. P., & Ong, T. Q. (2020). The invariance paradox: Using optimal test design to minimize bias. *Educational Measurement: Issues and Practice*, *39*(2), 48–57. https://doi.org/10.1111/emip.12298

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, *70*(351a), 631–639. https://doi.org/10.1080/01621459.1975.10482485

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling: A Multidisciplinary Journal*, *15*(1), 136–153. https://doi.org/10.1080/10705510701758406

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*(4), 319–342. https://doi.org/10.1111/j.1745-3984.2001.tb01130.x

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). American Council on Education and Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kang, T., & Chen, T. T. (2008). Performance of the generalized $S\text{-}X^2$ item fit index for polytomous IRT models. *Journal of Educational Measurement*, *45*(4), 391–406. https://doi.org/10.1111/j.1745-3984.2008.00071.x

Kang, T., & Chen, T. T. (2011). Performance of the generalized *S-X*^{2} item fit index for the graded response model. *Asia Pacific Education Review*, *12*(1), 89–96. https://doi.org/10.1007/s12564-010-9082-4

Karami, H. (2013). The quest for fairness in language testing. *Educational Research and Evaluation*, *19*(2–3), 158–169. https://doi.org/10.1080/13803611.2013.767618

Kellerer, H., Pferschy, U., & Pisinger, D. (2004). *Knapsack problems*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-24777-7

Kerber, A., Schultze, M., Müller, S., Rühling, R. M., Wright, A. G. C., Spitzer, C., Krueger, R. F., Knaevelsrud, C., & Zimmermann, J. (2020). Development of a short and ICD-11 compatible measure for DSM-5 maladaptive personality traits using ant colony optimization algorithms. *Assessment*. https://doi.org/10.1177/1073191120971848

Kim, E. S., & Yoon, M. (2011). Testing measurement invariance: A comparison of multiple-group categorical CFA and IRT. *Structural Equation Modeling*, *18*(2), 212–228. https://doi.org/10.1080/10705511.2011.557337

Kim, S.-H., & Cohen, A. S. (1998). Detection of differential item functioning under the graded response model with the likelihood ratio test. *Applied Psychological Measurement*, *22*(4), 345–355. https://doi.org/10.1177/014662169802200403

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, *220*(4598), 671–680. https://doi.org/10.1126/science.220.4598.671

Kleinman, M., & Teresi, J. A. (2016). Differential item functioning magnitude and impact measures from item response theory models. *Psychological Test and Assessment Modeling*, *58*(1), 79–98.

Kline, R. B. (2013). Assessing statistical aspects of test fairness with structural equation modelling. *Educational Research and Evaluation*, *19*(2–3), 204–222. https://doi.org/10.1080/13803611.2013.767624

Kopf, J., Zeileis, A., & Strobl, C. (2015). Anchor selection strategies for DIF analysis: Review, assessment, and new approaches. *Educational and Psychological Measurement*, *75*(1), 22–56. https://doi.org/10.1177/0013164414529792

Kuncel, N. R., & Klieger, D. M. (2012). Predictive bias in work and educational settings. In N. Schmitt (Ed.), *The Oxford handbook of personnel assessment and selection*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199732579.013.0020

Kunnan, A. J. (2000). Test fairness. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment*. Cambridge University Press.

Kunnan, A. J. (2004). Test fairness. In M. Milanovic & C. Weir (Eds.), *European language testing in a global context: Proceedings of the ALTE Barcelona Conference*. Cambridge University Press.

Kunnan, A. J. (2007). Test fairness, test bias, and DIF. *Language Assessment Quarterly*, *4*(2), 109–112. https://doi.org/10.1080/15434300701375865

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lai, K., & Green, S. B. (2016). The problem with having two watches: Assessment of fit when RMSEA and CFI disagree. *Multivariate Behavioral Research*, *51*(2–3), 220–239. https://doi.org/10.1080/00273171.2015.1134306

Lei, P.-W., Chen, S.-Y., & Yu, L. (2006). Comparing methods of assessing differential item functioning in a computerized adaptive testing environment. *Journal of Educational Measurement*, *43*(3), 245–264. https://doi.org/10.1111/j.1745-3984.2006.00015.x

Leite, W. L., Huang, I.-C., & Marcoulides, G. A. (2008). Item selection for the development of short forms of scales using an ant colony optimization algorithm. *Multivariate Behavioral Research*, *43*(3), 411–431. https://doi.org/10.1080/00273170802285743

Levis, A. W., Harel, D., Kwakkenbos, L., Carrier, M.-E., Mouthon, L., Poiraudeau, S., Bartlett, S. J., Khanna, D., Malcarne, V. L., Sauve, M., van den Ende, C. H. M., Poole, J. L., Schouffoer, A. A., Welling, J., Thombs, B. D., & and the Scleroderma Patient-Centered Intervention Network Investigators. (2016). Using optimal test assembly methods for shortening patient-reported outcome measures: Development and validation of the Cochin Hand Function Scale-6: A Scleroderma Patient-centered Intervention Network Cohort Study. *Arthritis Care & Research*, *68*(11), 1704–1713. https://doi.org/10.1002/acr.22893

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, *3*(3), 635–694. https://doi.org/10.2466/pr0.1957.3.3.635

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

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Information Age Publishing.

Luecht, R. M. (1998a). A reaction to "Moderating possibly irrelevant multiple mean score differences on a test of mathematical reasoning." *Journal of Educational Measurement*, *35*(3), 223–225.

Luecht, R. M. (1998b). Computer-assisted test assembly using optimization heuristics. *Applied Psychological Measurement*, *22*(3), 224–236. https://doi.org/10.1177/01466216980223003

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods*, *49*(4), 1494–1502. https://doi.org/10.3758/s13428-016-0809-y

Luo, X. (2020). Automated test assembly with mixed-integer programming: The effects of modeling approaches and solvers. *Journal of Educational Measurement*, *57*(4), 547–565. https://doi.org/10.1111/jedm.12262

Magis, D. (2015). A note on the equivalence between observed and expected information functions with polytomous IRT models. *Journal of Educational and Behavioral Statistics*, *40*(1), 96–105. https://doi.org/10.3102/1076998614558122

Magis, D., Raîche, G., Béland, S., & Gérard, P. (2011). A generalized logistic regression procedure to detect differential item functioning among multiple groups. *International Journal of Testing*, *11*(4), 365–386. https://doi.org/10.1080/15305058.2011.602810

Makhorin, A. (2012). *GLPK (GNU Linear Programming Kit)*. Free Software Foundation.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*(4), 719–748. https://doi.org/10.1093/jnci/22.4.719

Marc, L. G., Raue, P. J., & Bruce, M. L. (2008). Screening performance of the 15-item Geriatric Depression Scale in a diverse elderly home care population. *The American Journal of Geriatric Psychiatry*, *16*(11), 914–921. https://doi.org/10.1097/JGP.0b013e318186bd67

Marcoulides, G. A. (1994). Selecting weighting schemes in multivariate generalizability studies. *Educational and Psychological Measurement*, *54*(1), 3–7.

Marcoulides, G. A., & Drezner, Z. (2004). Tabu search variable selection with resource constraints. *Communications in Statistics - Simulation and Computation*, *33*(2), 355–362. https://doi.org/10.1081/SAC-120037240

Marcoulides, K. M. (2018). Automated latent growth curve model fitting: A segmentation and knot selection approach. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(5), 687–699. https://doi.org/10.1080/10705511.2018.1424548

Marcoulides, K. M. (2020). Latent growth curve model selection with Tabu search. *International Journal of Behavioral Development*, 1–7. https://doi.org/10.1177/0165025420941170

Marcoulides, K. M., & Falk, C. F. (2018). Model specification searches in structural equation modeling with R. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(3), 484–491. https://doi.org/10.1080/10705511.2017.1409074

Marsh, H. W., Hau, K.-T., & Grayson, D. (2005). Goodness of fit in structural equation models. In A. Maydeu-Olivares & J. McArdle (Eds.), *Multivariate applications book series. Contemporary psychometrics: A festschrift for Roderick P. McDonald* (pp. 275–340). Lawrence Erlbaum Associates Publishers.

Martín-Fernández, M., Gracia, E., & Lila, M. (2021). A short measure of acceptability of intimate partner violence against women: Development and validation of the A-IPVAW-8 scale. *Assessment*, 107319112110001. https://doi.org/10.1177/10731911211000110

Maydeu-Olivares, A. (2015). Evaluating the fit of IRT models. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment*. Routledge.

Maydeu-Olivares, A., & Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research*, *49*(4), 305–328. https://doi.org/10.1080/00273171.2014.911075

McDonald, R. P. (2013). *Test theory: A unified treatment*. Routlege, Taylor & Francis Group.

McNamara, T., & Ryan, K. (2011). Fairness versus justice in language testing: The place of English literacy in the Australian Citizenship Test. *Language Assessment Quarterly*, *8*(2), 161–178. https://doi.org/10.1080/15434303.2011.565438

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*(2), 127–143. https://doi.org/10.1016/0883-0355(89)90002-5

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. https://doi.org/10.1007/BF02294825

Meredith, W., & Millsap, R. E. (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychometrika*, *57*(2), 289–311. https://doi.org/10.1007/BF02294510

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education/Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research*, *45*(1/3), 35–44. https://doi.org/10.1023/A:1006964925094

Michell, J. (1990). *An introduction to the logic of psychological measurement*. L. Erlbaum Associates.

Michell, J. (2021). "The art of imposing measurement upon the mind": Sir Francis Galton and the genesis of the psychometric paradigm. *Theory & Psychology*, 095935432110176. https://doi.org/10.1177/09593543211017671

Mills, J. D., Olejnik, S. F., & Marcoulides, G. A. (2005). The Tabu search procedure: An alternative to the variable selection methods. *Multivariate Behavioral Research*, *40*(3), 351–371. https://doi.org/10.1207/s15327906mbr4003_4

Millsap, R. E. (1997). Invariance in measurement and prediction: Their relationship in the single-factor case. *Psychological Methods*, *2*(3), 248. https://doi.org/10.1037/1082-989X.2.3.248

Millsap, R. E. (2007). Invariance in measurement and prediction revisited. *Psychometrika*, *72*(4), 461–473. https://doi.org/10.1007/s11336-007-9039-7

Millsap, R. E. (2010). Testing measurement invariance using item response theory in longitudinal data: An introduction. *Child Development Perspectives*, *4*(1), 5–9. https://doi.org/10.1111/j.1750-8606.2009.00109.x

Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement* (1st ed.). Routledge. https://doi.org/10.4324/9781315871691

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, *8*.

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074–2102. https://doi.org/10.1002/sim.8086

Mosier, C. I. (1951). Symposium: The need and means of cross-validation. I. Problem and designs of cross-validation. *Educational and Psychological Measurement*, *11*, 5–11.

Moss, P. A. (2013). Validity in action: Lessons from studies of data use. *Journal of Educational Measurement*, *50*(1), 91–98. https://doi.org/10.1111/jedm.12003

Newman, J. R. (1977). Differential weighting in multiattribute utility measurement: When it should not and when it does make a difference. *Organizational Behavior and Human Performance*, *20*(2), 312–325. https://doi.org/10.1016/0030-5073(77)90010-1

Newton, P. E., & Baird, J.-A. (2016). The great validity debate. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 173–177. https://doi.org/10.1080/0969594X.2016.1172871

Nguyen, H. V., & Waller, N. G. (2022). Local minima and factor rotations in exploratory factor analysis. *Psychological Methods*. https://doi.org/10.1037/met0000467

Nydick, S. (2014). *catIrt: An R package for simulating IRT-based computerized adaptive tests* (0.50-0) [R].

Oehlert, G. W. (2018). *A first course in design and analysis of experiments*. Author.

Olaru, G., & Danner, D. (2021). Developing cross-cultural short scales using ant colony optimization. *Assessment*, *28*(1), 199–210. https://doi.org/10.1177/1073191120918026

Olaru, G., & Jankowsky, K. (2021). The HEX-ACO-18: Developing an age-invariant HEXACO short scale using ant colony optimization. *Journal of Personality Assessment*, 1–12. https://doi.org/10.1080/00223891.2021.1934480

Olaru, G., Schroeders, U., Wilhelm, O., & Ostendorf, F. (2018). A confirmatory examination of age-associated personality differences: Deriving age-related measurement-invariant solutions using ant colony optimization. *Journal of Personality*, *86*(6), 1037–1049. https://doi.org/10.1111/jopy.12373

Olaru, G., Wilhelm, O., Nordin, S., Witthöft, M., & Köteles, F. (2019). Modern health worries: Deriving two measurement invariant short scales for cross-cultural research with ant colony optimization. *PLOS ONE*, *14*(2), e0211819. https://doi.org/10.1371/journal.pone.0211819

Oliveri, M. E., Nastal, J., & Slomp, D. (2020). Reflections on equity-centered design. *ETS Research Report Series*, *2020*(1), 1–11. https://doi.org/10.1002/ets2.12307

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*(1), 50–64. https://doi.org/10.1177/01466216000241003

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S-X^2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement*, *27*(4), 289–298. https://doi.org/10.1177/0146621603027004004

Oshima, T. C., & Morris, S. B. (2008). Raju's Differential Functioning of Items and Tests (DFIT). *Educational Measurement: Issues and Practice*, *27*(3), 43–50. https://doi.org/10.1111/j.1745-3992.2008.00127.x

Pattanaik, S., John, M. T., Kohli, N., Davison, M. L., Chung, S., Self, K., Naik, A., & Flynn, P. M. (2020). Item and scale properties of the Oral Health Literacy Adults Questionnaire assessed by item response theory. *Journal of Public Health Dentistry*, jphd.12434. https://doi.org/10.1111/jphd.12434

Pauwels, E., Claes, L., Dierckx, E., Debast, I., Van Alphen, S. P. J. (Bas), Rossi, G., Schotte, C., Santens, E., & Peuskens, H. (2014). Age neutrality of the Young Schema Questionnaire in patients with a substance use disorder. *International Psychogeriatrics*, *26*(8), 1317–1326. https://doi.org/10.1017/S1041610214000519

Piedmont, R. L., & Weinstein, H. P. (1993). A psychometric evaluation of the new NEO–PIR facet scales for Agreeableness and Conscientiousness. *Journal of Personality Assessment*, *60*(2), 302–318. https://doi.org/10.1207/s15327752jpa6002_8

Poe, M., & Elliot, N. (2019). Evidence of fairness: Twenty-five years of research in Assessing Writing. *Assessing Writing*, *42*, 100418. https://doi.org/10.1016/j.asw.2019.100418

Popham, W. J. (2005). Consequential validity: Right concern-wrong concept. *Educational Measurement: Issues and Practice*, *16*(2), 9–13. https://doi.org/10.1111/j.1745-3992.1997.tb00586.x

R Core Team. (2021). *R: A language and environment for statistical computing* [R]. https://www.R-project.org/

Raborn, A. W., & Leite, W. L. (2018). ShortForm: An R package to select scale short forms with the ant colony optimization algorithm. *Applied Psychological Measurement*, *42*(6), 516–517. https://doi.org/10.1177/0146621617752993

Raborn, A. W., Leite, W. L., & Marcoulides, K. M. (2020). A comparison of metaheuristic optimization algorithms for scale short-form development. *Educational and Psychological Measurement*, *80*(5), 910–931. https://doi.org/10.1177/0013164420906600

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology*, *87*(3), 517–529. https://doi.org/10.1037/0021-9010.87.3.517

Raju, N. S., Van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement*, *19*(4), 353–368. https://doi.org/10.1177/014662169501900405

Randall, J. (2021). "Color-neutral" is not a thing: Redefining construct definition and representation through a justice-oriented critical antiracist lens. *Educational Measurement: Issues and Practice*, *40*(4), 82–90. https://doi.org/10.1111/emip.12429

Randall, J., Slomp, D., Poe, M., & Oliveri, M. E. (2022). Disrupting white supremacy in assessment: Toward a justice-oriented, antiracist validity framework. *Educational Assessment*, 1–9. https://doi.org/10.1080/10627197.2022.2042682

Rindskopf, D. (2001). Reliability: Measurement. In *International encyclopedia of the social & behavioral sciences* (pp. 13023–13028). Elsevier. https://doi.org/10.1016/B0-08-043076-7/00722-1

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement*, *17*(2), 105–116. https://doi.org/10.1177/014662169301700201

Rozeboom, W. W. (1992). The glory of suboptimal factor rotation: Why local minima in analytic optimization of simple structure are more blessing than curse. *Multivariate Behavioral Research*, *27*(4), 585–599. https://doi.org/10.1207/s15327906mbr2704_5

Sahai, H., & Ageel, M. I. (2000). *The Analysis of variance: Fixed, random and mixed models*.

Sahin, A., & Anil, D. (2017). The effects of test length and sample size on item parameters in item response theory. *Educational Sciences: Theory & Practice*, *17*(1), 321–335. https://doi.org/10.12738/estp.2017.1.0270

Schauberger, G., & Mair, P. (2020). A regularization approach for the detection of differential item functioning in generalized partial credit models. *Behavior Research Methods*, *52*(1), 279–294. https://doi.org/10.3758/s13428-019-01224-2

Schroeders, U., Wilhelm, O., & Olaru, G. (2016). Meta-heuristics in short scale construction: Ant colony optimization and genetic algorithm. *PLOS ONE*, *11*(11), e0167110. https://doi.org/10.1371/journal.pone.0167110

Schultze, M., & Eid, M. (2018). Identifying measurement invariant item sets in cross-cultural settings using an automated item selection procedure. *Methodology*, *14*(4), 177–188. https://doi.org/10.1027/1614-2241/a000155

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Shepard, L. A. (2005). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, *16*(2), 5–24. https://doi.org/10.1111/j.1745-3992.1997.tb00585.x

Sheppard, R., Han, K., Colarelli, S. M., Dai, G., & King, D. W. (2006). Differential item functioning by sex and race in the Hogan Personality Inventory. *Assessment*, *13*(4), 442–453. https://doi.org/10.1177/1073191106289031

Sireci, S. G., & Rios, J. A. (2013). Decisions that make a difference in detecting differential item functioning. *Educational Research and Evaluation*, *19*(2–3), 170–187. https://doi.org/10.1080/13803611.2013.767621

Slomp, D. (2016). An integrated design and appraisal framework for ethical writing assessment. *Journal of Writing Assessment*, *9*(1). https://escholarship.org/uc/item/4bg9003k

Smith, G. T. (2005). On construct validity: Issues of method and measurement. *Psychological Assessment*, *17*(4), 396–408. https://doi.org/10.1037/1040-3590.17.4.396

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology*, *91*(6), 1292–1306. https://doi.org/10.1037/0021-9010.91.6.1292

Stillwell, W. G., Seaver, D. A., & Edwards, W. (1981). A comparison of weight approximation techniques in multiattribute utility decision making. *Organizational Behavior and Human Performance*, *28*(1), 62–77. https://doi.org/10.1016/0030-5073(81)90015-5

Stocking, M. L., Jirele, T., Lewis, C., & Swanson, L. (1998). Moderating possibly irrelevant multiple mean score differences on a test of mathematical reasoning. *Journal of Educational Measurement*, *35*(3), 199–221. https://doi.org/10.1111/j.1745-3984.1998.tb00534.x

Stocking, M. L., Lawrence, I., Feigenbaum, M., Jirele, T., Lewis, C., & Van Essen, T. (2002). An empirical investigation of impact moderation in test construction. *Journal of Educational Measurement*, *39*(3), 235–252. https://doi.org/10.1111/j.1745-3984.2002.tb01176.x

Stone, C. A., & Zhang, B. (2003). Assessing goodness of fit of item response theory models: A comparison of traditional and alternative procedures. *Journal of Educational Measurement*, *40*(4), 331–352. https://doi.org/10.1111/j.1745-3984.2003.tb01150.x

Stucky, B. D., Edelen, M. O., Tucker, J. S., Shadel, W. G., Cerully, J., Kuhfeld, M., Hansen, M., & Cai, L. (2014). Development of the PROMIS(R) negative psychosocial expectancies of smoking item banks. *Nicotine & Tobacco Research*, *16*(Suppl 3), S232–S240. https://doi.org/10.1093/ntr/ntt282

Su, S., Wang, C., & Weiss, D. J. (2021). Performance of the $S-\chi^2$ statistic for the multidimensional graded response model. *Educational and Psychological Measurement*, *81*(3), 491–522. https://doi.org/10.1177/0013164420958060

Suh, Y., & Cho, S.-J. (2014). Chi-square difference tests for detecting differential functioning in a multidimensional IRT model: A Monte Carlo study. *Applied Psychological Measurement*, *38*(5), 359–375. https://doi.org/10.1177/0146621614523116

Svicher, A., Romanazzo, S., De Cesaris, F., Benemei, S., Geppetti, P., & Cosci, F. (2019). Mental Pain Questionnaire: An item response theory analysis. *Journal of Affective Disorders*, *249*, 226–233. https://doi.org/10.1016/j.jad.2019.02.030

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361–370. https://doi.org/10.1111/j.1745-3984.1990.tb00754.x

Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*(3), 393–408. https://doi.org/10.1007/BF02294363

Talbi, E.-G. (2009). *Metaheuristics: From design to implementation*. John Wiley & Sons.

Taylor, J. A. (1953). A personality scale of manifest anxiety. *The Journal of Abnormal and Social Psychology*, *48*(2), 285–290. https://doi.org/10.1037/h0056264

Teresi, J. A., Kleinman, M., & Ocepek-Welikson, K. (2000). Modern psychometric methods for detection of differential item functioning: Application to cognitive assessment measures. *Statistics in Medicine*, *19*(11–12), 1651–1683.

Teresi, J. A., Wang, C., Kleinman, M., Jones, R. N., & Weiss, D. J. (2021). Differential item functioning analyses of the Patient-Reported Outcomes Measurement Information System (PROMIS®) measures: Methods, challenges, advances, and future directions. *Psychometrika*, *86*(3), 674–711. https://doi.org/10.1007/s11336-021-09775-0

Theussl, S., Hornik, K., Buchta, C., Schwendinger, F., & Schuchardt, H. (2019). *Rglpk: R/GNU linear programming kit interface*.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 147–172). Lawrence Erlbaum Associates.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning*. Lawrence Erlbaum Associates, Inc.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tierney, R. D. (2014). Fairness as a multifaceted quality in classroom assessment. *Studies in Educational Evaluation*, *43*, 55–69. https://doi.org/10.1016/j.stueduc.2013.12.003

Van den Broeck, J., Rossi, G., Dierckx, E., & De Clercq, B. (2012). Age-neutrality of the NEO-PI-R: Potential differential item functioning in older versus younger adults. *Journal of Psychopathology and Behavioral Assessment*, *34*(3), 361–369. https://doi.org/10.1007/s10862-012-9287-4

van der Linden, W. J. (1998). Optimal assembly of psychological and educational tests. *Applied Psychological Measurement*, *22*(3), 195–211. https://doi.org/10.1177/01466216980223001

van der Linden, W. J. (2005). *Linear models for optimal test design*. Springer New York. https://doi.org/10.1007/0-387-29054-0

van der Linden, W. J., & Adema, J. J. (1998). Simultaneous assembly of multiple test forms. *Journal of Educational Measurement*, *35*(3), 185–198. https://doi.org/10.1111/j.1745-3984.1998.tb00533.x

van der Linden, W. J., & Li, J. (2016). Comment on three-element item selection procedures for multiple forms assembly: An item matching approach. *Applied Psychological Measurement*, *40*(8), 641–649. https://doi.org/10.1177/0146621616664075

Varni, J. W., Thissen, D., Stucky, B. D., Liu, Y., Magnus, B., Quinn, H., Irwin, D. E., DeWitt, E. M., Lai, J.-S., Amtmann, D., Gross, H. E., & DeWalt, D. A. (2014). PROMIS® Parent Proxy Report Scales for children ages 5–7 years: An item response theory analysis of differential item functioning across age groups. *Quality of Life Research*, *23*(1), 349–361. https://doi.org/10.1007/s11136-013-0439-0

Veldkamp, B. P. (1999). Multiple objective test assembly problems. *Journal of Educational Measurement*, *36*(3), 253–266. https://doi.org/10.1111/j.1745-3984.1999.tb00557.x

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed). Springer.

von Winterfeldt, D., & Edwards, W. (1986). *Decision analysis and behavioral research*. Cambridge University Press.

Wang, C., Zhu, R., & Xu, G. (2022). Using lasso and adaptive lasso to identify DIF in multidimensional 2PL models. *Multivariate Behavioral Research*, 1–21. https://doi.org/10.1080/00273171.2021.1985950

Wang, W.-C., & Yeh, Y.-L. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement*, *27*(6), 479–498. https://doi.org/10.1177/0146621603259902

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, *6*(4), 473–492. https://doi.org/10.1177/014662168200600408

Weiss, D. J. (2004). Computerized adaptive testing for effective and efficient measurement in counseling and education. *Measurement and Evaluation in Counseling and Development*, *37*(2), 70–84. https://doi.org/10.1080/07481756.2004.11909751

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, *21*(4), 361–375. https://doi.org/10.1111/j.1745-3984.1984.tb01040.x

Whitely, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*(1), 179–197. https://doi.org/10.1037/0033-2909.93.1.179

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis* (2nd ed. 2016). Springer International Publishing. https://doi.org/10.1007/978-3-319-24277-4

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research.* (pp. 281–324). American Psychological Association. https://doi.org/10.1037/10222-009

Wilks, S. S. (1938). Weighting systems for linear functions of correlated variables when there is no dependent variable. *Psychometrika*, *3*(1), 23–40. https://doi.org/10.1007/BF02287917

Willingham, W. W. (1999). A systemic view of test fairness. In S. Messick (Ed.), *Assessment in higher education: Issues in access, quality, student development, and public policy* (pp. 213–242). Lawrence Erlbaum.

Willingham, W. W., & Cole, N. S. (2013). *Gender and fair assessment*. Routledge.

Woods, C. M. (2011). DIF testing for ordinal items with Poly-SIBTEST, the Mantel and GMH Tests, and IRT-LR-DIF when the latent distribution is nonnormal for both groups. *Applied Psychological Measurement*, *35*(2), 145–164. https://doi.org/10.1177/0146621610377450

Wu, T.-J., & Sepulveda, A. (1998). The weighted average information criterion for order selection in time series and regression models. *Statistics & Probability Letters*, *39*(1), 1–10. https://doi.org/10.1016/S0167-7152(98)00003-0

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing*, *27*(2), 147–170. https://doi.org/10.1177/0265532209349465

Yang, X.-S. (2018). *Optimization techniques and applications with examples*. John Wiley & Sons.

Yarkoni, T. (2010). The abbreviation of personality, or how to measure 200 personality scales with 200 items. *Journal of Research in Personality*, *44*(2), 180–198. https://doi.org/10.1016/j.jrp.2010.01.002

Yuan, K.-H., Liu, H., & Han, Y. (2021). Differential item functioning analysis without a priori information on anchor items: QQ plots and graphical test. *Psychometrika*, *86*(2), 345–377. https://doi.org/10.1007/s11336-021-09746-5

Zwick, R. (2009). The investigation of differential item functioning in adaptive tests. In W. J. van der Linden & C. Glas (Eds.), *Elements of adaptive testing* (pp. 331–352). Springer.

Zwick, R. (2019). Fairness in measurement and selection: Statistical, philosophical, and public perspectives. *Educational Measurement: Issues and Practice*, *38*(4), 34–41. https://doi.org/10.1111/emip.12299

Appendices

Appendix A. Positionality Statement

Researchers' background and experiences can (knowingly or unknowingly) influence the ways in which research questions are posed, data are analyzed, and conclusions are drawn. Although positionality statements are becoming commonplace in qualitative research spaces, such self-reflexivity is equally important for quantitative research. I therefore offer additional information about my positionality to contextualize my approach to this research.

I am a White, cisgender woman born in the United States. I completed my doctoral degree on Miní Sóta Makhóčhe, the homelands of the Dakhóta Oyáte. Although my gender is historically underrepresented in quantitative spaces, most other aspects of my identity have afforded me significant opportunity throughout my life thus far. It is imperative that I acknowledge the ways in which my privilege facilitates my progress in academic spaces.

I believe that fairness is an integral component of test construction and administration. Moreover, I believe that established psychometric methods should be critically examined to ensure that psychosocial tests appropriately represent and support all experiences. I have focused my recent research on equity in test construction, but I cannot speak for, or over, scholars from historically marginalized backgrounds who have been moving this field forward for years. I am continually unlearning implicit biases as I strive to create inclusive spaces for *all* individuals.

Appendix B. Supplementary Figures

Figure A1. Average False Positive Rates for Regularized DIF in Conditions with No Simulated DIF

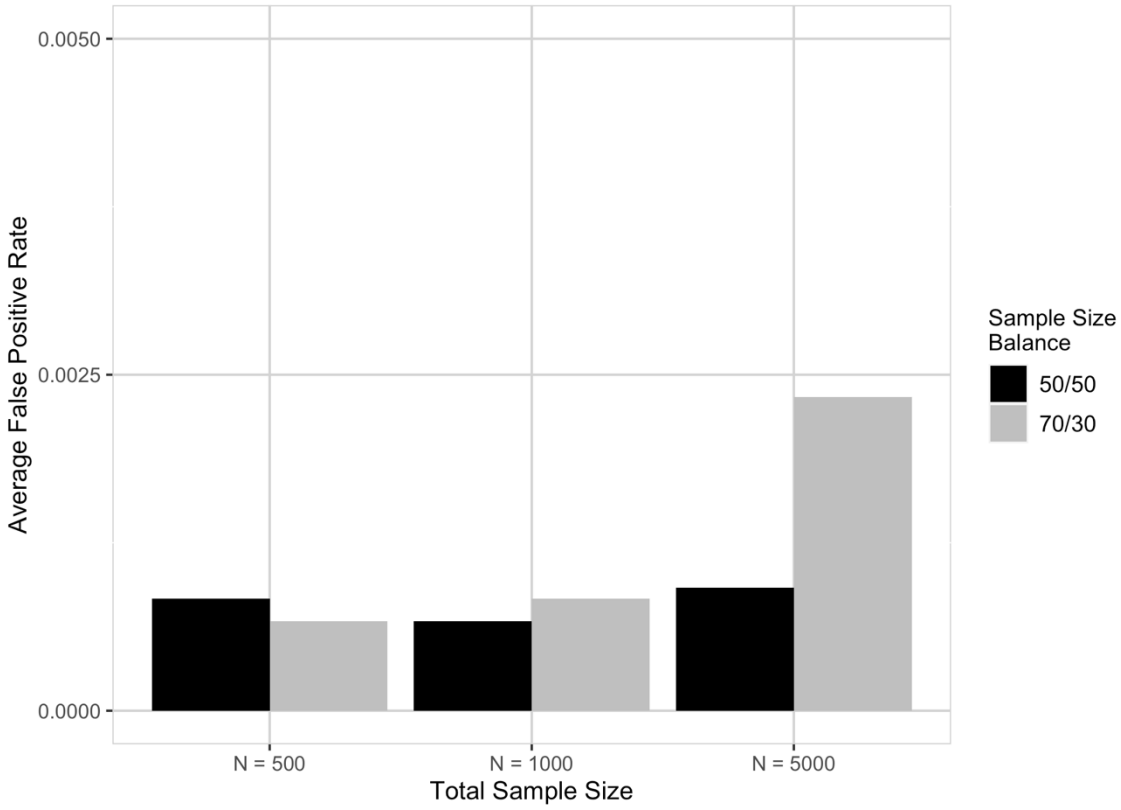


Figure A2. Average Full-Sample RMSEA for Weak MI Models Across DIF Characteristics and Sample Sizes

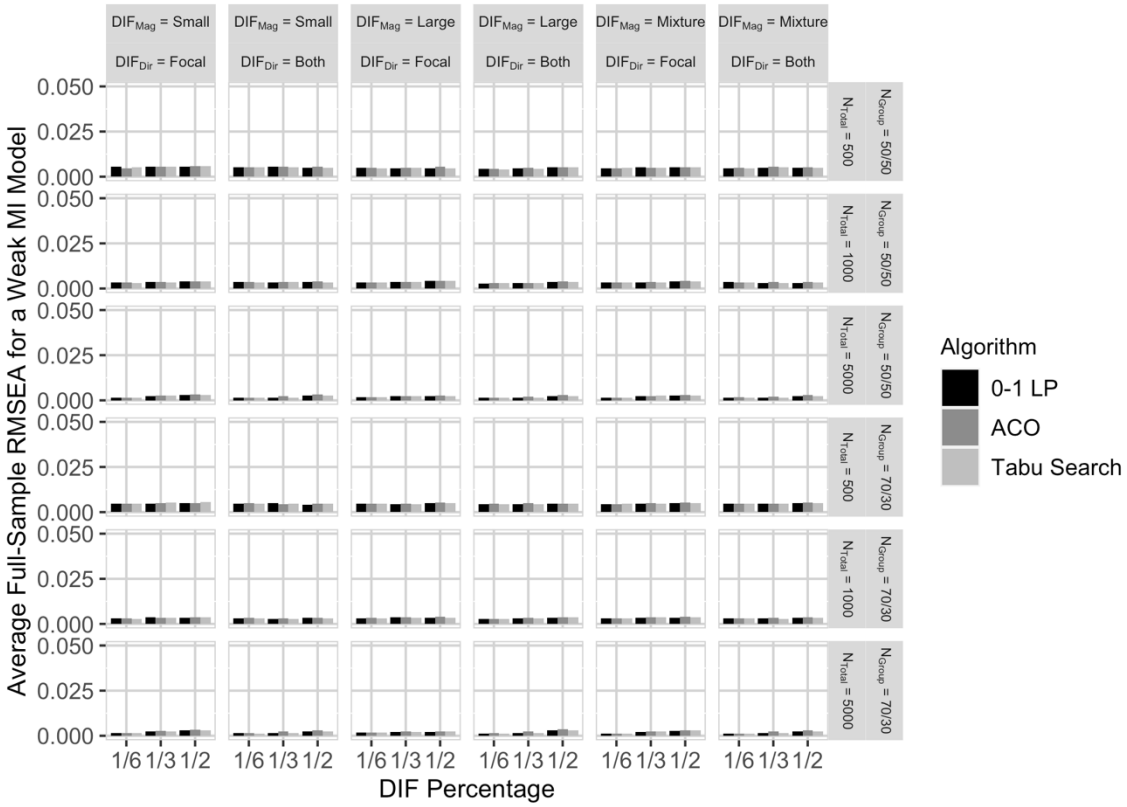


Figure A3. Average Full-Sample RMSEA for Configural MI Models Across DIF Characteristics and Sample Sizes

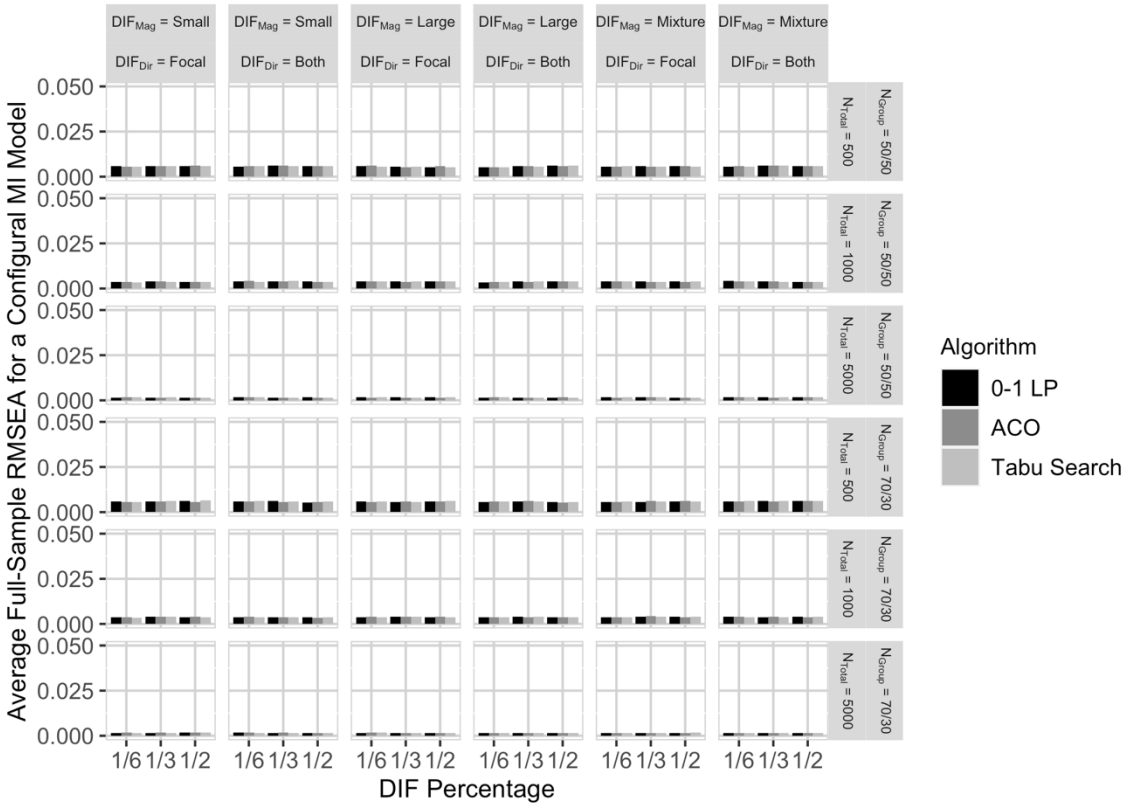


Figure A4. Average Group-Level SRMSR for Weak MI Models Across DIF Characteristics and Sample Sizes

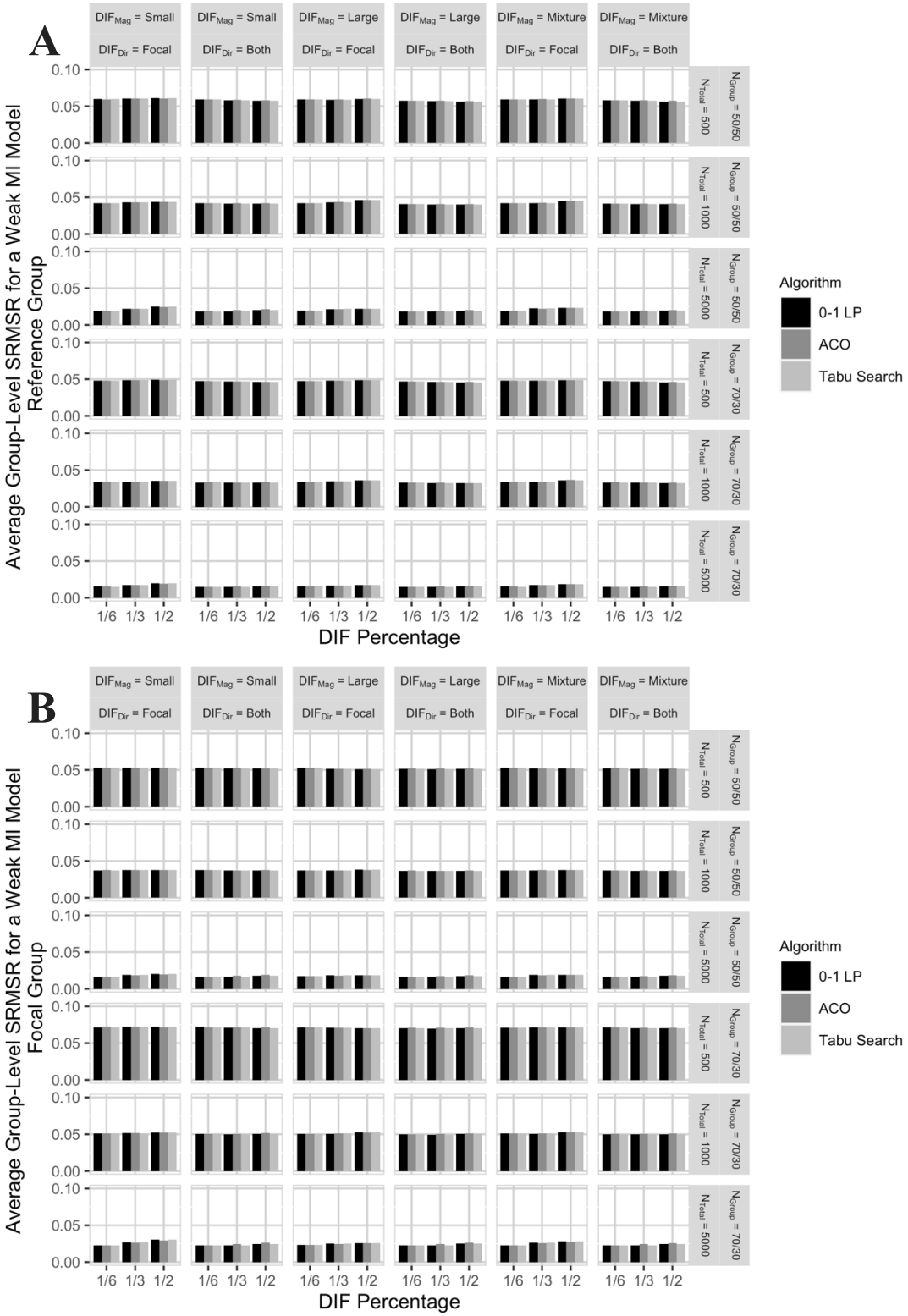


Figure A5. Average Group-Level SRMSR for Configural MI Models Across DIF Characteristics and Sample Sizes

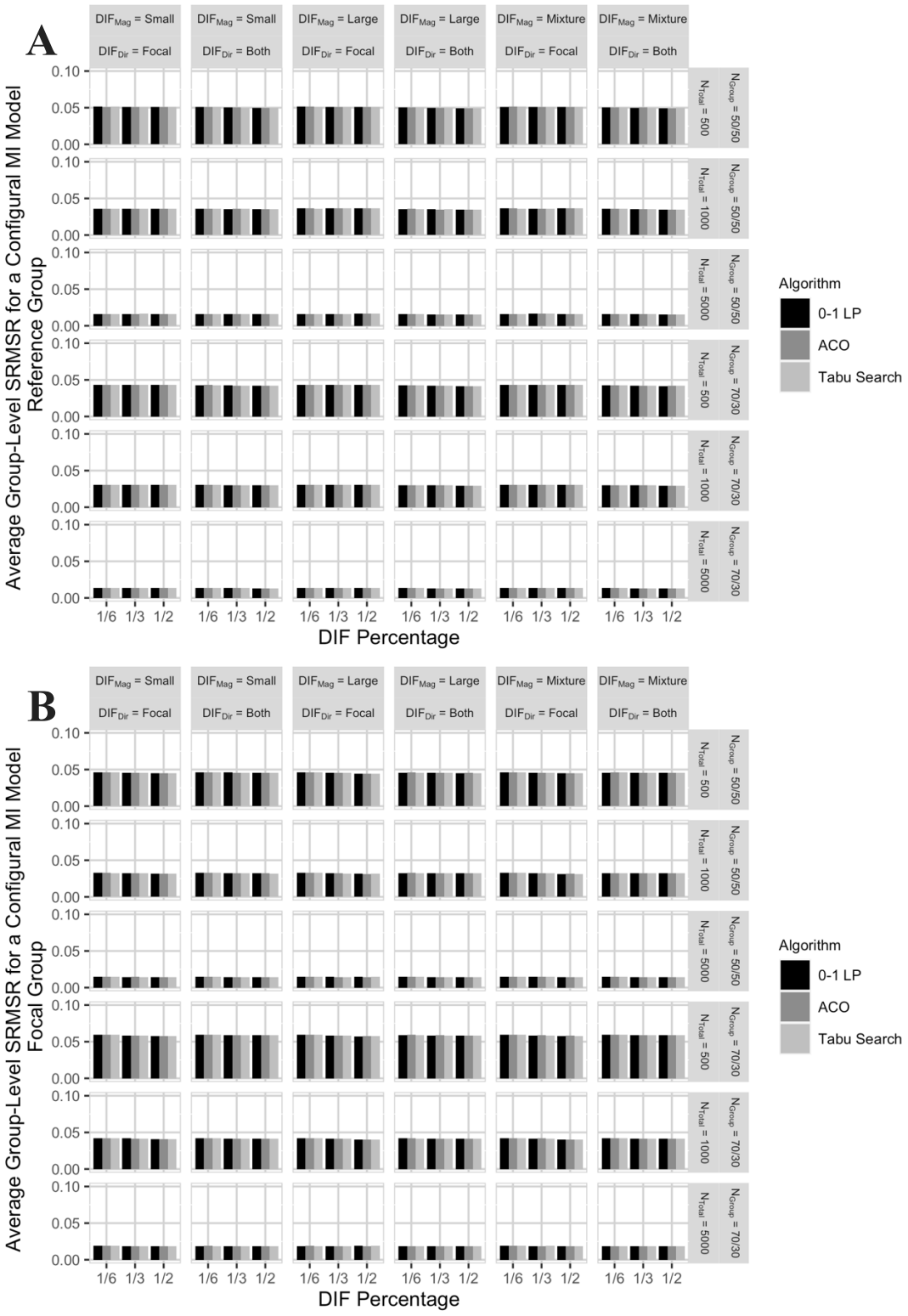


Figure A6. Average Test Information Values for Selected Tests by Sample Size, Estimation Type, and DIF Characteristics

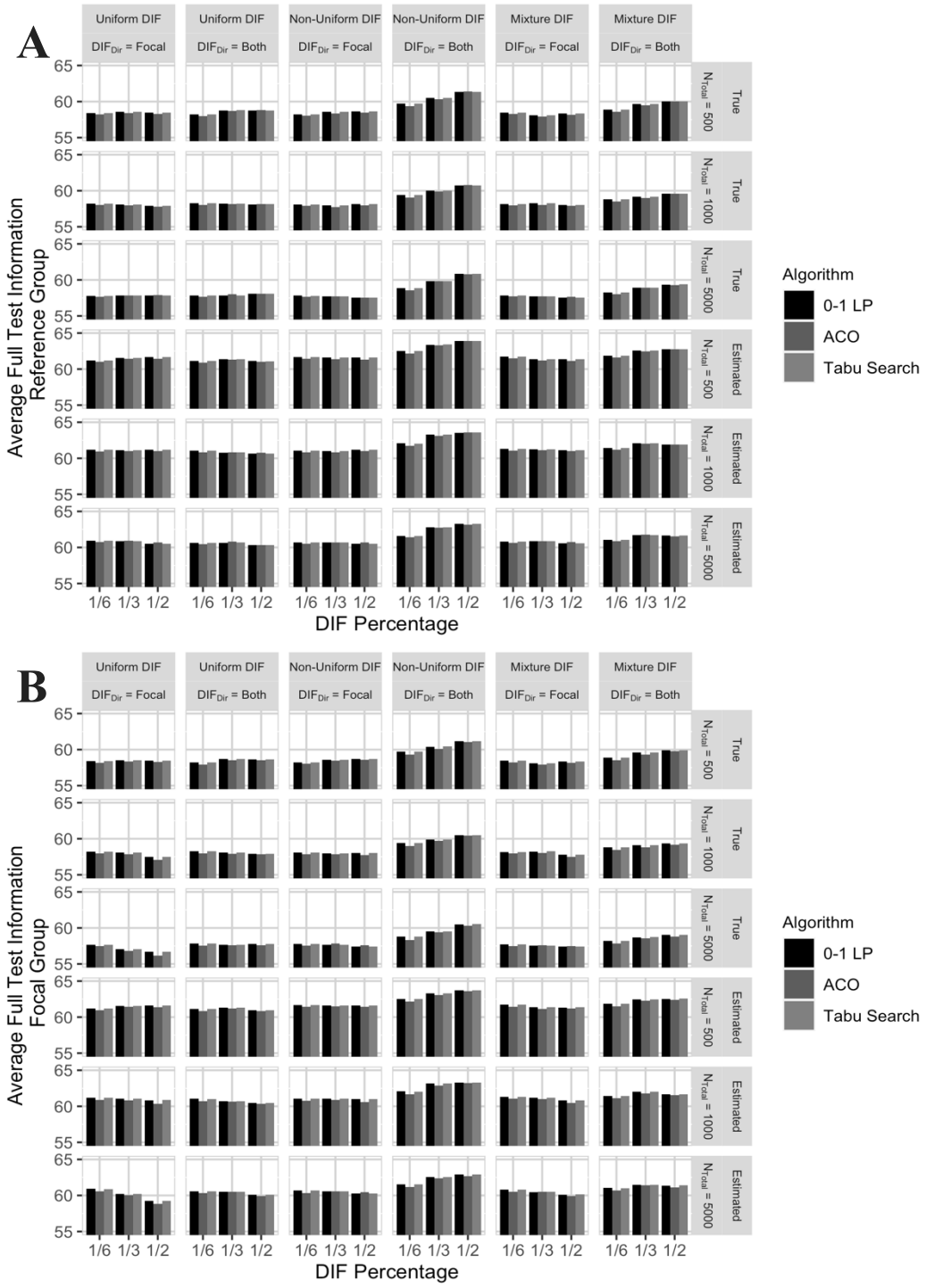


Figure A7. Average Estimated Discrimination and Difficulty Parameters for Selected Items when using IRT-LRT

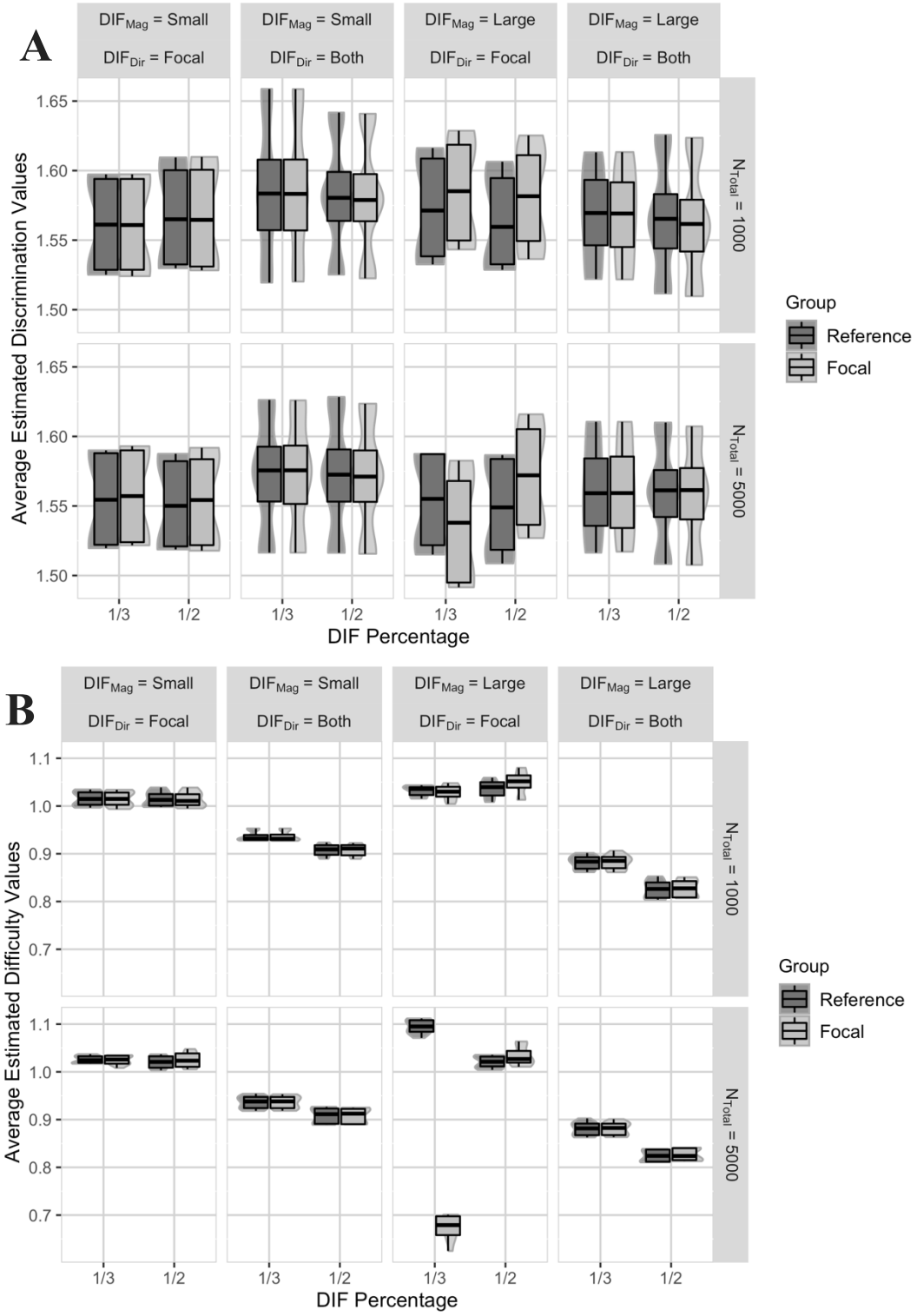


Figure A8. Average Number of Invariant and Differentially Functioning Items in Selected Tests when using IRT-LRT

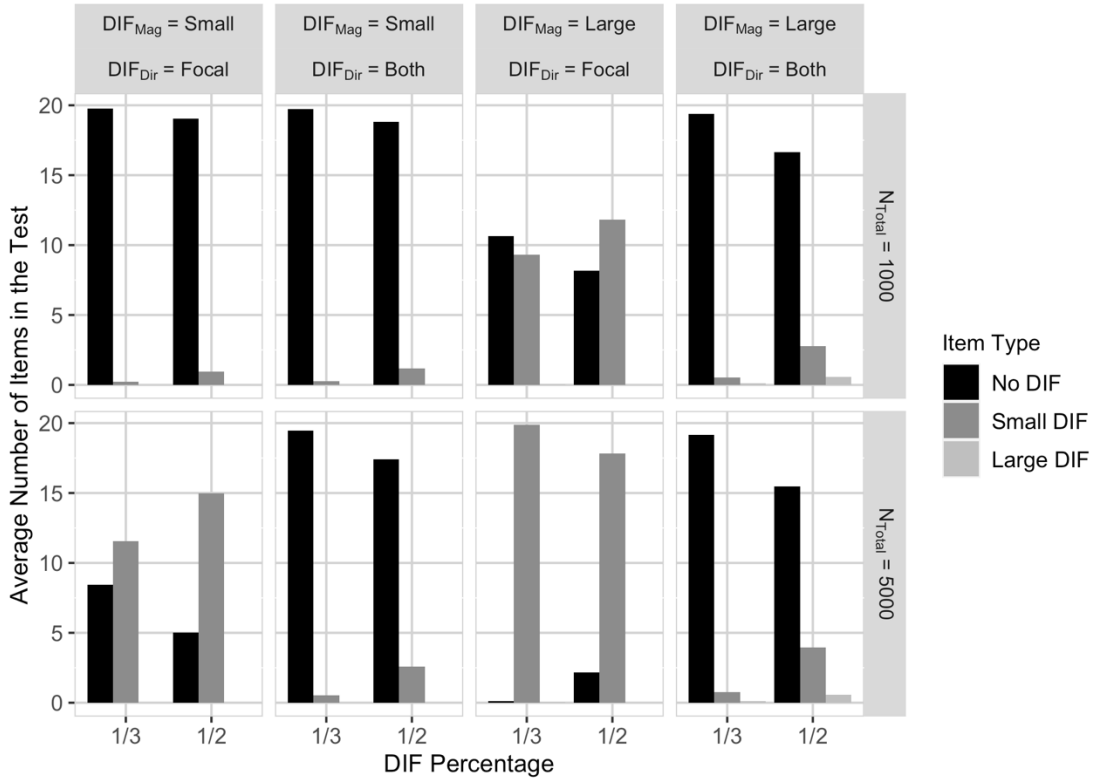


Figure A9. Average uDTF Effect Size for Selected Tests when using IRT-LRT

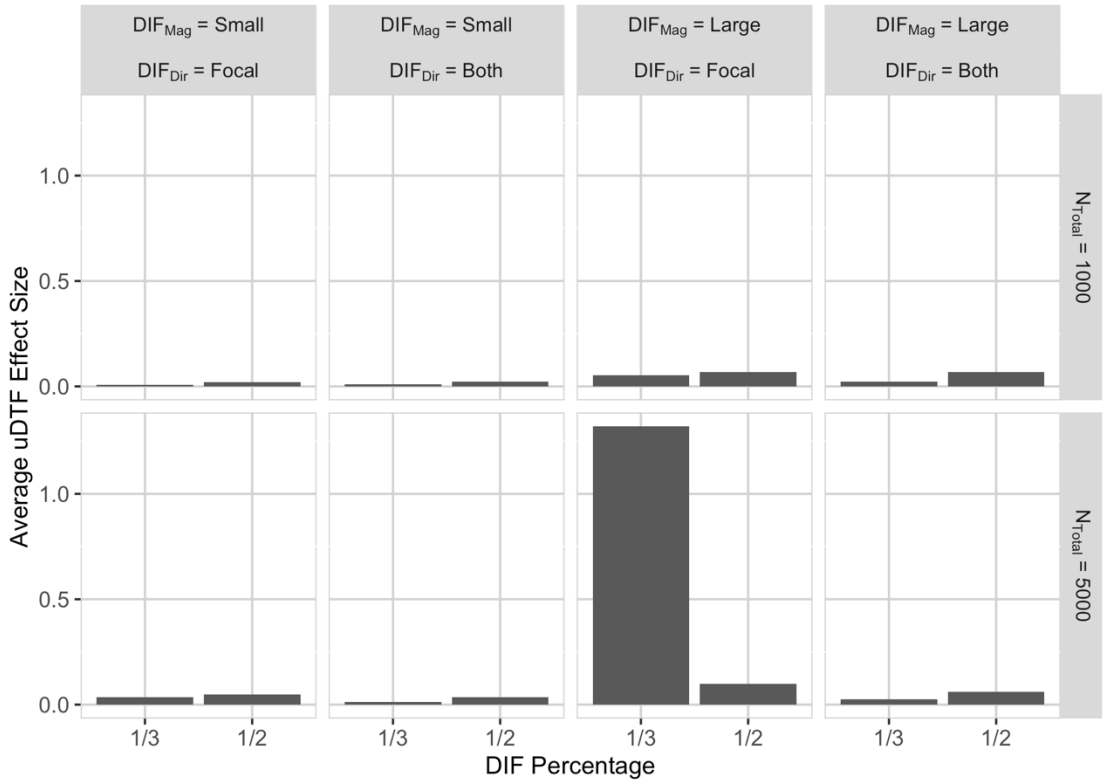


Figure A10. Average Full-Sample RMSEA for Strong MI Models Fit when using IRT-LRT

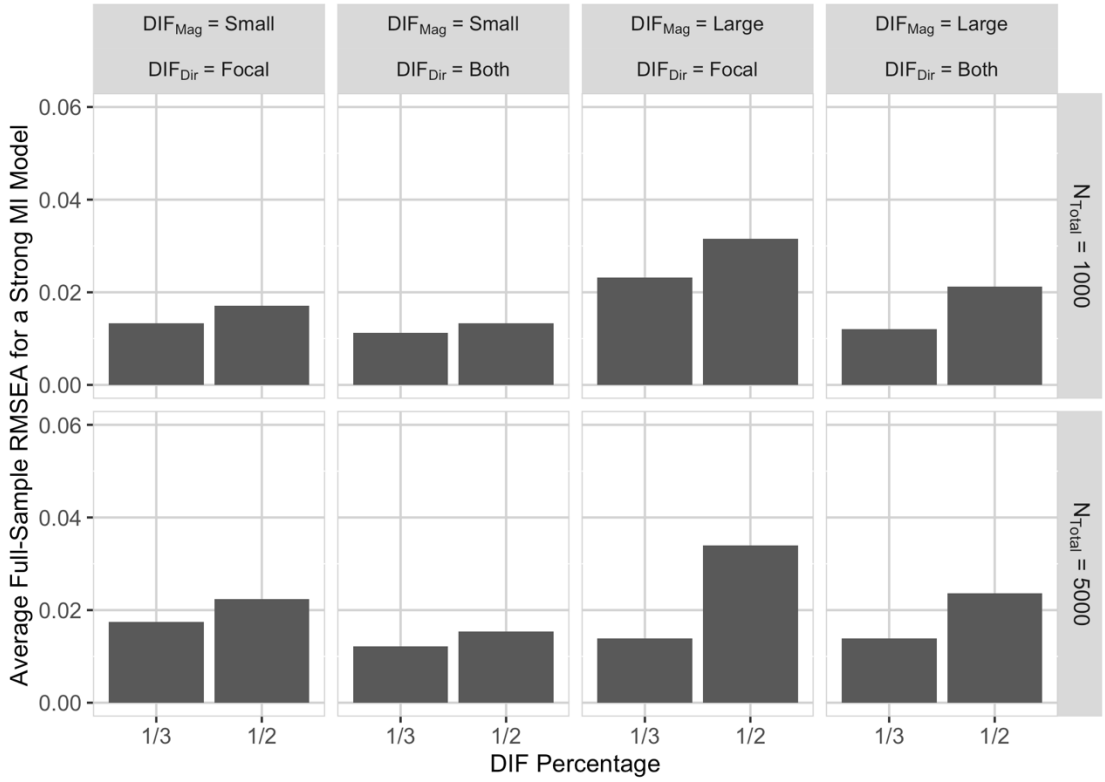


Figure A11. Average Group-Level SRMSR for Strong MI Models when using IRT-LRT

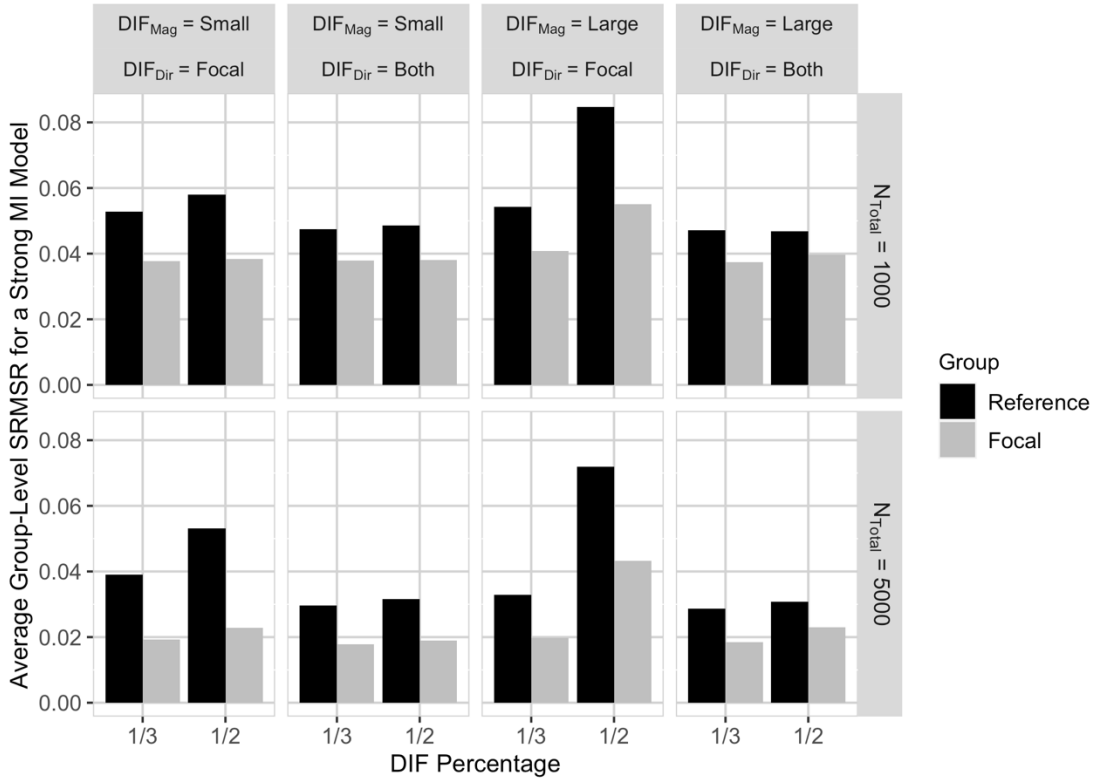


Figure A12. Average Test Information Function Deviations for Selected Tests when using IRT-LRT

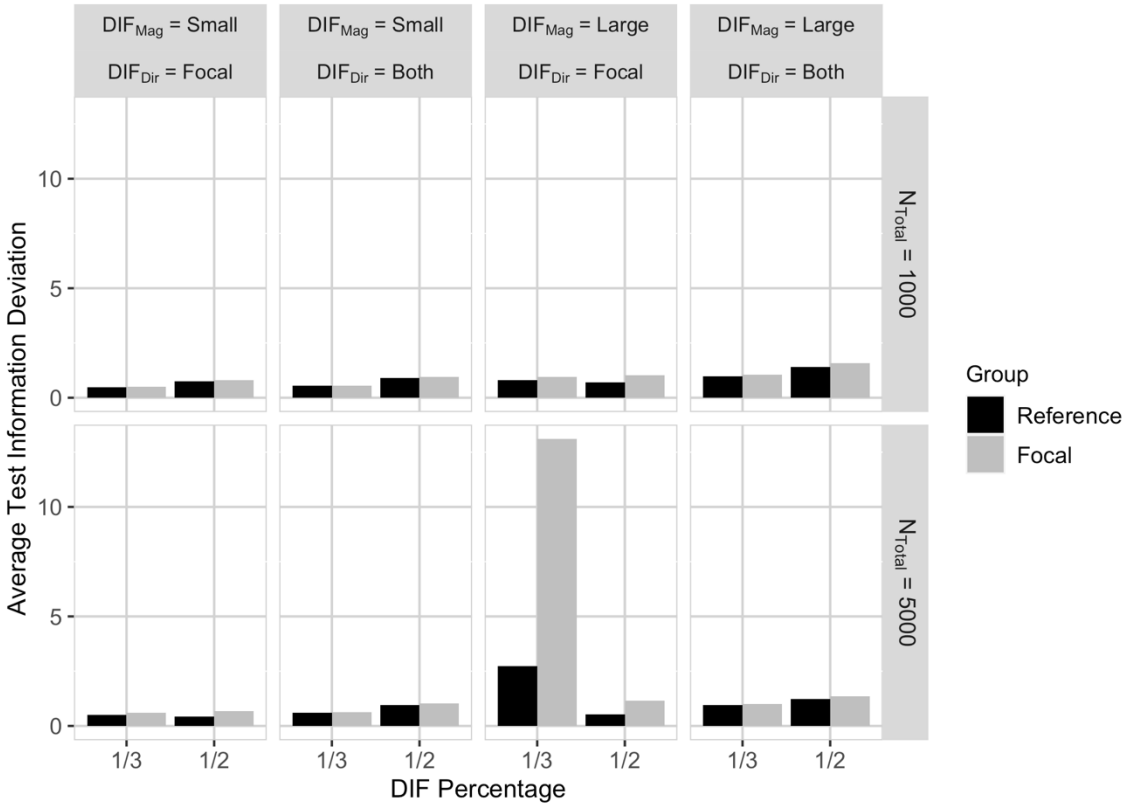


Figure A13. Average Discrimination and Difficulty Values for Items in the Selected Tests Across Weighting Schemes

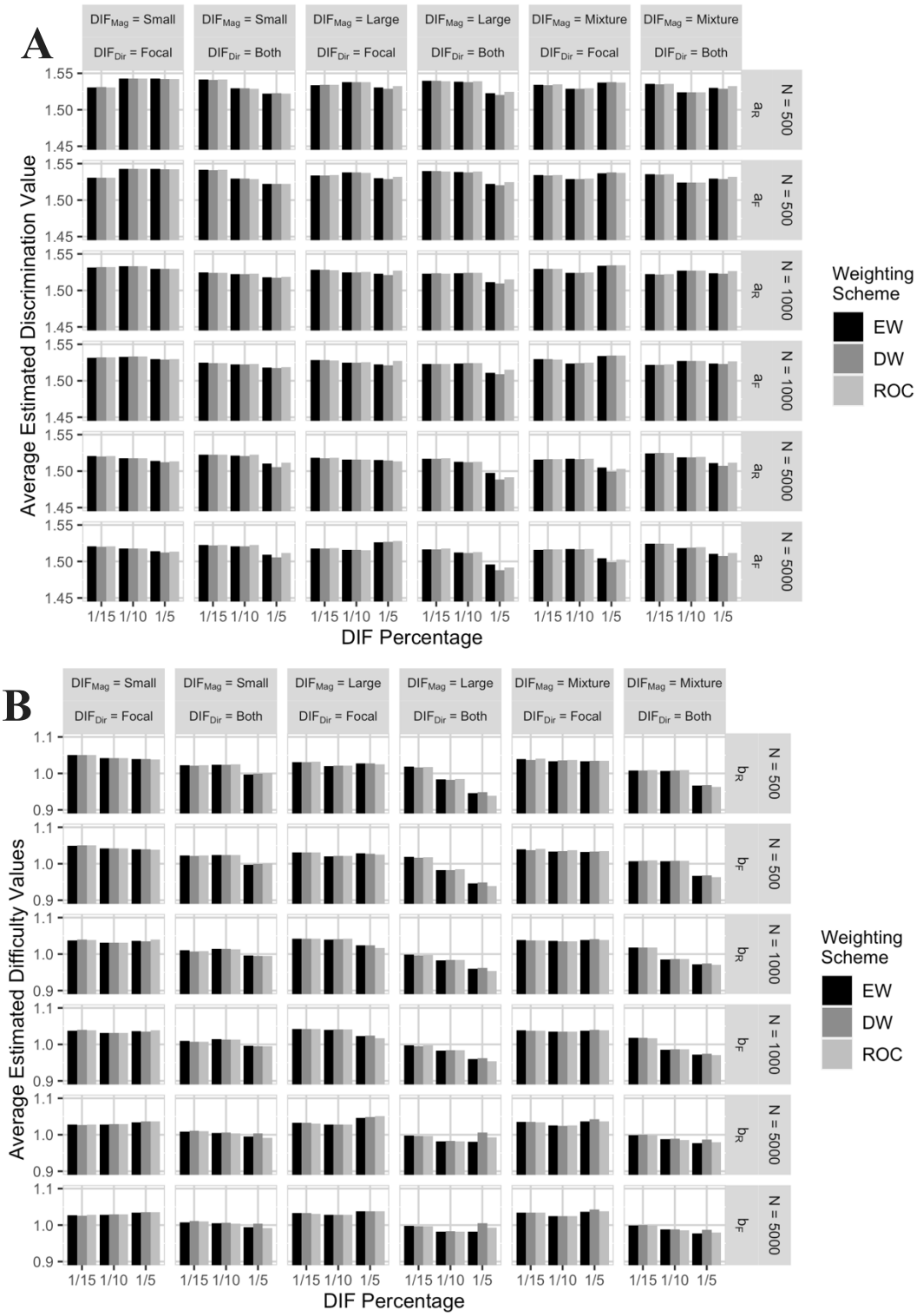


Figure A14. Average Full-Sample RMSEA for Strong MI Models Across Weighting Schemes

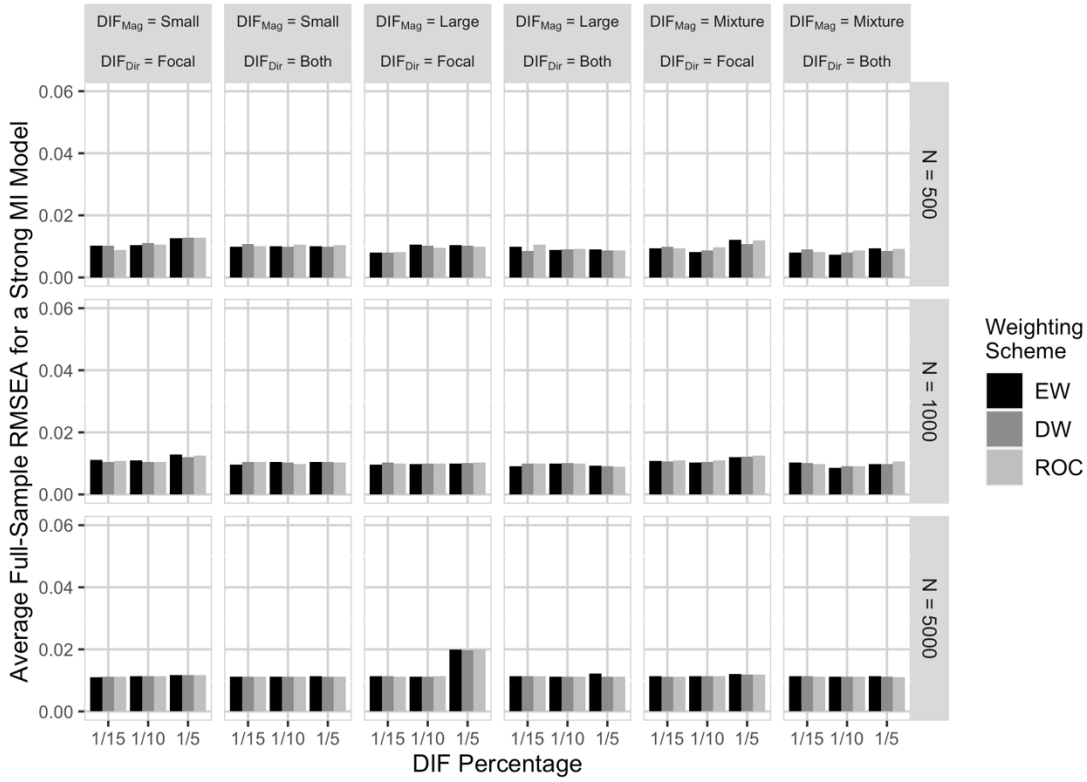


Figure A15. Average Full-Sample RMSEA for Weak MI Models Across Weighting Schemes

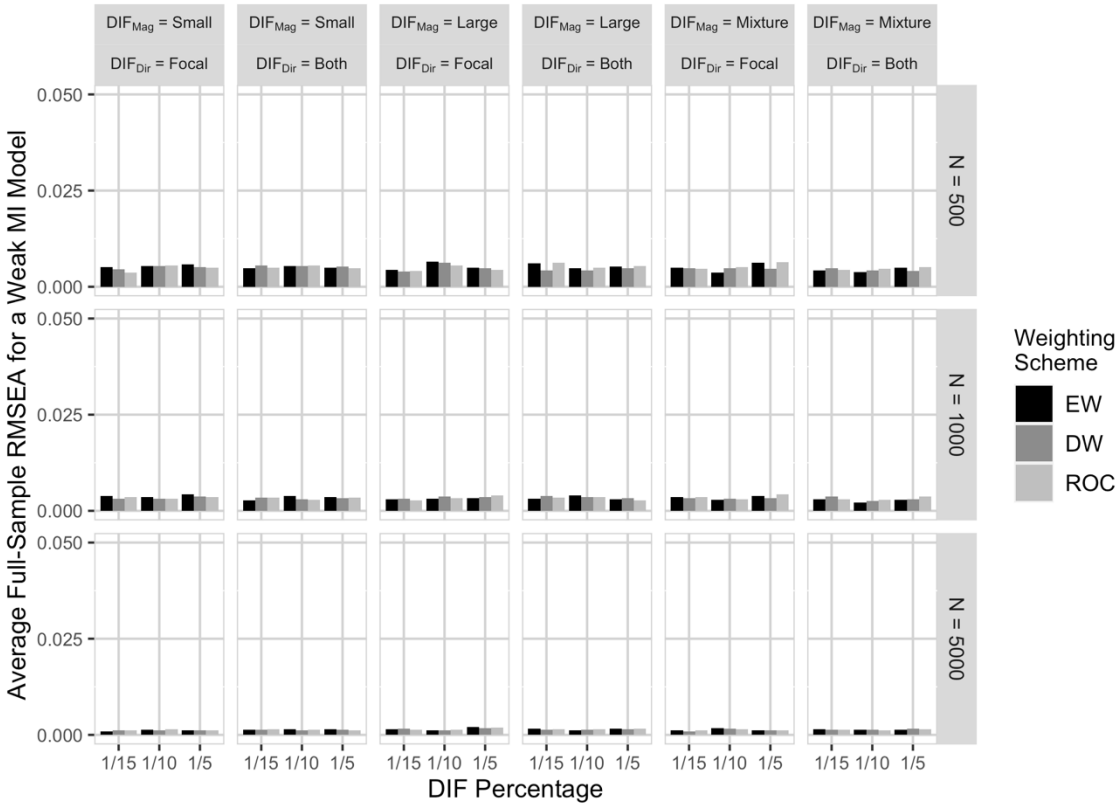


Figure A16. Average Full-Sample RMSEA for Configural MI Models Across Weighting Schemes

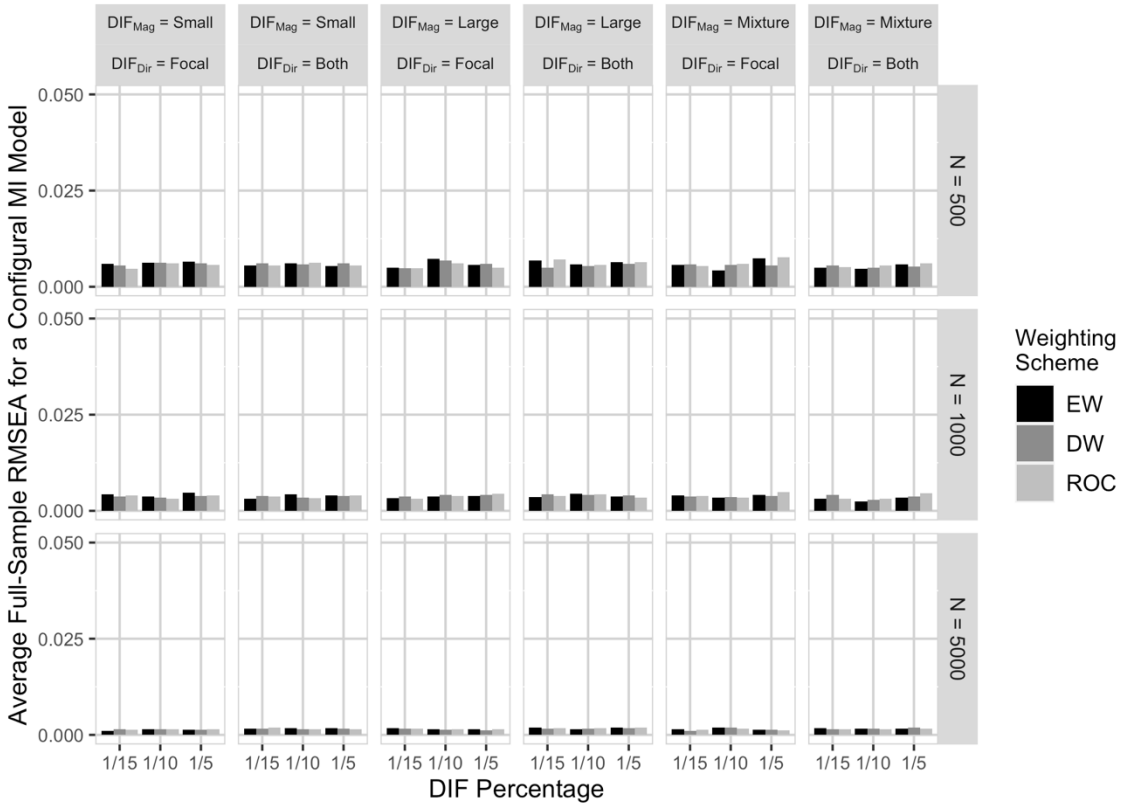


Figure A17. Average Group-Level SRMSR for Strong MI Models Across Weighting Schemes

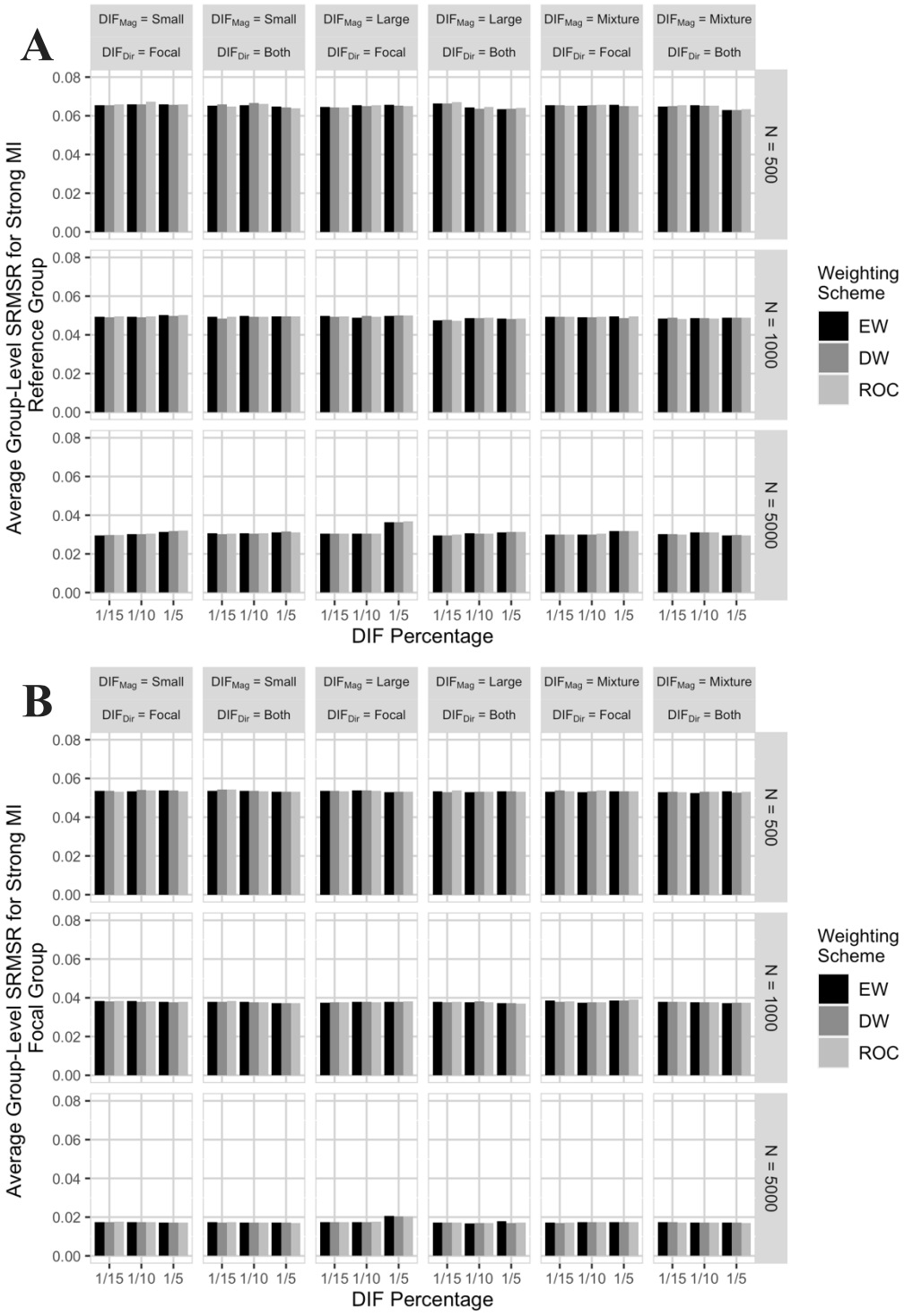


Figure A18. Average Group-Level SRMSR for Weak MI Models Across Weighting Schemes

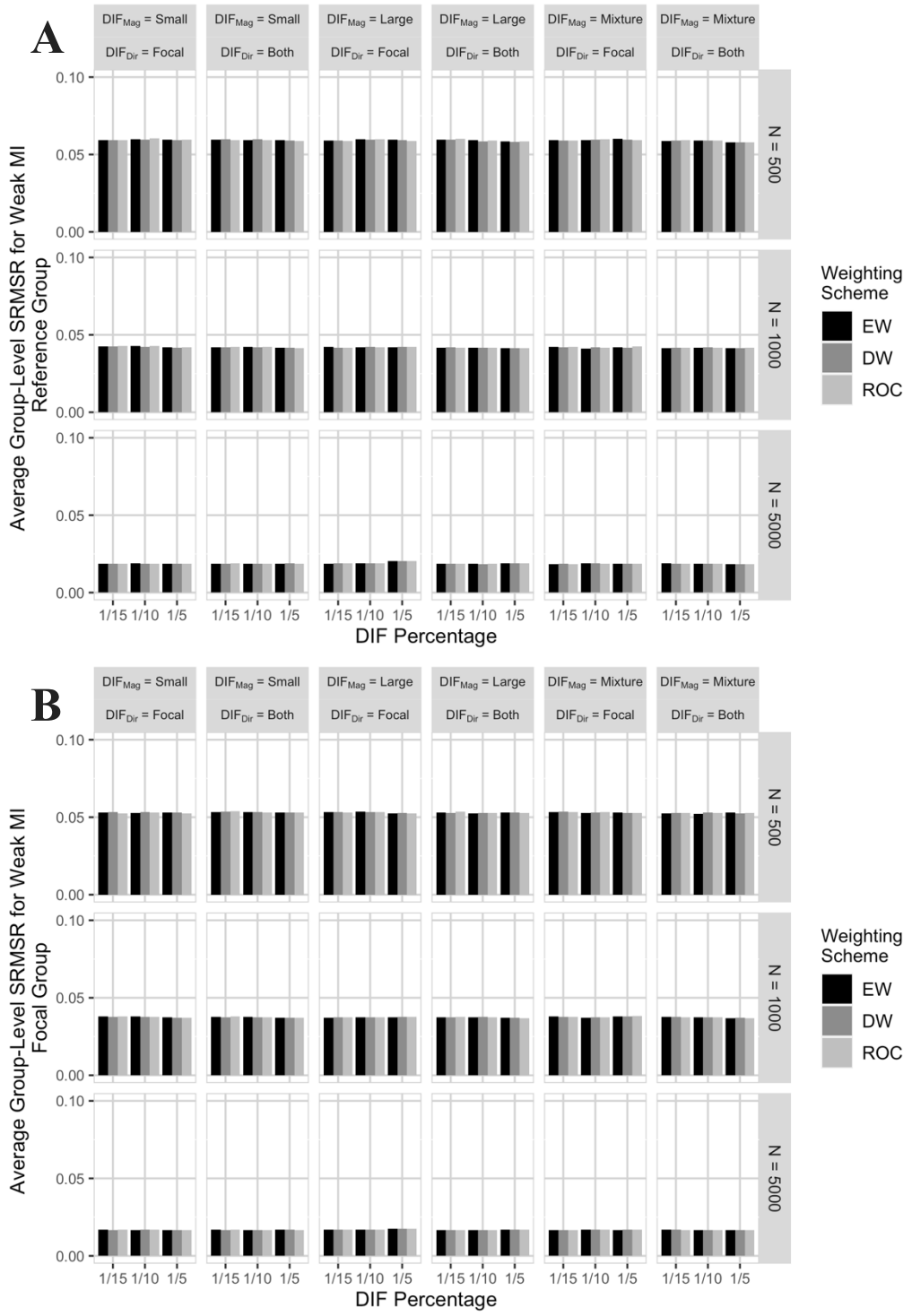


Figure A19. Average Group-Level SRMSR for Weak MI Models Across Weighting Schemes

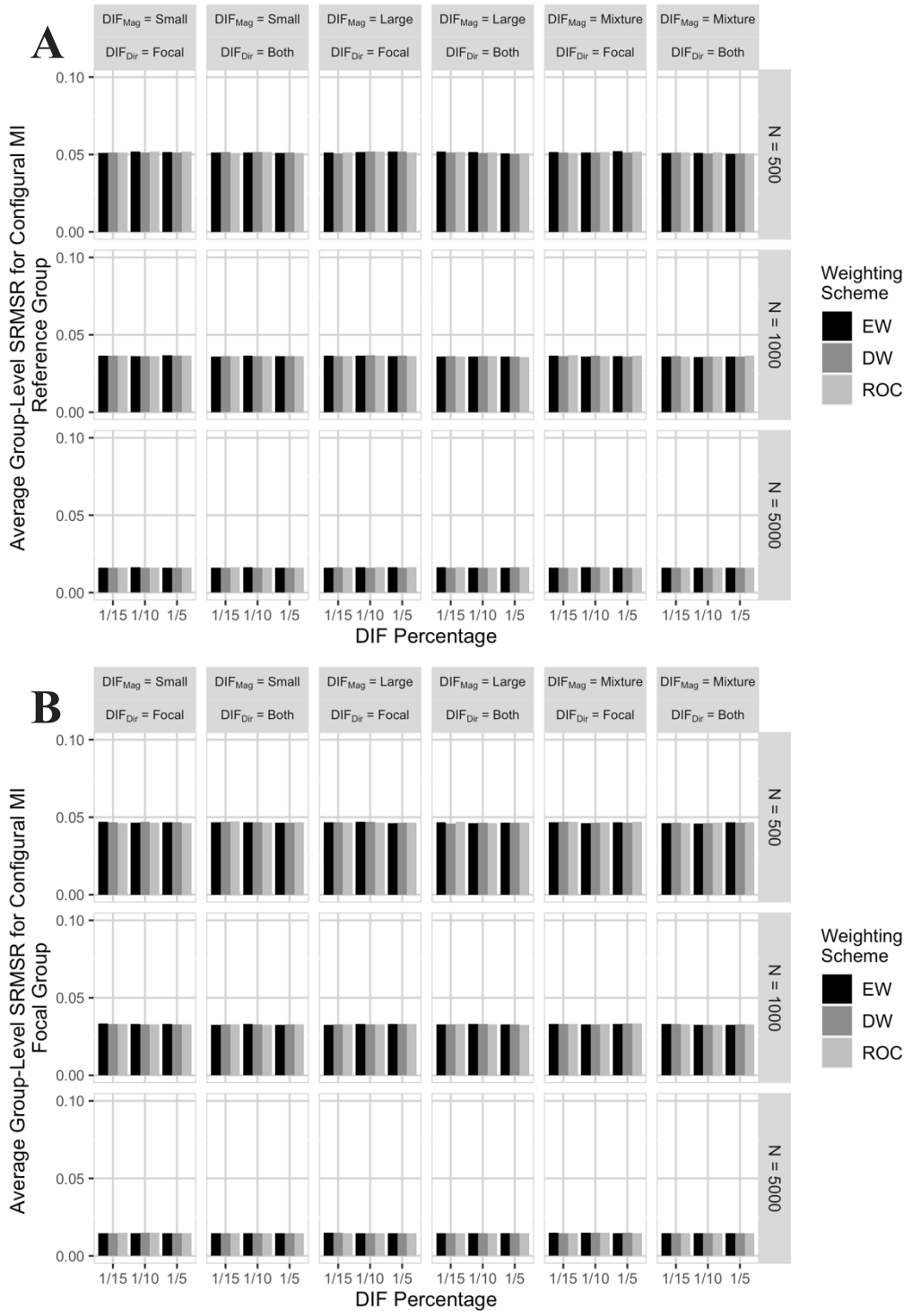


Figure A20. Average Full Test Information Values for Selected Tests Across Weighting Schemes

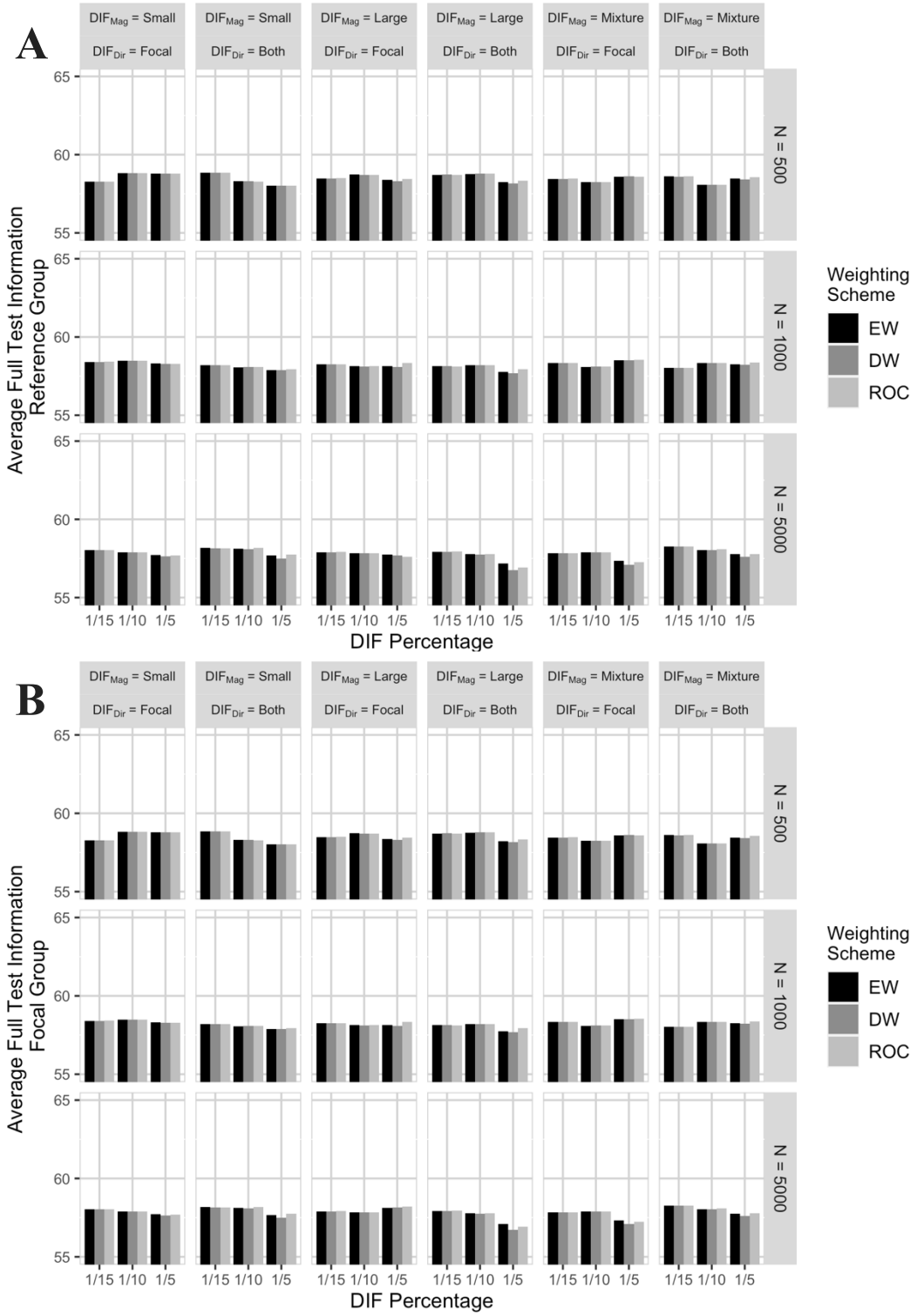


Figure A21. Average Correlations between Selected Tests and an External Criterion Measure Across Weighting Schemes

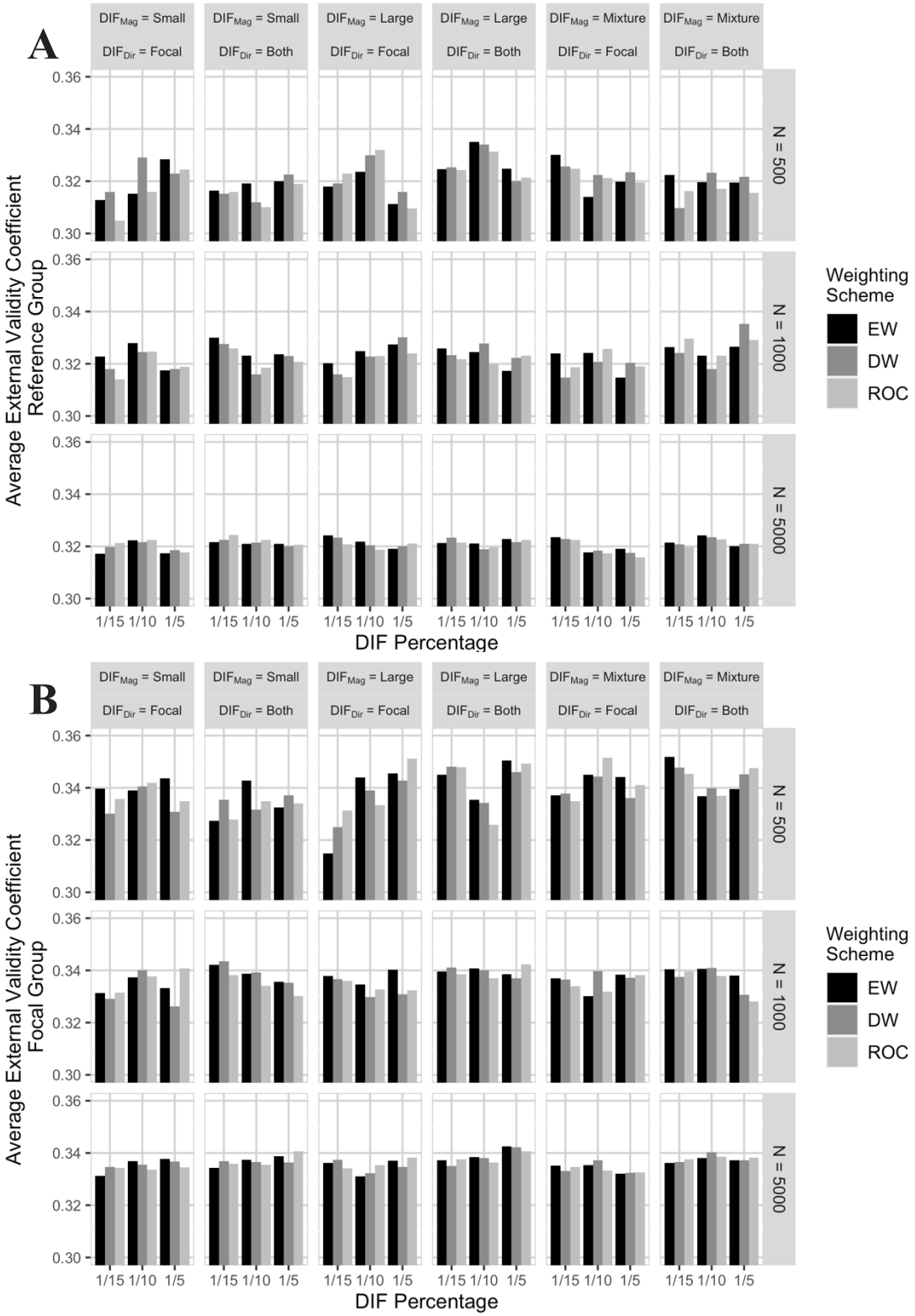


Figure A22. Average False Positive Rates for Regularized DIF with DIF Percentages Ranging from One-Fifteenth to One-Fourth of the Item Bank

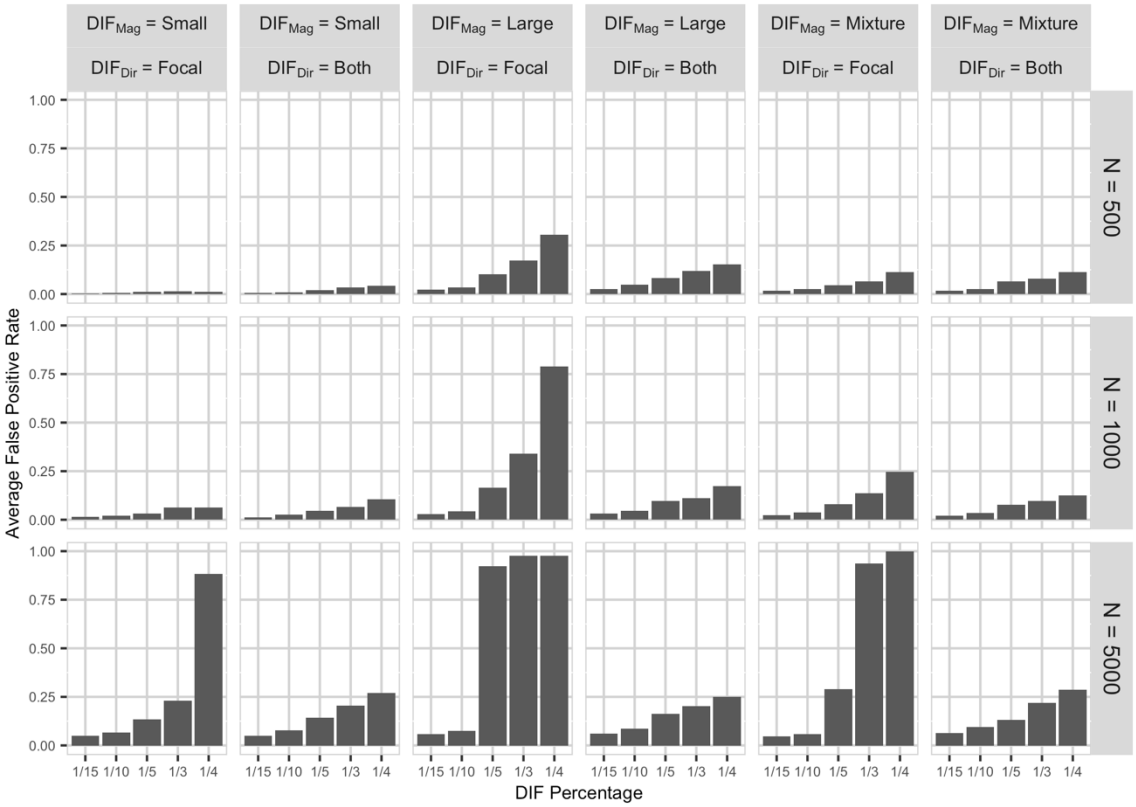


Figure A23. Average Differences in uDTF Effect Sizes between Double and Equal Weighting or between ROC and Equal Weighting

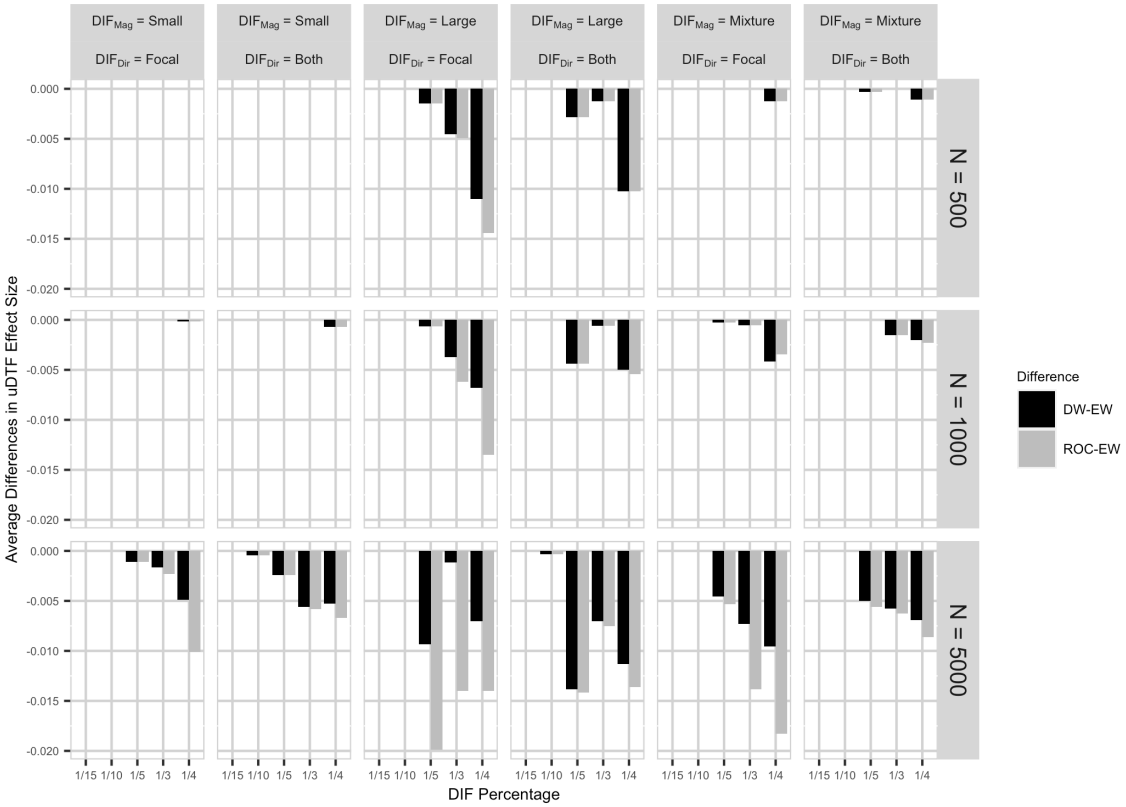


Figure A24. Average Differences in Test Information Function Deviations between Double and Equal Weighting or between ROC and Equal Weighting

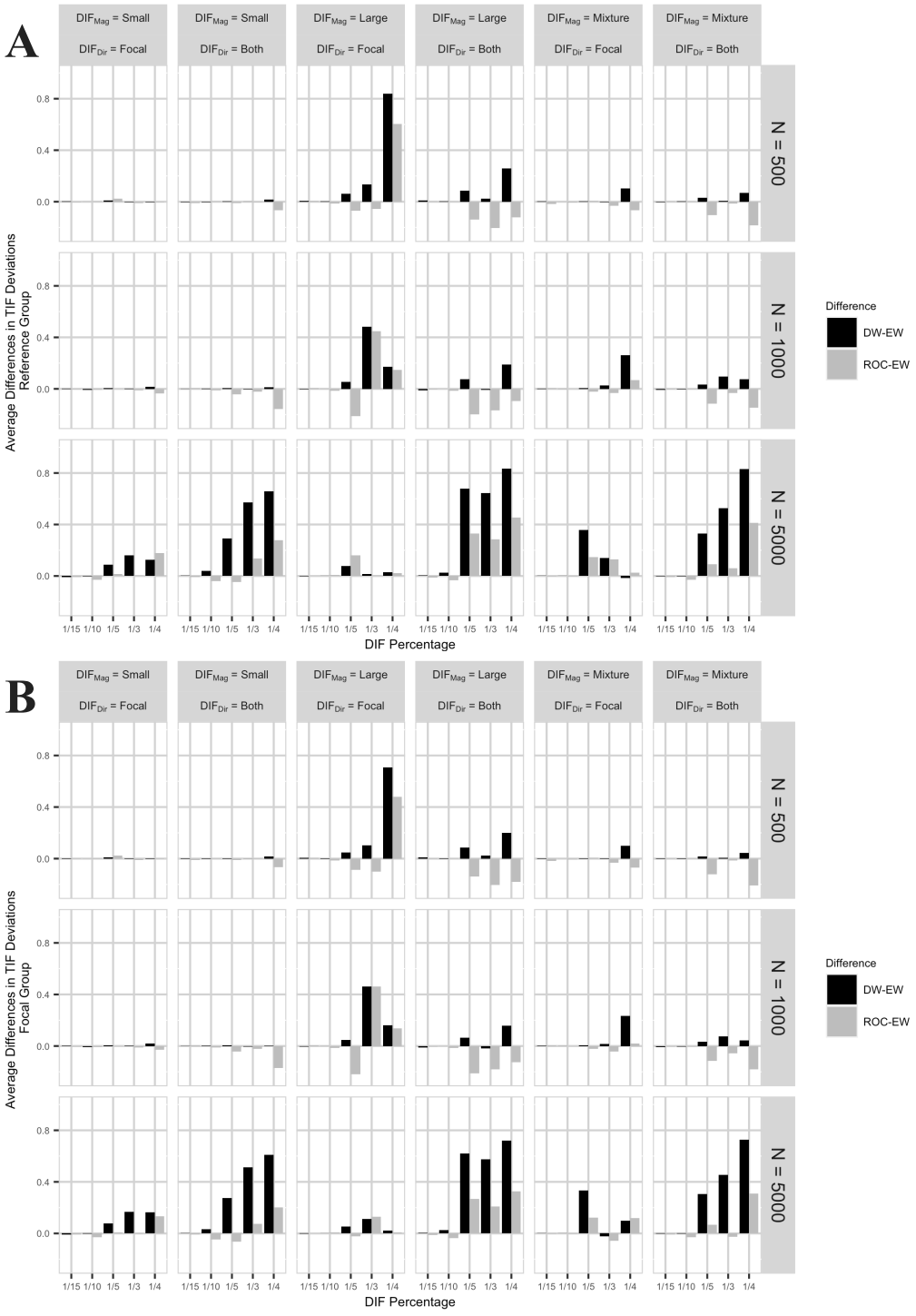


Figure A25. Average Differences in the Number of Well-Fitting Items between Double and Equal Weighting or between ROC and Equal Weighting

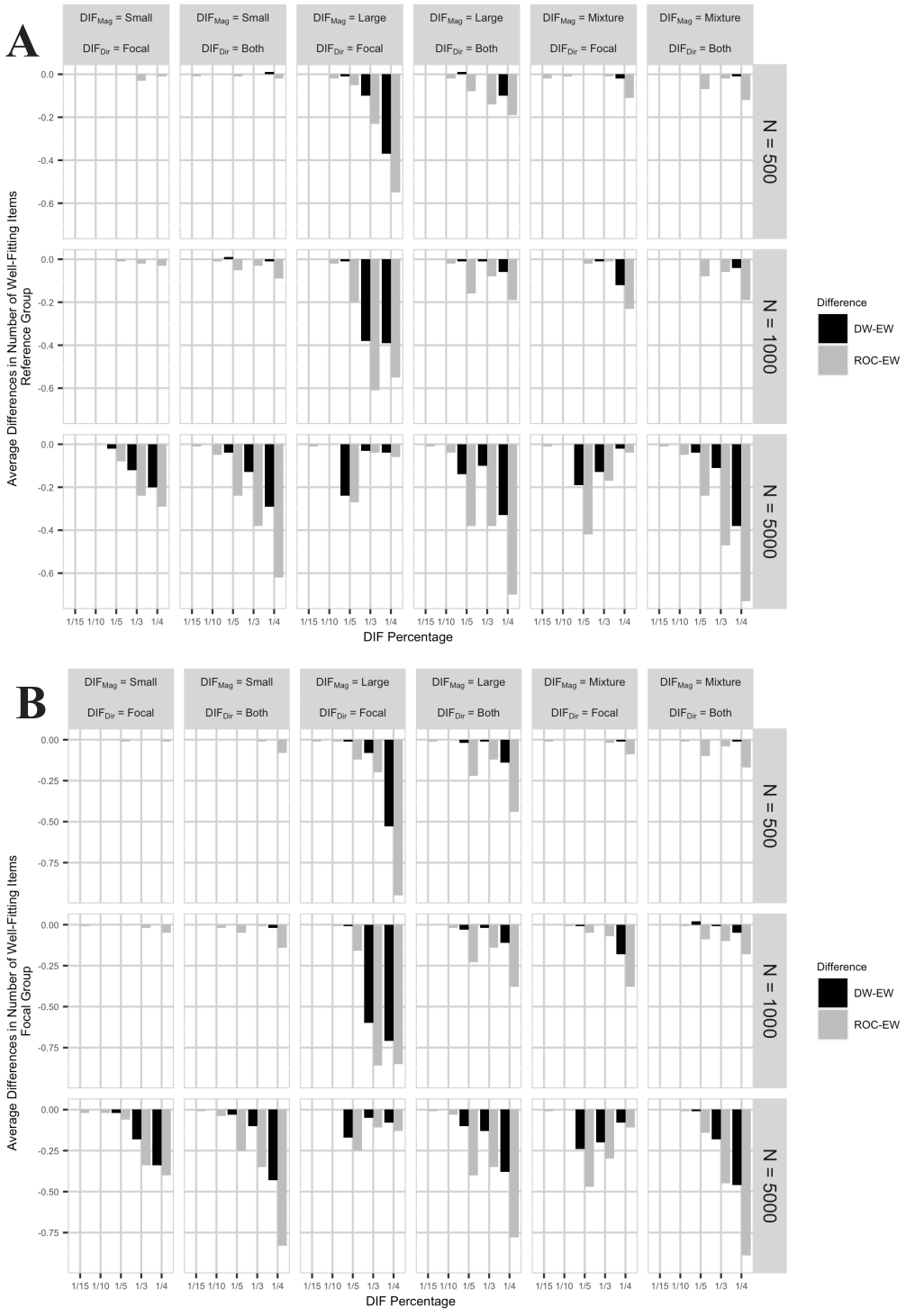


Figure A26. Average False Positive Rates for Regularized DIF with Smaller DIF Percentages in Study 3

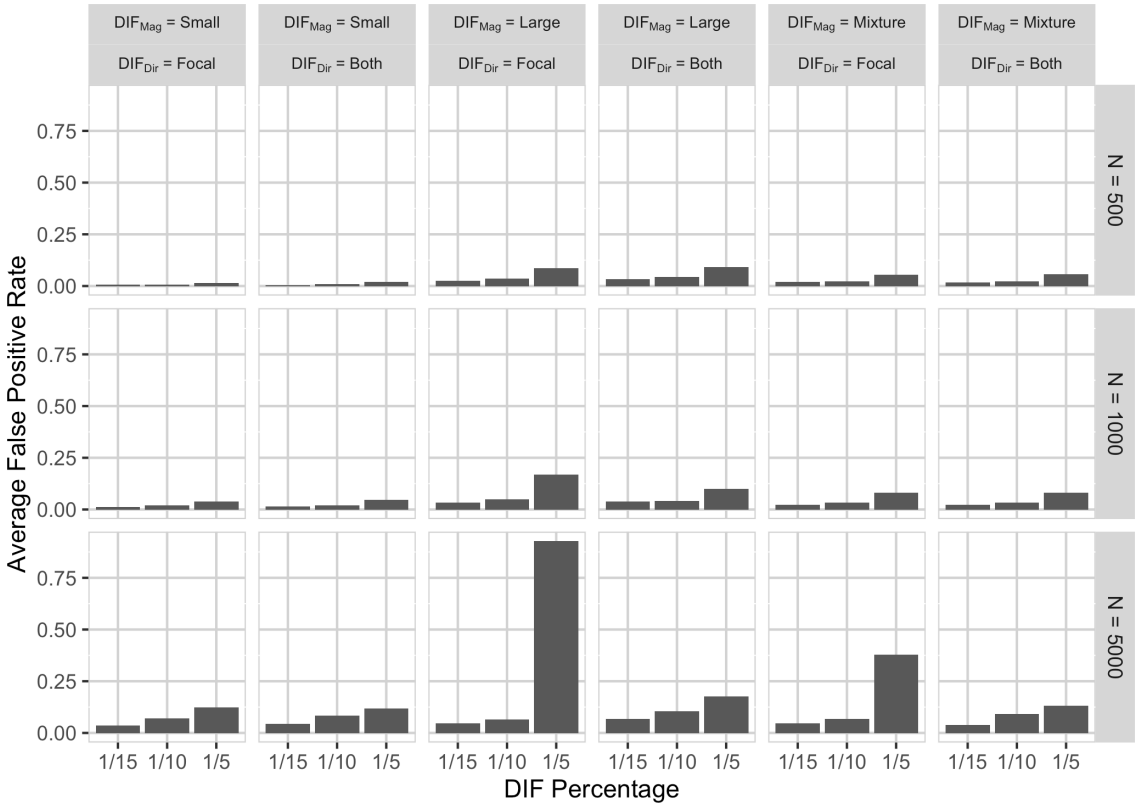


Figure A27. Average True Positive Rates for Regularized DIF with Smaller DIF Percentages in Study 3

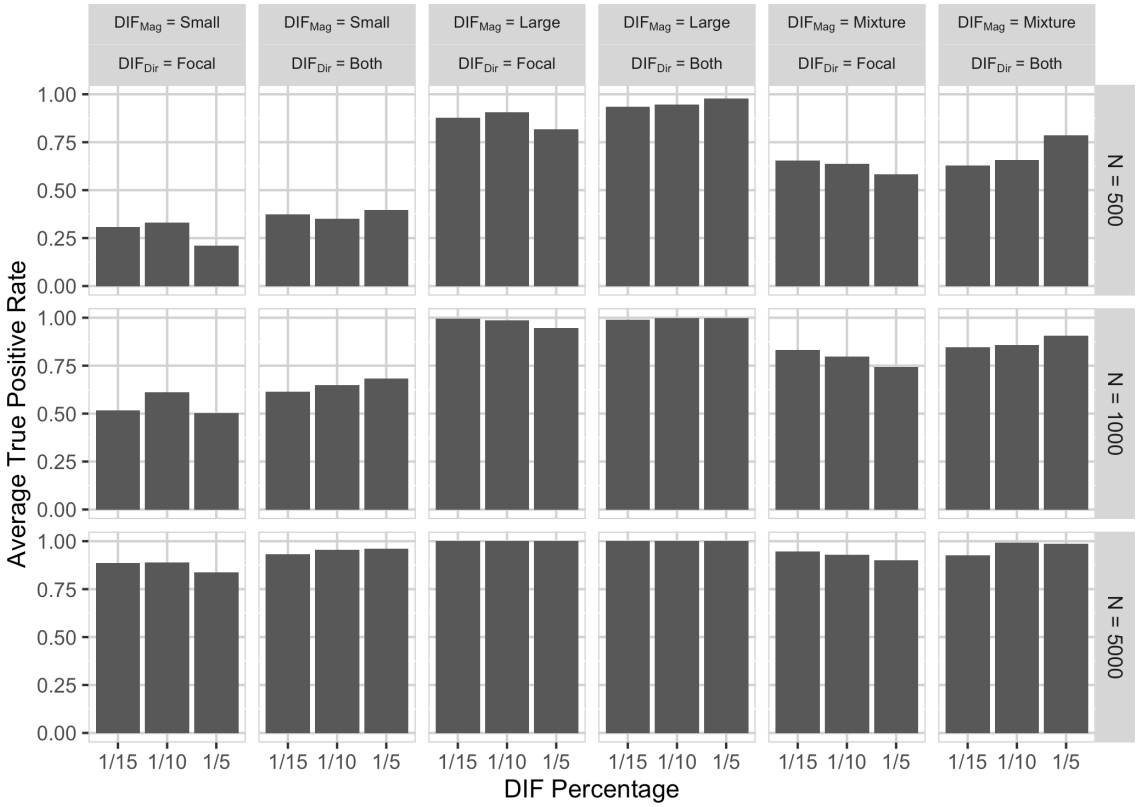


Figure A28. Average Discrimination and Difficulty Values for Items in the Selected Tests by Objective Function Types

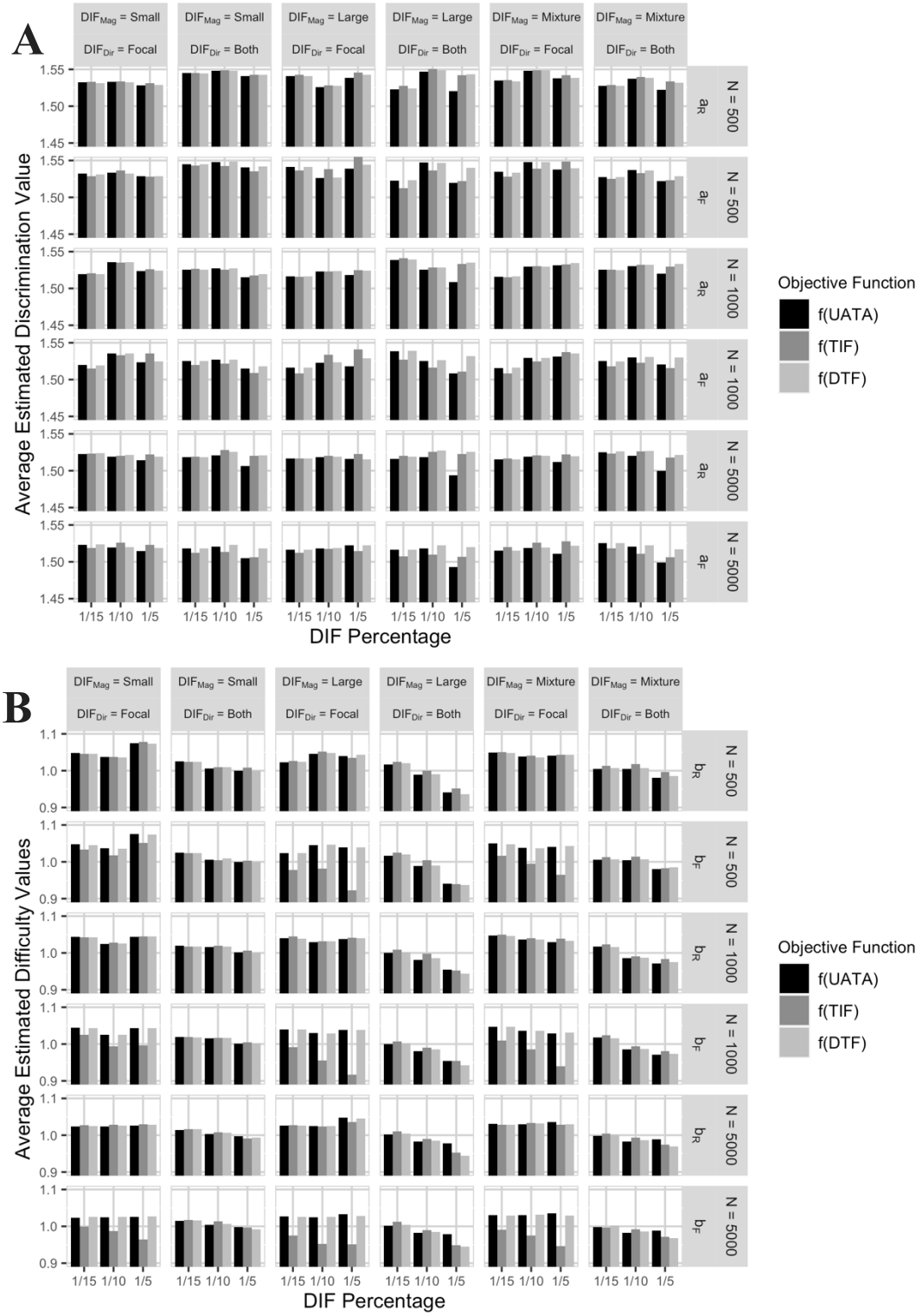


Figure A29. Average Full-Sample RMSEA for Strong MI Models Across Objective Function Types

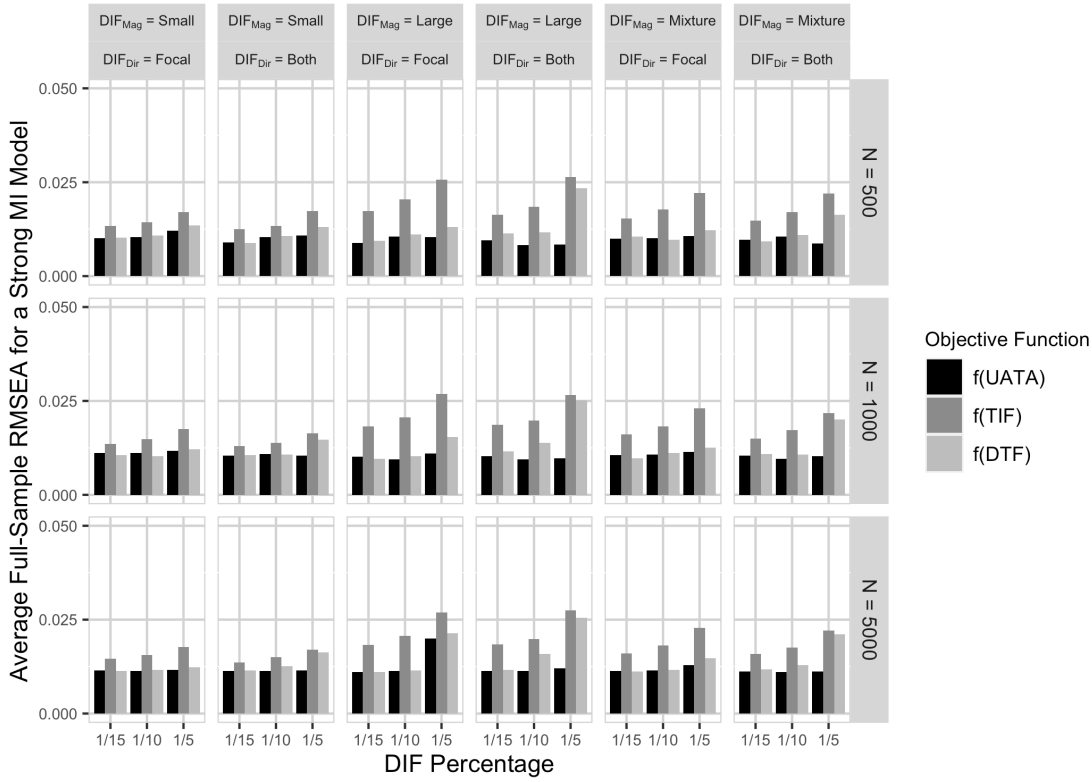


Figure A30. Average Full-Sample RMSEA for Weak MI Models Across Objective Function Types

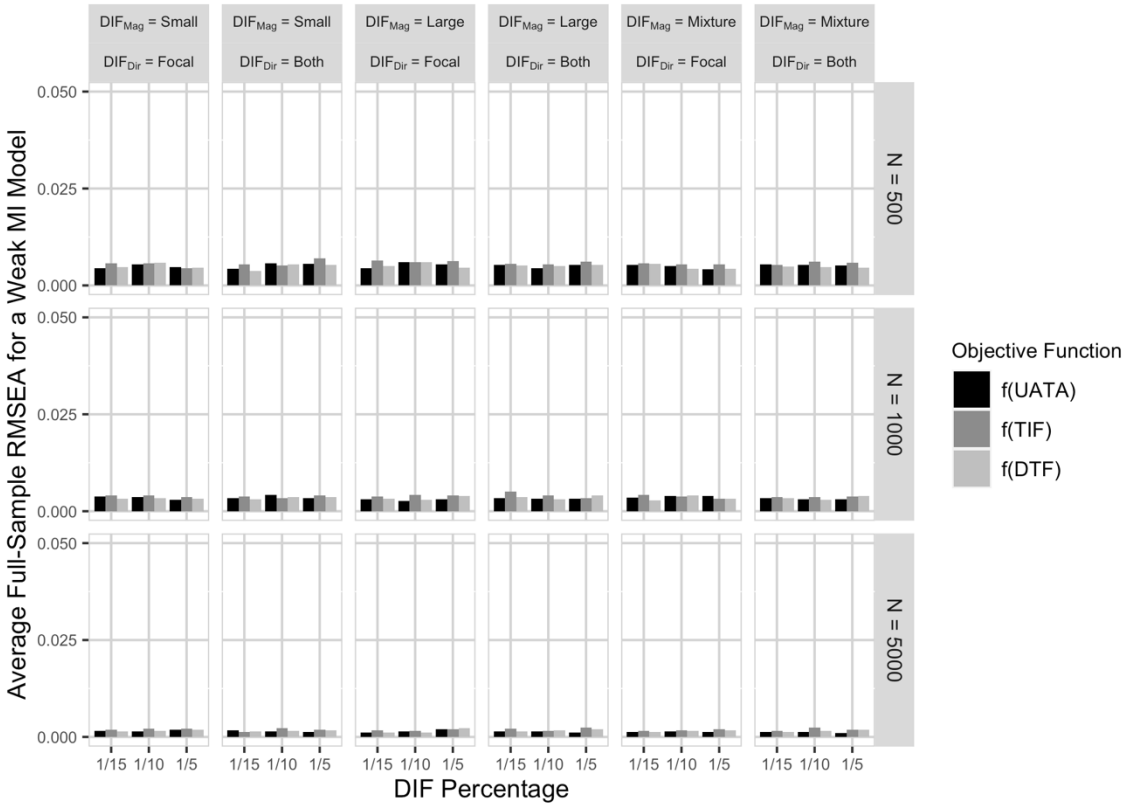


Figure A31. Average Full-Sample RMSEA for Configural MI Models Across Objective Function Types

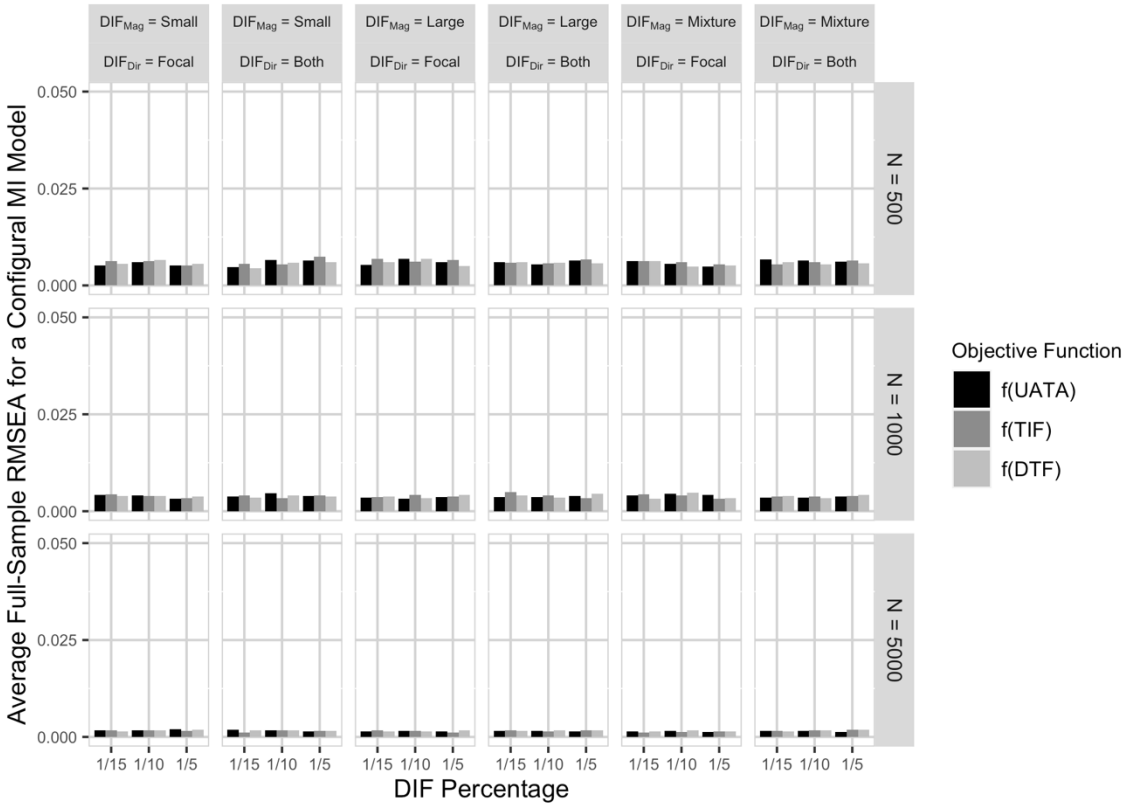


Figure A32. Average Group-Level SRMSR for Strong MI Models Across Objective Function Types

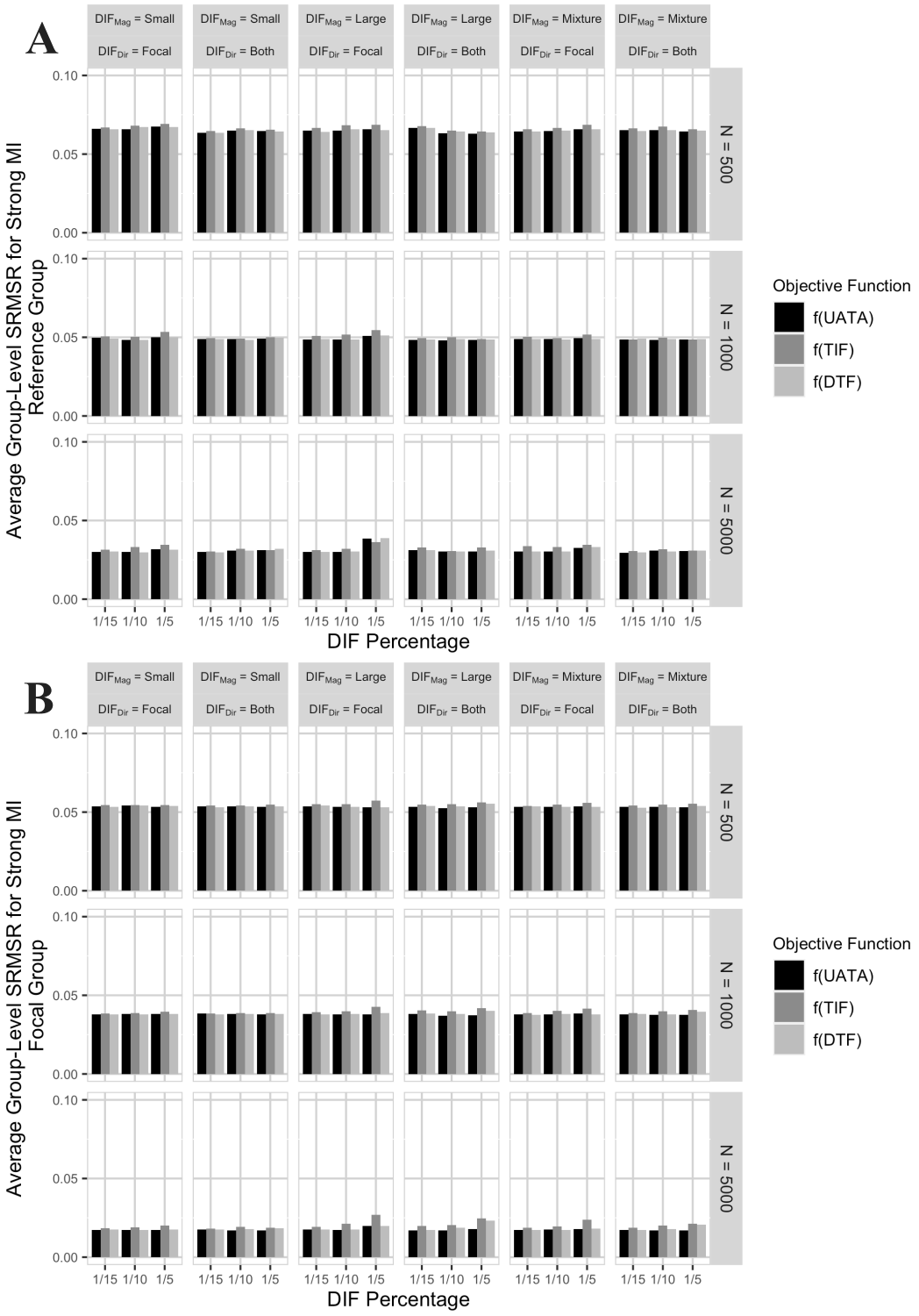


Figure A33. Average Group-Level SRMSR for Weak MI Models Across Objective Function Types

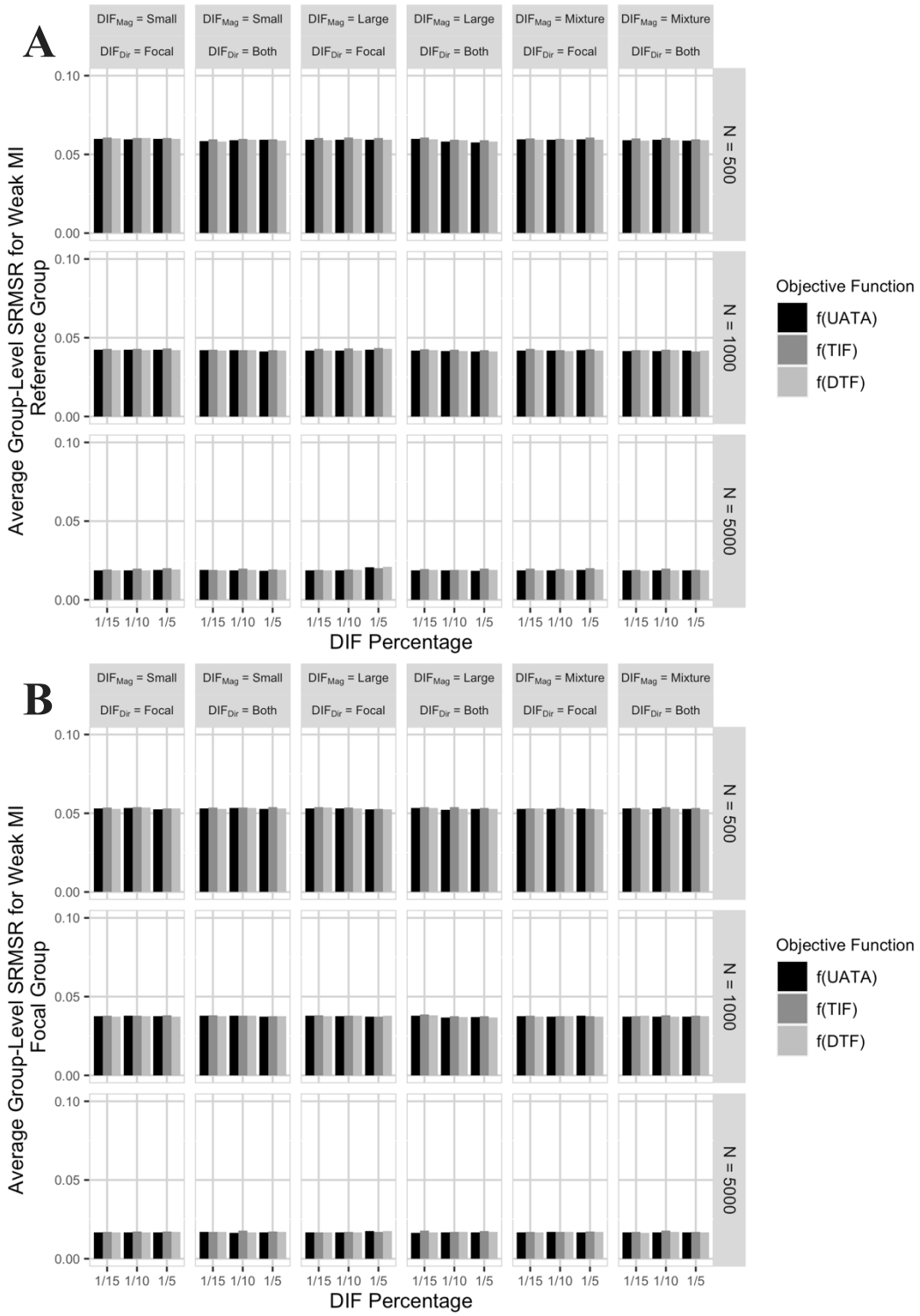


Figure A34. Average Group-Level SRMSR for Configural MI Models Across Objective Function Types

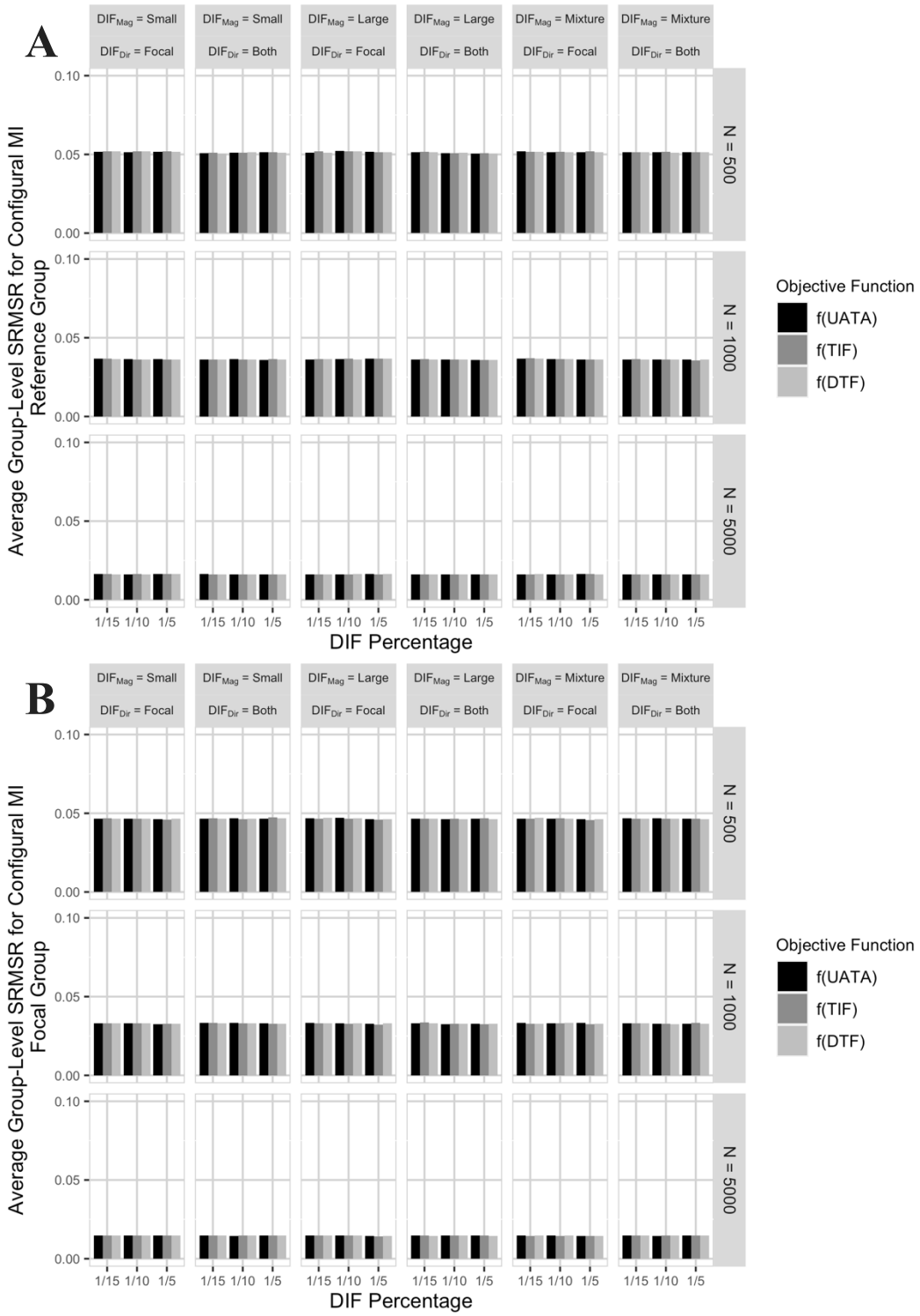


Figure A35. Average Full Test Information Values for Selected Tests Across Objective Function Types

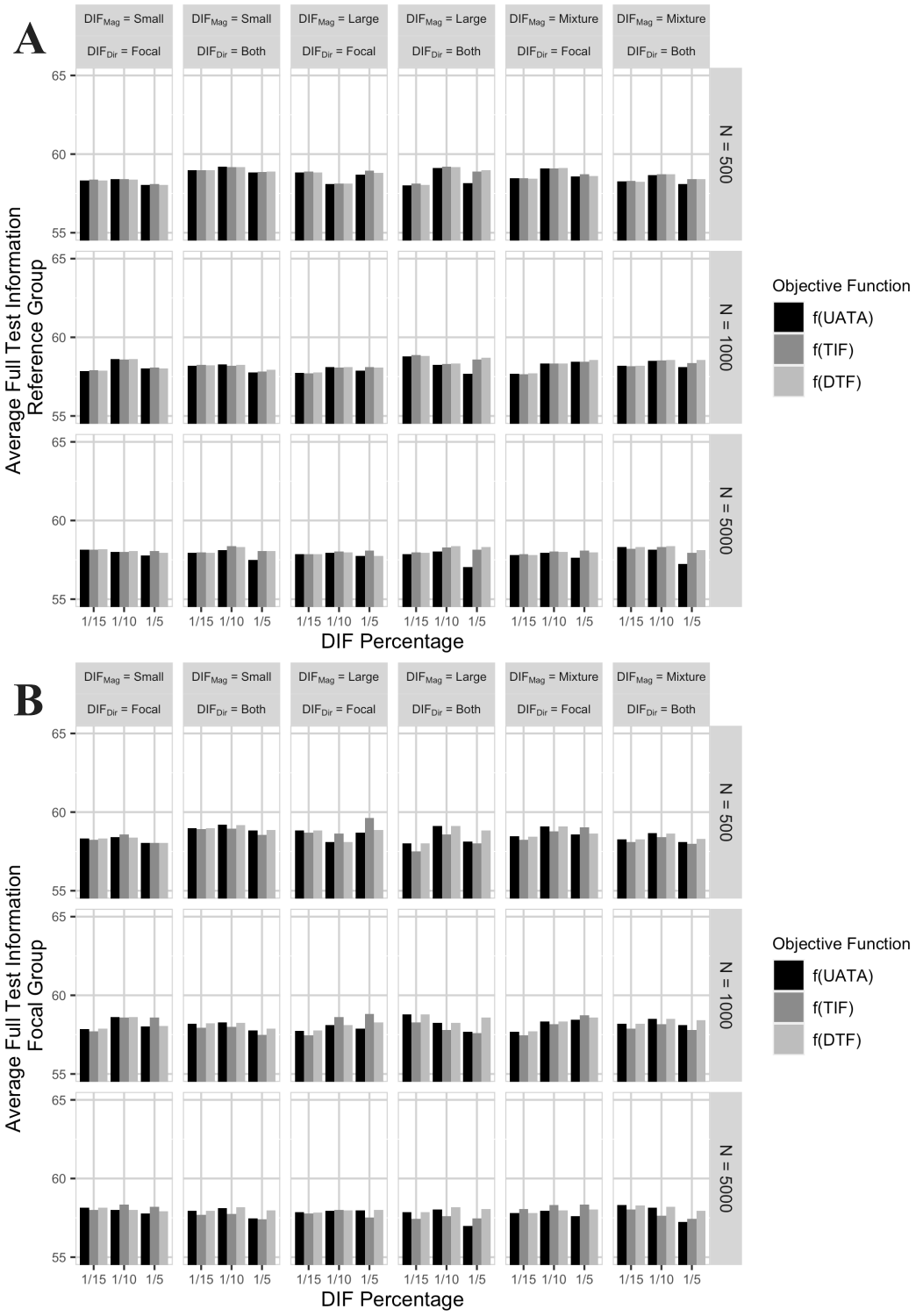


Figure A36. Average External Validity Coefficients for Selected Tests Across Objective Function Types