UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

DIAGNOSTIC CLASSIFICATION IN THE PRESENCE OF INTERSECTIONAL DIFFERENTIAL ITEM FUNCTIONING: A MONTE CARLO STUDY ON IRT AND RANDOM FOREST APPROACHES

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

CATHERINE M. BAIN

Norman, Oklahoma

2026

DIAGNOSTIC CLASSIFICATION IN THE PRESENCE OF INTERSECTIONAL DIFFERENTIAL ITEM FUNCTIONING: A MONTE CARLO STUDY ON IRT AND RANDOM FOREST APPROACHES

A DISSERTATION APPROVED FOR THE

DEPARTMENT OF PSYCHOLOGY

BY THE COMMITTEE CONSISTING OF

Dr. Patrick Manapat, Chair

Dr. Hairong Song

Dr. Danielle Manapat

Dr. Lauren Ethridge

Dr. Jie Cao

Dedication

This is for all those dreams I believed in. This is for all those doubts in my mind.

To the finish lines I have crossed and the starting lines I have yet to toe.

Acknowledgements

To everyone who helped me cross this finish line, thank you. First and foremost, to my committee, thank you for your time, your questions, and your dedication to my growth as a scholar. To Hairong, thank you for welcoming me to the OU Quant program with open arms. To Danielle, thank you for always having an open door and a listening ear. Your passion for students and teaching is contagious. To Lauren, especially, who went out of her way to invest into me as a scholar. The amount of gratitude I have for what you've taught me and the role you played in my education cannot be fully expressed in the acknowledgements section. Please know I am honored to have had the opportunity to learn from and work with you. Finally, to Patrick, words cannot adequately express my gratitude towards you. So, I will just say this: thank you for taking a chance on me, for believing in me, for your time and effort, for your ability to proofread, and for always pushing me to be the best academic (and best person) I can be.

To Ankur Gupta, Krista Cline, Brian Giesler, Robert Terry, Dingjing Shi, Jordan Loeffelman, Yutian Thompson, Jane Silovsky, Lara Mayeux, Jenel Cavazos, and George Bogaski; you are all proof that it takes a village.

Thank you to the friends I've gained. Jordan, I am incredibly grateful to know and be known by you. Thank you for the long walks, runs, and conversations. Thank you for always listening to me, for reminding me that I do in fact know things, and for being the best co-author a girl could ask for. I am honored to have toed this starting line with you, and I feel so incredibly lucky to have you here at the finish line. To Raechel, for always asking what we both think would be simple questions that never fail to make me read at least three new papers. I am a better academic because of you. To Destiny, thank you for always helping me find the words. I wish you could help me find these. To Brenna, thank you for keeping me sane and reminding me that

sitting outside can solve a lot of problems. To Hannah, thank you always being willing to cause a ruckus or send a box of wasps. To Josie and Dana, thank you for always being down to yap, no matter what about. To Keelyn, thank you for being the best officemate a girl can have. To Irene, thank you for the late-night coworking sessions. Many a word of this dissertation were written during those.

Thank you to the friends who were here before and are still here now. To Emma, to be known is to be loved. Thank you for knowing me. To Ash, thank you for our semi-annual phone calls that remind me why I do this. To Margaret, Katie, and Liv, thank you for pulling me out of the grad school bubble. To my parents, thank you for your support and for always letting me chase my dreams. To my siblings, thank you for always humbling me—I think it's safe to say I'm a certified nerd now. To Terri and Phil, Rob and Ronnie, and Pete and Emma, thank you for cheering me on these last few years.

Lastly, thank you to Maggie, I would not have gotten here without you. Thank you for reminding me that life lives in the little moments. You are the best partner I could ask for, and I am so thankful to know and be known by you. Thank you for your patience, for the early morning phone calls, and for the letters. Thank you for being you and cheering me on the whole time.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Abstract

Psychological researchers frequently use self-report scales to classify individuals into diagnostic categories. Item response theory (IRT) is the recommended family of mathematical models for scoring psychological scales. The accuracy of IRT scores is dependent upon the extent to which the assumptions invoked by the IRT model are met. Techniques such as multi-group IRT allow for more accurate estimation of scores when the assumption of item invariance is violated. However, if the grouping variable causing differential item functioning (DIF) is complex, unmeasured, or unknown, modeling DIF through the IRT framework becomes unreasonable. Random forests (RF), as a nonparametric machine learning approach, do not assume item invariance and may be a reasonable approach to classification when DIF is complex, unmeasured, or unknown. Using a Monte Carlo simulation, this study examined the performance of three IRT-based approaches and one RF-based approach to classification when data contained complex DIF caused by multiple demographic variables. Classification performance was evaluated across six metrics: accuracy, sensitivity, specificity, precision, negative predictive value, and F1 score. Results indicate that the classification performance of the second-order IRT model-based approach was higher than that of either of the multidimensional IRT model-based approaches. Higher classification performance was observed from the RF-based approach than the second-order IRT-based approach in most conditions. However, in conditions where the focal group was simulated to have the same or higher levels of the latent trait than the reference group, second-order IRT-based classifications were more sensitive than RF classifications. Additionally, the classification performance of the RF approach was strongly associated with measurement error, such that more measurement error led to lower classification performance.

Introduction

Psychological self-report scales are widely used in research to stratify individuals into diagnostic classes. In research contexts, these scales often supplement or replace more extensive diagnosis tools, informing decisions about participant inclusion, group comparisons, and variable modeling in large-scale studies (de Lauzon et al., 2004; Jacobsen et al., 2024; Öztürk Özkan et al., 2025). For example, researchers have used the three-factor eating questionnaire (de Lauzon et al., 2004) to stratify individuals into diagnostic groups when investigating the relationship between social media use and disordered eating (Öztürk Özkan et al., 2025), the depression anxiety stress scale (Jacobsen et al., 2024) to compare coping strategies among mental health practitioners with and without depression or anxiety (Stals et al., 2024), and the schizotypal personality questionnaire (Chmielewski & Watson, 2008) to examine social, cognitive, and emotional functioning in individuals with schizotypal personality disorder (Aghvinian & Sergi, 2018). In some research designs, diagnostic class membership determines study condition assignment—as in regression discontinuity designs where treatment is assigned based on a symptom severity threshold (e.g., Krantz, 2022).

In the context of clinical diagnostic assessment, the best available estimate of an individual's true level of the diagnostic construct is typically a structured clinical interview or laboratory test.[1] In practice, the results of these tests are often treated as the "gold standard" estimate against which all other measures of the diagnostic construct should be judged. Direct assessment via this gold standard, however, usually requires testing and documentation of symptoms by a licensed clinician and is often costly, time-intensive, and potentially invasive. For

[1] I highlight that these are best understood as the best available estimate, not a fixed ground truth, as even such assessments do not result in error-free measurements of the latent trait (Buss et al., 2023; Haslam et al., 2020; Watts et al., 2024; Widiger & Samuel, 2005).

these reasons, the use of a gold standard may be unreasonable in many research contexts. Instead, researchers use the items of a psychological scale as observed variables to estimate an individual's level of the latent construct (Borsboom et al., 2003; Lord & Novick, 2008). These estimates are known to be imperfect because they contain some degree of noise resulting from question wording, participant fatigue, or inattention; differences in the latent variable and observed variable measure; or additional factors (Jacobucci & Grimm, 2020; Lord & Novick, 2008; McNeish, 2022; Rhemtulla et al., 2020; Vowels, 2023). When assessing the utility of a scale, researchers must quantify the degree of noise (i.e., measurement error) that is present in these estimates. Put another way, it is important to determine if scores from a given scale reasonably approximate those of the latent construct (DeVellis, 2012; Flake et al., 2022; Monaghan et al., 2021). In the context of scales designed to measure a clinical construct, best practices recommend assessing the degree to which estimates from the scale predict individuals into the same diagnostic class as would the gold standard.

Some amount of misclassification is to be expected, as no scale score is without measurement error (Furr, 2020; Gonzalez, 2021; Jacobucci & Grimm, 2020; McNeish, 2022; Rhemtulla et al., 2020; Vowels, 2023) and self-report scales are not synonymous with a clinical diagnosis (Kohrt et al., 2025). However, misclassification can also occur as a function of misspecification of the scoring model. Scoring models invoke certain assumptions, and the extent to which these assumptions are met (or violated) partially determines how precisely individual trait levels can be estimated and, consequently, how accurately individuals can be classified (Emons et al., 2007; Flake et al., 2017, 2022; Flake & Fried, 2020; I. Han et al., 1996; Hussey & Hughes, 2020; Messick, 1995; Salloum et al., 2024; Shaw et al., 2020; Vowels, 2023).

Often in practice, psychological scales are assumed to be unidimensional, determining classifications based on a single, continuous score (Avila et al., 2015; McNeish, 2024, 2026; McNeish & Wolf, 2020; Stickle & Weems, 2006). As such, the literature on classifying individuals based on scores from self-report scales has largely focused on the unidimensional case (Bain et al., 2026; Gonzalez, 2021; Gonzalez et al., 2021; Templin et al., 2008). However, many psychological constructs, and their corresponding scales, are multidimensional (e.g., Bianchini et al., 2008; Brooks et al., 2022; Vitoratou et al., 2021). When a scale is multidimensional but is scored as though it were unidimensional, the resulting estimates of the latent trait are less precise and less diagnostically accurate (Groen, 2014; Lai & Zhang, 2022; Reise et al., 2007; Soland et al., 2024; Vowels, 2023; Walker & Beretvas, 2003).

Multidimensionality can manifest at two levels: within-item and between-item (Figure 1; Wilson & Hoskens, 2005). Between-item multidimensionality occurs when different subsets of items from a given scale reflect distinct latent constructs, but each item loads onto one and only one factor (e.g., Model A). Within-item multidimensionality occurs when individual items themselves are indicators of more than one construct simultaneously, either through cross-loadings (e.g., Model B) or general-specific structures (e.g., Model C). Within-item multidimensionality can arise from genuine psychological complexity—a depression scale item asking about fatigue may load on both a depressive affect factor and a somatic symptom factor—or from measurement artifacts such as the wording effects introduced by negatively keyed items, which can produce a secondary response factor distinct from the construct of interest.

A particularly consequential form of within-item multidimensionality occurs when a dimension is associated with observable group membership. In this case, items may exhibit differential item functioning (DIF): item responses vary systematically across groups (e.g.,

racial, cultural, or gender groups) beyond what is explained by differences in the target construct. Both uniform DIF, in which item severity differs across groups, and non-uniform DIF, in which the strength of association between the item and the construct differs across groups, can decrease classification accuracy and introduce systematic error into group comparisons.

Despite growing methodological attention to dimensionality and measurement bias (Black et al., 2024; Fallon et al., 2021; Gonzalez & Pelham III, 2021; Lai & Tse, 2024; Luong & Flake, 2023; Navarro-González et al., 2024; Tay et al., 2022), the most common approach to classification in psychological research remains the application of a single fixed cut-point to an unweighted summed score. Such an approach invokes a set of psychometric assumptions that, when violated, bias assessment scores and thereby decrease the accuracy of the resulting classifications (e.g., De Ron et al., 2021; Haslbeck et al., 2022; Lundström et al., 2022; Ramsay et al., 2020; Wu et al., 2014). Specifically, use of an unweighted summed score assumes that all items are equally associated with the construct and that the latent trait is unidimensional. Stratifying individuals based on a single fixed cut-point assumes that this cut-point ignores measurement uncertainty in score estimates. These assumptions are pragmatically appealing but are rarely met by observed data. Violations of these assumptions can decrease diagnostic accuracy when constructs are multidimensional, when items differ in their severity or strength of association with the construct, or when a non-compensatory classification rule is more theoretically justified—that is, when high scores on one dimension should not be allowed to offset low scores on another.

Two alternative approaches to deriving classifications from diagnostic assessments are discussed in this paper. Item response theory (IRT) is a parametric family of models in which item responses are used to estimate a latent trait score, which is then compared to a threshold to

stratify individuals into diagnostic classes. IRT models also invoke assumptions, discussed in detail in Chapter 2, but these assumptions are less restrictive than those invoked by an unweighted summed score. One such assumption is that of item invariance, or the assumption that two individuals with the same level of a latent trait will respond the same way to a given item. When this assumption is not met, a scale is said to exhibit differential item functioning (DIF). Extensions of IRT, such as multi-group IRT and mixture IRT, have been developed to allow for more accurate estimation of scores when a scale contains DIF. However, if the grouping variable causing differential item functioning (DIF) is complex, unmeasured, or unknown, the use of such approaches is unreasonable. More discussion as to why can be found in Chapter 2. In such cases, DIF is left unmodeled, and estimated IRT scores are biased. This bias is then carried forward into IRT-based diagnostic classes. As such, a practically reasonable alternative is needed when the cause of DIF is complex, unknown, or unmeasured.

Prior work has found that RF may be a viable approach in such scenarios (Bain et al., 2026). However, this prior simulation invoked the simplest case of unmodeled DIF, when DIF is caused by a single, binary grouping variable. Additionally, only unidimensional scales were simulated. Whether RF classification remains a viable approach when DIF is caused by an interaction of multiple grouping variables or when scales are multidimensional has not been examined.

While this study investigates if RF offers a viable approach to classification problems, the author is not advocating for RF as a substitute for IRT or other traditional psychometric models in all cases. IRT remains essential to the field of psychology. IRT produces scores that are independent of the particular items administered and test information functions that quantify measurement precision across the full range of the construct rather than assuming a uniform

standard error for all respondents (Embretson, 1996; Hambleton et al., 2011). This has direct consequences for clinical practice: when measurement error varies systematically across trait levels, as it typically does for psychological measures with categorical response formats, using a single standard error can attenuate effect sizes, reduce statistical power, and impair diagnostic sensitivity and specificity (Embretson, 1996; Gilbert, 2025). IRT also allows for differential item functioning analysis to detect item-level bias across demographic groups (Battauz, 2019; Fallon et al., 2021; Paulsen et al., 2020; Wald, 1947), supports scale linking and equating that places scores from different instruments on a common metric, and provides the psychometric foundation for computerized adaptive testing, which achieves equivalent precision with substantially fewer items and reduces respondent burden (Gibbons et al., 2016; Huang et al., 2012; Van Groen et al., 2019). Beyond these applied advantages, IRT remains the gold standard approach to scoring in psychology, and the proposed RF approach is best understood as a complement to, rather than a replacement for, IRT.

Purpose of the Study

This dissertation aims to assess if RF is a viable approach to classification under conditions of between-item multidimensionality and complex DIF. Using a Monte Carlo simulation, this dissertation systematically manipulates key psychometric features, such as scale dimensionality, DIF prevalence, and DIF strength, to evaluate classification performance across three research questions. First, when a scale is multidimensional, should classification decisions be made using a general scale score or using the scores of each dimension separately? Second, how does the number of dimensions in a scale affect the performance of both IRT-based classification and RF classification approaches? Third, how does complex DIF affect the performance of both IRT-based classification and RF-based classification approaches?

These questions have both methodological and applied implications. They address ongoing concerns about construct complexity and measurement bias and fairness in assessment. By systematically evaluating how classification accuracy is affected by realistic violations of assumptions invoked by psychometric models, this study aims to provide proscriptive guidance for researchers who rely on diagnostic assessments to make classification decisions in psychological research.

This dissertation is organized into six parts. The first discusses the role that psychological scales play in research-based diagnostic assessment as well as measures that can be used to quantify classification performance of a scale. The second describes different approaches to classification with IRT scores. The third presents the use of RF for classification. The fourth outlines study goals and the Monte Carlo simulation design. The fifth presents the results from the Monte Carlo simulation. Finally, the sixth section discusses the implications, strengths, limitations, and future directions of this research.

Chapter 1: Diagnostic Assessment via Psychological Scales

Psychological scales are comprised of a series of items quantifying observable (e.g., behavioral or cognitive) indicators of an underlying psychological construct (e.g., depression, anxiety, or disordered eating). A scale score is estimated by modeling the relationship between the underlying latent trait and the observed indicators.[2] The latent trait is assumed to follow a continuous, normal distribution. However, the goal of diagnostic assessment is to partition these continuous trait values, or more accurately the obtained estimates of the trait values, into discrete classes (e.g., "depressed" or "not depressed"; Liu, 2012). The method used to stratify participants directly affects the accuracy of the resulting classifications and the validity, reliability, and

[2] For an overview of the history of latent variables and latent variable modeling, see Borsboom et al., (2003), Spearman (1904), or Thissen and Thissen-Roe (2022).

generalizability of research findings (Flake, 2021; Gonzalez, 2021; Jacobucci & Grimm, 2020; X. Liu, 2012; J. Wang & Tian, 2024). As such, accurate measurement and scoring are foundationally important in psychology.

This chapter examines the psychometric foundations of diagnostic classification using individuals' responses to psychological scales, beginning with approaches to scoring and latent trait estimation and progressing through the challenges of transforming continuous scores into categorical classifications. Two approaches to predicting diagnostic classes are then compared: a parametric, psychometric approach and a nonparametric approach. The chapter also addresses how the choice between compensatory and non-compensatory scoring frameworks affects classification accuracy and presents key measures of classification performance used to evaluate diagnostic decisions. Finally, the consequences of misclassification are discussed.

Psychological Scale Scoring and Latent Trait Estimation

As mentioned, psychological constructs are rarely directly measurable. As such, researchers rely on observed indicators, typically in the form of self-report items, to estimate a respondent's level of a latent trait. In this context, the term "latent trait" refers to the unobservable psychological construct of interest, and "trait estimation" refers to the process of estimating an individual's level of the construct from their observed item responses. Most items in psychological scales use Likert-type response formats, which provide ordinal data (Blanton & Jaccard, 2018; J. K. Lee et al., 2023).

The most widely used scoring method is the summed score, which sums observed item responses to estimate the latent trait (Ferrando & Lorenzo-Seva, 2021). This approach has the advantage of being simple to implement, easy to repeat across studies, not restricted by sample size, and able to be implemented in any statistical software program. However, use of a summed

score makes strong assumptions about the relationship between the scale items and the latent construct of interest (e.g., McNeish, 2024; McNeish & Wolf, 2020; Widaman & Revelle, 2023, 2024).

McNeish and Wolf (2020) showed that computing a summed score is mathematically equivalent to fitting an extremely constrained unidimensional confirmatory factor analytic model that makes two major assumptions. First, it assumes that all the loadings across items are equal (and oftentimes set to one). Second, it assumes that the residual variances are also equal across items. Put another way, in order for the use of summed scores to be justified, a scale must meet not only the assumption of tau-equivalence but also the more restrictive assumption of parallelism (Graham, 2006; Lord & Novick, 2008; Raykov, 1997b, 1997a). Additionally, the use of summed scores assumes that measurement invariance holds across samples (McNeish, 2022; Widaman & Revelle, 2024).

In practice, however, these assumptions are often violated. First and foremost, when a scale is multidimensional, treating all items as indicators of a single construct ignores important distinctions between symptom domains (Ayearst & Bagby, 2010; R. Liu, 2018; G. T. Smith et al., 2009; Stickle & Weems, 2006). When tau-equivalence is violated, summed scores give equal weight to items that may differ substantially in their relationship to the underlying trait, potentially over-weighting weak indicators and under-weighting strong ones (McNeish & Wolf, 2020). When the assumption of measurement invariance is violated, summed scores systematically over- or under-estimate trait levels for certain groups, introducing bias into both individual classifications and group comparisons (Widaman & Revelle, 2024). When any of these violations occur, summed scores produce biased and imprecise trait estimates (Borsboom, 2006; McNeish, 2022; McNeish & Wolf, 2020; Raykov & Zhang, 2025; Widaman & Revelle,

2023, 2024). This error can carry forward, resulting in misclassification of respondents. Despite these limitations, summed scores continue to be widely used (Sijtsma et al., 2024).

Recent advances in latent variable modeling have provided researchers with access to measurement models that require fewer constraints than the model implied by the summed score. Scores estimated via these more complex models are often referred to broadly as factor scores. Moreover, McNeish (2022) demonstrated that the psychometric properties—correlations with true scores, classification accuracy, and reliability estimates—of summed scores and factor scores differ even when their correlation is 0.98, such that factor scores have higher correlations with true scores, higher sensitivity, and higher reliability.

The majority of latent variable modeling approaches used in applied psychological research fall within one of two broad frameworks: item response theory (IRT) or structural equation modeling [see Borsboom and Cramer (2013) for an alternative network perspective that has recently grown in popularity]. This dissertation focuses on the IRT framework. Further technical details about IRT estimation and multidimensional models are provided in Chapter 2. For the present purposes, it is sufficient to note that trait estimation, regardless of latent variable modeling approach, results in a continuous output. To make a diagnostic decision, discrete classes must be predicted from this continuous estimate. The following section details possible approaches to predict these classes.

Classification Strategies

Two possible approaches to predicting these classes include (1) a multi-step psychometric approach or (2) a nonparametric classification approach (Giannouli & Kampakis, 2024; Gonzalez, 2021; Holden et al., 2011). In the psychometric approach, a researcher first uses a mathematical model to estimate an individual's level of the latent trait (e.g., the θ estimate from

an IRT model) and then determines a decision rule that can be applied to assign an individual into a predicted class based on their estimated score. For example, one decision rule is a simple cut-point such that individuals with latent trait levels above the cut-point are placed in one class (e.g., diagnosed with the disorder), while those with scores below the cut-point are placed in the other (e.g., the non-diagnosed class). The nonparametric approach does not estimate an individual's latent trait level. Rather, it predicts the probability of an individual's class membership from their item responses and then uses a set of decision rules applied to this probability to assign an individual into a predicted class based on this probability. This dissertation considers IRT as the parametric approach and RF as the nonparametric approach. The choice of decision rule corresponds to one of two general sets of scoring frameworks.

Compensatory vs. Non-Compensatory Scoring

Compensatory scoring rules allow for high scores in one dimension to offset lower scores in another. In contrast, non-compensatory scoring rules require scores across all assessed dimensions to exceed a set threshold (Templin & Henson, 2006). When a scale is unidimensional, these rules are the same, as there is only one assessed dimension.

When applied to diagnostic classification, compensatory scoring strategies assume that elevations on one symptom dimension can offset lower levels on another. To implement this approach, information first must be combined across dimensions, and then a single threshold is applied at the scale level to predict classification (Bovin et al., 2016; Byllesby & Palmieri, 2023; Meyers, 2018). However, compensatory scoring invokes the assumption that the dimensions are all interchangeable indicators of an overarching construct, such that higher severity in one domain contributes equivalently to classification as severity in another (Fried & Nesse, 2015; McNeish & Wolf, 2020). This logic is embedded in many multidimensional scales that are

recommended to be scored with a unidimensional model for diagnostic purposes (Baker et al., 2019; Bovin et al., 2016; Hailu Gebrie, 2018; Raymond & Butler, 2025; Rosenthal et al., 2021).

For example, Raymond and Butler's (2025) MisoQuest scale treats misophonia as a unidimensional construct, recommending scoring the scale by taking a total score and thresholding such that scores greater than or equal to 61 indicate the presence of clinically significant misophonia. However, much of the previous research on misophonia indicates that it is a complex disorder comprised of many unique dimensions (Bain et al., 2025; Rosenthal et al., 2021; Schröder et al., 2013). Thus, by using a single score, MisoQuest reduces items measuring multiple theoretical domains of misophonia (e.g., impairment, affective and cognitive reactions, and coping and avoidance behaviors) into a latent dimension of misophonia severity (Silva et al., 2024). Although this approach is supported by psychometric evidence, it assumes compensatory interchangeability among symptoms (e.g., anger compensating for avoidance; Raymond & Butler, 2025; Silva et al., 2024).

In contrast, non-compensatory scoring strategies invoke the assumption that each dimension is unique, and one cannot be exchanged for the other. Put another way, severity levels need to be elevated across all dimensions for a diagnosis to be assigned, and no amount of severity in one dimension can compensate for the absence of severity in another. Mathematically, this means that thresholds are conjunctive rather than additive, and multiple cut-points are needed to make a classification decision (Kroc & Olvera Astivia, 2022). The proposed diagnostic criteria for misophonia exemplify this structure, requiring elevated levels of symptomology in all six different domains (Table 1; Schröder et al., 2013). Low symptom severity in one of the six domains precludes a diagnosis, regardless of symptom severity elsewhere.

Regardless of whether a compensatory or non-compensatory scoring rule is applied, researchers need to assess the clinical utility of a given scale by quantifying how well scale scores predict gold standard diagnostic classes. The following section provides a brief overview of different metrics that can be used to assess classification performance.

Measures of Classification Performance

A confusion matrix can be used to describe the relationship between an individual's gold standard diagnostic status and their predicted classification. An illustrative example is shown in Table 1.

The confusion matrix summarizes classification decisions by cross-tabulating observed, gold standard diagnostic status against predicted classification. Four outcomes are possible: true positives (TP; correctly predicted to be diagnosed), false negatives (FN; incorrectly predicted to be not diagnosed), false positives (FP; incorrectly predicted to be diagnosed), and true negatives (TN; correctly predicted to be not diagnosed). All illustrative confusion matrix can be found in Table 2.

Below are key statistics derived from the confusion table that are commonly used to assess classification performance. Accuracy measures the proportion of correctly classified instances (both true positives and true negatives) out of the total instances. It is computed as:

$$Accuracy = \frac{\mathrm{TP+TN}}{\mathrm{TP+FN+FP+TN}}. \quad (1)$$

Sensitivity quantifies the models' ability to correctly identify positive instances. It is calculated as:

$$Sensitivity = \frac{\mathrm{TP}}{\mathrm{TP+FN}}. \quad (2)$$

Specificity measures the models' ability to correctly identify negative instances. It is computed in the following way.

$$Specificity = \frac{\mathrm{TN}}{\mathrm{TN+FP}}. \quad (3)$$

Precision assesses the proportion of true positive predictions among all positive predictions. It is given by:

$$Precision = \frac{TP}{\mathrm{TP + FP}}. \quad (4)$$

Negative predictive value (NPV) quantifies the proportion of true negative predictions among all negative predictions. It is calculated as:

$$NPV = \frac{\mathrm{TN}}{\mathrm{TN + FN}}. \quad (5)$$

All statistics discussed so far are influenced by the relative sizes of the diagnostic groups. Therefore, a more balanced measure, the F1 score (F1), is often reported. The F1 is the harmonic mean of precision and sensitivity:

$$F1\ score = 2 \times \frac{Precision \times Sensitivity}{Precision + Sensitivity}. \quad (6)$$

Determining the location of the diagnostic cut-point(s) is vital as it affects not only overall classification accuracy but also the relative accuracy (and error) for either diagnostic class. For example, it is important to maximize the proportion of individuals correctly identified as meeting diagnostic criteria (sensitivity) as well as the proportion of individuals correctly identified as not meeting diagnostic criteria (specificity). If the cut-point is set too high, sensitivity will decrease, resulting in more FNs. If the cut-point is set too low, specificity will decrease, resulting in more FPs.

In a model with no classification error, the above statistics would all be equal to one. There are several approaches available in the literature to optimize classification accuracy (Gibbons et al., 2016; X. Liu, 2012; Youngstrom, 2013; Zweig & Campbell, 1993). One widely used method is receiver operating characteristic (ROC) curve analysis (Youngstrom, 2013). ROC

curve analysis allows researchers to visualize and select a cut-point that balances sensitivity and specificity appropriately for their specific research.

ROC curves are constructed by calculating sensitivity and specificity across a range of possible thresholds. These values are then plotted, with the curve showing how classification performance changes as a function of the threshold (Figure 2). For example, to balance sensitivity and specificity, a researcher may select the threshold that maximizes the Youden index (Liu, 2012). The Youden index is the point in the ROC curve that is farthest away from the chance diagonal. In practice, researchers may not value sensitivity and specificity equally and may wish to choose the threshold that prioritizes one over the other. For example, in high-risk scenarios (e.g., suicide risk), maximizing sensitivity may be prioritized at the expense of specificity. Once the threshold is selected, it can be applied to predict the diagnostic classes from the continuous score.

One is also able to assess how well a scale discriminates between classes via the shape and height of the ROC curve. Often, a reference line is included in the plot to illustrate the chance diagonal (a 45-degree line from the bottom left of the plot to the top right; Figure 2). The area under the ROC curve (AUC) can also be helpful in comparing the performance of classifiers. The AUC provides a single summary statistic of classification performance across all possible thresholds. Mathematically, the AUC can be expressed as:

$$AUC = \int_0^1 TPR\left(FPR^{-1}(x)\right)dx, \tag{7}$$

where TPR is the true positive rate (sensitivity) and FPR is the false positive rate (1 – specificity). The AUC is equal to the probability that a randomly selected diagnosed individual has a higher predicted score than a randomly selected non-diagnosed individual (Hanley & McNeil, 1982). It can also be interpreted as the average sensitivity across all values of

specificity, or the reverse, the average specificity across all values of sensitivity. A curve that follows the chance diagonal will have an AUC of .50, while curves that illustrate stronger performance will have AUC values closer to 1.0.

Consequences of Misclassification in Research

Misclassification has both interpretive and theoretical consequences (Bradford et al., 2024; Staccio, 2025; Wakefield, 2007, 2010). False positives may result in overpathologizing normative behavior, contributing to inflated prevalence estimates and potentially stigmatizing individuals without clinical symptoms (Wakefield, 2010, 2015, 2016). Conversely, false negatives reduce statistical power, underestimate associations between constructs, and omit clinically relevant cases from analyses, potentially leading to underestimation of group differences or failure to detect meaningful effects (Blakely et al., 2013; Brennan et al., 2021; Chyou, 2007; Egleston et al., 2011; Höfler, 2005; Soland et al., 2024; Tiego et al., 2023; Valeri & Vanderweele, 2014; Vowels, 2023; Weiss & Dardick, 2021; Yland et al., 2022; Zheng & Tian, 2005).

When misclassification is systematic (i.e., more likely in one group than another), bias can occur in either direction, potentially overestimating, underestimating, or even incorrectly estimating the direction of observed associations (Chyou, 2007; Gallop & Weschle, 2019; Höfler, 2005; Imai & Yamamoto, 2010, 2010). This creates challenges for group comparison studies, where unequal classification accuracy across groups can produce spurious findings or mask true differences. One possible explanation for systematic misclassification is differential item functioning (DIF). If items function differently across groups, but the same scoring model is used for both groups, trait estimates become biased, and classification thresholds that are appropriate for one group may be inappropriately stringent or lenient for another (Gonzalez &

Georgeson, 2020; Gonzalez & Pelham III, 2021). This can result in systematic over-diagnosis in some groups and under-diagnosis in others, perpetuating health disparities and contributing to inequitable treatment access.

The impact of DIF on classification accuracy is a function of both the magnitude of item bias and the number of biased items in the scale (Ozcan & Lai, 2025; Paulsen et al., 2020; Rome & Zhang, 2018). Even small amounts of DIF across multiple items can accumulate to produce substantial differences in classification rates between groups (Rome & Zhang, 2018). For example, if a depression scale contains items that are more likely to be endorsed by one cultural group at equivalent levels of depression, individuals from that group may be disproportionately classified as depressed, while individuals from other groups may be systematically missed. Such patterns can reinforce stereotypes, contribute to over-pathologization of certain populations, and undermine trust in psychological assessment practices (Droubay et al., 2025; Jennings et al., 2022; Mongelli et al., 2020; Schwartz & Blankenship, 2014; Teplin et al., 2023; Wakefield, 2010, 2015, 2016). Ensuring measurement fairness is therefore not only a psychometric necessity but also a matter of equity regarding ethical assessment practice.

Another possible explanation for misclassification is improper selection of a scoring rule. When non-compensatory disorders are measured with compensatory scoring systems, the scoring model is misspecified, and the risk of misclassification increases (R. Liu, 2018; Tamano et al., 2025; Vowels, 2023). Individuals may be labeled as diagnosed based solely on extreme scores in a single dimension despite low or moderate scores on other dimensions, producing false positives. On the other hand, individuals with moderate levels across all domains may fail to exceed the total score threshold, producing false negatives (Tamano et al., 2025; Tan et al., 2022). If a unidimensional total score is used, it may fail to exceed the recommended diagnostic

threshold, even though the pattern meets all diagnostic criteria (Avila et al., 2015; Karukivi et al., 2017; Linde et al., 2022; Ravand & Baghaei, 2019; D. Wang et al., 2019).

Given the substantial consequences of misclassification, researchers must carefully consider cut-point selection and classification performance during both scale development and selection (Gonzalez, 2021; Kyriazos & Stalikas, 2018; Stefana et al., 2025). Cut-point selection should be justified based on the specific goals and consequences of the research context (H. E. Jones et al., 2019; Monaghan et al., 2021; Sheldrick et al., 2015). ROC analysis and cross-validation techniques should be employed to evaluate and optimize classification performance (LeDell et al., 2015; Song et al., 2021; Sweet et al., 2023; Youngstrom, 2013; Zweig & Campbell, 1993).Additionally, researchers should report a set of classification performance metrics (e.g., accuracy, sensitivity, specificity, negative predictive value, etc.) rather than accuracy alone. These metrics provide a more complete picture of how a scale performs across different classification thresholds and population base rates.

Classification accuracy depends not only on the choice of scoring rule or decision cut-point, but also on the quality of the underlying trait estimates from which the classifications are derived. Traditional summed score methods, while easy to implement, often fail to account for key psychometric complexities such as between-item variation in discrimination and severity and the multidimensional structure of many constructs. Advanced variable models, particularly IRT-based scoring, address many of these limitations. The following chapter discusses IRT-based approaches to diagnostic classification, with a focus on the strengths of IRT scoring models over other advanced latent variable modeling approaches to scoring.

Chapter 2: IRT-Based Diagnostic Classification

Many papers on the development or evaluation of a given psychometric scale will provide a recommended cut-point for diagnostic classification that is based on a summed score (e.g., Carvalho et al., 2019; Jacobsen et al., 2024; Ohi et al., 2024; Williams et al., 2022; Wortmann et al., 2016). As such, diagnostic classifications are often derived using a summed score-based approach. However, it is well documented that item response theory (IRT) models have distinct scoring advantages over summed scores. As such, IRT-based approaches to diagnostic classification could be used instead. The overall process for IRT-based classification is outlined in Figure 3.

According to Edwards (2009), IRT is a family of latent variable models that seek to uncover the underlying process that influences responses to observed variables. This relationship is driven by properties of individuals (i.e., θ, which corresponds to the level of the latent variable or construct of interest) and the properties of the items. IRT models are used for the development and validation of many existing diagnostic scales (Bain et al., 2025; Bradshaw & Templin, 2014; Dawson et al., 2010; Edelen & Reeve, 2007; Lo et al., 2017; Nikan et al., 2018; Rosenthal et al., 2021; Thissen & Wainer, 2001). There are many benefits of working within the IRT framework for scale development and evaluation. A detailed discussion can be found in Edwards (2009), Embretson and Reise (2000), and Thissen and Steinberg (2009). I focus here on the specific benefits that are related to scoring.

Advantages of IRT Scoring

The research highlighting the general advantages of score estimation via advanced latent variable models over estimation via a simple summed score was discussed in Chapter 1. Therefore, this section is focused on the advantages of IRT-based scoring over factor analytic

model-based scoring.[3] As noted by Thissen and Thissen-Roe (2022), although both IRT and factor analysis are members of the larger family of advanced latent variable modeling strategies, they have largely separate histories. IRT models were mostly developed to estimate test scores from observed discrete data. Factor analysis models were developed to estimate factor loadings, residual variances, and possibly factor means from observed continuous data. Factor analytic scoring procedures are based on regression (Bartlett, 1937; Thomson, 1935, 1936, 1938; Thurstone, 1935, 1947), and the literature on factor score estimates is often about the specific properties of the estimates, such as whether they are conditionally unbiased or whether they have the right variances and correlations with other variables, rather than the estimates themselves (e.g., Devlieger et al., 2016; Hoshino & Bentler, 2013, 2013). To borrow the words of Thissen and Thissen-Roe, "in the factor analytic framework, factor score estimates have been a stepchild, often unwanted." (Thissen & Thissen-Roe, 2022, p. 1). Here, three advantages of IRT scores are discussed: (1) conditional measurement precision, (2) sample (in)dependence of item and person parameters, and (3) modeling the non-linear item-trait relationship. A fourth advantage of IRT is formal frameworks for differential item functioning (DIF), but discussion of this advantage is left until the section on DIF later in this chapter.

Perhaps the most fundamental advantage of IRT scores is that measurement error is explicitly modeled as a function of the latent trait level. Under the factor analytic framework, measurement error is assumed to be completely random and uncorrelated with everything, including the level of the latent construct (Rhemtulla & Savalei, 2025). IRT models do not invoke this assumption. Rather, IRT quantifies the precision of the scale at every point along the

[3] Note that although factor analysis can be done in the IRT framework, the term factor analysis or factor analytic methods is used here to refer to the measurement model aspect of structural equation modeling. The term "factor scores" is here after used to refer to scores estimated via the factor analytic framework.

latent trait continuum through test information functions (TIFs; discussed in further detail later in this chapter). The inverse of a TIF is a function of conditional standard error across the latent trait continuum. This means a researcher using IRT scores knows not just that their measure has some test-level global reliability of 0.85, for example, but that it is less precise for individuals with lower trait levels. This information is not available with traditional factor scores.

Factor analytic item statistics (e.g., *p*-values and point-biserial correlations between the item and latent trait) are sample dependent, meaning if a researcher administered the same test to a group with high average trait levels, items appear less meaningful than if administered to a group with low average trait levels. This creates problems for cross-study comparisons. IRT parameters, assuming correct model fit, are theoretically sample-invariant (up to a linear transformation of the scale), and person parameters are item-invariant (Edwards, 2009).

Factor scores assume a linear relationship between the latent construct and the observed item responses. As argued by Embretson (1996), this assumption is almost certainly violated for binary and ordinal items, particularly in the tail ends of the distribution of the latent trait. IRT does not make this assumption; it models the relationship between the estimate of the latent trait and responses to ordinal items with a logistic curve, explicitly capturing the nonlinearity. Two of the most common IRT models used in the psychological sciences are Samejima's graded response model (GRM; 1997) and its multidimensional extension (MGRM; see De Ayala, 1994).

Graded Response Model (GRM)

The GRM is commonly used in the development of diagnostic scales based on Likert-type item responses. It can be expressed as:

$$P(x_j = c \mid \theta) = \frac{1}{1 + \exp\left[-a_j(\theta - b_{cj})\right]} - \frac{1}{1 + \exp\left[a_j(\theta - b_{(c+1)j})\right]} \qquad (8)$$

which represents the probability of endorsing response option *c* given the latent trait (θ), resulting in the observed response (x_j) to item *j* (Samejima, 1997). The GRM uses differences in cumulative probability to determine the probability of answering with a specific response option on each item. Specifically, the GRM compares the probability of responding with category *c* and the probability of responding with successive *c* + 1 categories. For example, using a typical 5-point Likert scale (i.e., *C* = 5), there are five possible responses: {0, 1, 2, 3, 4}. As such, the GRM will compute the comparisons: {0} vs. {1,2,3}; {0,1} vs. {2,3}; {0,1,2} vs. {3,4}; and {0,1,2,3} vs. {4}. For an item with *C* categories, this results in *C* – 1 *b*-parameters, each representing the point on the latent continuum where the probability of selecting a higher category surpasses that of the current one.

The *a*-parameter, or discrimination parameter, represents the slope for item *j* and reflects how well an item distinguishes among individuals with different levels of the latent trait. A steeper slope suggests that a larger portion of the variability in item responses is explained by differences in the latent construct (i.e., a stronger relationship between the item and latent construct). The *b*-parameter represents the threshold for response option *c* for item *j*. In the context of clinical diagnostic assessment, *b*-parameters are often referred to as severity parameters, as they indicate the level of severity of the disorder of interest needed for a respondent to select a specific response option, with higher values corresponding to higher severity levels required to endorse an option (Samejima, 1969).

The GRM assumes unidimensionality, meaning that a single latent trait explains the relationship among items; essentially, that each item reflects one underlying diagnostic construct. When scales are comprised of multiple latent constructs, the GRM must be extended to a multidimensional model.

Multidimensional Graded Response Model (MGRM)

When items reflect more than one underlying trait (e.g., multiple diagnostic dimensions), the GRM must be extended to the MGRM. In the MGRM, the probability that person *i* selects category *c* or higher on item *j*, conditional on their latent traits $\boldsymbol{\theta}_i$ is given by:

$$P(Y_{ij} \geq c \,|\boldsymbol{\theta_i}) = \frac{1}{1 + exp[-\boldsymbol{a}_j(\theta_{d(j)i} - b_{jc})]} \quad (9)$$

where a_j is the item's discrimination parameter vector, $\theta_{d(j)}$ is the latent trait dimension associated with item *j*, and b_{jk} are ordered category thresholds. For item *j*, the number of discrimination parameters relates directly to the dimensionality (*D*) of the latent space, meaning that a_j is a *D* length vector. As *D* increases, the number of necessary discrimination parameters increases accordingly, while severity parameters depend on the number of response categories. Thus, scale length, dimensionality, and item response format all impact the number of parameters that must be estimated for a given model.

Challenges in item parameter interpretation aside, the MGRM extends the unidimensional GRM and invokes the same assumptions (Reckase, 2009). However, parameter estimation becomes more computationally intensive in higher dimensions, resulting from the increased number of integrals required (Bock & Aitkin, 1981; Schilling & Bock, 2005).

In the context of classification, multidimensional IRT (MIRT) models raise a key challenge: individuals are no longer represented by a single θ score but by a vector of scores (one per latent dimension). While this provides a more nuanced profile of symptoms and traits, it complicates the task of classification. Therefore, one must consider how the individual dimension scores can be combined or whether they should be combined at all. The MGRM allows for non-compensatory classification by setting a separate threshold for each dimension. To implement a compensatory classification strategy, however, use of a MIRT model is required

(Stucky & Edelen, 2014), and a cut-point would be identified for the second-level factor in the same manner as discussed previously (e.g., through use of an ROC curve). This study focuses on the use of a second-order MGRM as the multi-level MIRT model.

Second-Order Multidimensional Graded Response Models

In a second-order model, specific factors $\theta_1, \ldots, \theta_D$ are themselves associated with a higher-order general factor θ_G:

$$\theta_d = \lambda_d \theta_G + \zeta_d \tag{10}$$

where λ_d is the loading of specific factor *d* on the general factor, and $\zeta_d \sim N(0, \psi_d)$ represents the residual variance in the specific factor not explained by the general factor (Cai, 2010a). Items load on specific factors, which in turn load on the general factor, creating a two-level structure (see Figure 1C). The general factor only influences item responses indirectly through the specific factors.

The second-order MGRM extends the MGRM framework by imposing this second-order factor structure on the latent trait space. For an item *j* that loads on specific factor *d*, substituting Equation 10 into Equation 9 yields the probability that person *i* selects category *c* or higher on item *j*:

$$P(Y_{ij} \geq c \mid \theta_{Gi}, \zeta_{id}) = \frac{1}{1+\exp[-a_j(\lambda_d \theta_{Gi} + \zeta_{id} - b_{jc}]}, \tag{11}$$

where a_j is the discrimination of item *j* on its specific factor *d*, $\lambda_d \theta_{Gi}$ is the indirect contribution of the general factor mediated through specific factor *d*, and ζ_{id} is the person- and factor-specific residual. The two-level structure implies that the correlation between any two specific factors, θ_d and $\theta_{d'}$, is determined entirely by their mutual dependence on θ_G:

$$Cor(\theta_d, \theta_{d'}) = \frac{\lambda_d \lambda_{d'}}{\sqrt{(\lambda_d^2 + \psi_d)(\lambda_{d'}^2 + \psi_{d'})}}, \tag{12}$$

where $\psi_d = Var(\zeta_d)$. After conditioning on θ_G, the specific factors are mutually independent, since their residuals, ζ_d, are assumed uncorrelated—a property known as local independence at the second-order level (Cai, 2010a; Holzinger & Swineford, 1937; Stucky & Edelen, 2014).

An important practical consideration is that standard software implementations, including the *mirt* package in R, cannot directly estimate second-order GRMs (Chalmers, 2012). Instead, second-order models must be specified as constrained bifactor models, where the constraints enforce the mathematical equivalence between the two parameterizations. In the bifactor representation, each item *j* loads simultaneously on both the general factor and its specific factor *d*, with loadings a_{jG} and a_{jG}, respectively, so that the item response function takes the form:

$$\mathrm{P}\left(\mathrm{Y}_{ij} \geq \mathrm{c} \mid \theta_{Gi}, \xi_{id}\right) = \frac{1}{1 + \exp\left[-a_{jG}\theta_{Gi} + a_{jd}\xi_{id} - b_{jc}\right]}, \tag{13}$$

where ξ_{id} is the person's score on specific factor *d*, which in the bifactor model is orthogonal to θ_G by construction. The equivalence between Equations 11 and 13 is established by the $a_{iG} = \lambda_d a_{jd}$ for all items *j* loading on specific factor *d*. Under this constraint, the effective loading on the general factor is proportional to the item's loading on its specific factor, scaled by λ_d, which reproduces the indirect path structure of Equation 10. Additionally, the total variance attributable to the specific factor *d* is correctly partitioned between the general and residual components as defined in Equation 10 (Rijmen, 2010).

Information Functions

In addition to generating scale scores, IRT provides tools for evaluating item- and test-level measurement precision. IRT conditions precision across levels of the latent trait. IRT characterizes precision using several metrics, including Fisher information (hereafter called information). Information is provided at the item level through an item information function (IIF), often referred to as an item information curve (IIC) when plotted. IICs illustrate where

across a range of θ values each item is best able to discriminate, or put another way, at which levels of θ a given item is most precise. For example, the IIC in Figure 4 illustrates that the item is most informative for θ = [0.5, 1.5]. IIFs are additive, so a test information function (TIF) can be computed by summing all IIFs. TIFs can be plotted to illustrate where on the θ continuum the scale is most precise. For example, the TIF in Figure 4 illustrates that the scale is most informative in the θ = [0, 1] range, indicating θ estimates within this range are most precise. IIFs and TIFs can be used in scale development to select items that ensure the scale has specific measurement properties. For example, researchers creating a diagnostic scale may wish to maximize the number of items in the scale that are most precise within a specific range of θ containing the diagnostic θ threshold.

Model Estimation and Scoring

Estimation of MIRT models presents significant computational challenges, as both item parameters and individual latent trait levels must be estimated simultaneously. Estimation complexity increases substantially with the number of dimensions, as the likelihood function requires numerical integration over the multidimensional latent trait space. This section describes the algorithms used for parameter and trait estimation in this study. Particular attention is paid to the relevance of dimensionality and model structure to the choice of optimal estimation and scoring algorithms.

Expectation-Maximization Algorithm

The expectation-maximization (EM) algorithm is the most widely used approach for estimating item parameters in IRT models (Bock & Aitkin, 1981). In general, the EM algorithm alternates between the following two steps from a set of initial parameter estimates, say $\gamma^{(0)}$, and it generates a sequence of parameter estimates, $\gamma^{(0)}, \ldots, \gamma^{(k)}$, where $\gamma^{(k)} = (\alpha_1^{(k)}, \ldots, \alpha_1^{(k)}, \beta_1^{(k)}, \ldots,$

$\beta_1^{(k)}$), that converges under some very general conditions to the maximum likelihood estimate (MLE) of γ as the number of cycles k tends to infinity (Cai & Thissen, 2014):

E-step. Given $\gamma^{(k)}$, evaluate the conditional expected complete data log-likelihood, $Q(\gamma \mid Y; \gamma^{(k)})$, which is taken to be a function of γ.

M-step. Maximize $Q(\gamma \mid Y; \gamma^{(k)})$ to yield updated parameter estimates $\gamma^{(k+1)}$. Go back to E-step and repeat. The cycles are terminated when the estimates from adjacent cycles stabilize.

For unidimensional and low-dimensional models, the EM algorithm implements numerical integration using adaptive quadrature (Chalmers, 2012). Adaptive quadrature approximates the integral over the latent trait distribution by evaluating the integrand at carefully selected quadrature points, with more points placed in regions where the integrand changes rapidly. For models with few dimensions, this approach provides accurate integration with reasonable computational cost. However, even with a moderate number of quadrature points, the exponentially increasing size of the grid as the number of dimensions increases presents major computational challenges. Adaptive quadrature helps somewhat by requiring fewer points than fixed quadrature, but it does not solve the problem completely. Various authors have referred to this as the "curse of dimensionality" (Y. Liu et al., 2018, p. 456).

The EM algorithm is particularly well-suited for second-order and bifactor models, even when these models have a relatively high number of dimensions, because the constrained structure of these models reduces the effective dimensionality of the integration problem (Rijmen, 2010). The orthogonality constraints in bifactor models create conditional independence structures that allow for more efficient computation. For these reasons, EM

remains the preferred estimation method for second-order multidimensional graded response models and other constrained multidimensional structures (Rijmen, 2010).

Metropolis-Hastings Robbins-Monro Algorithm

For high-dimensional models without the simplified structure of the second-order or bifactor parameterizations, the Metropolis-Hastings Robbins-Monro (MH-RM) algorithm provides a more computationally efficient alternative to EM (Cai, 2010c, 2010b). The MH-RM algorithm avoids the computational challenges of the EM algorithm by replacing numerical integration with stochastic sampling combined with stochastic approximation. MH-RM is considered a state-of-the-art estimation technique for high-dimensional IRT models and is particularly effective for correlated factor models with three or more factors (Chen et al., 2019; K. T. Han & Paek, 2014).

The MH-RM algorithm consists of two steps (Cai & Thissen, 2014)[4]. In the first phase (burn-in), MH-RM uses Metropolis-Hastings sampling to draw samples from the posterior distribution of latent traits conditional on current item parameter estimates. These samples replace the numerical integration required in the E-step of EM, avoiding the exponential growth in quadrature points. In the second phase (stochastic approximation), the algorithm uses a Robbins-Monro procedure to update item parameters based on the sampled sufficient statistics. The Robbins-Monro procedure employs a decreasing sequence of step sizes, which allows the algorithm to converge to maximum likelihood estimates while maintaining computational efficiency (Robbins & Monro, 1951).

[4] Note that a pre-processing imputation step is required if the data contains missing data. As missing data is not within the scope of this research, that step is not discussed here. For more on the imputation step and technical details regarding parameter estimation, Metropolis-Hastings sampling, and Robbins-Monro stochastic approximation filtering, see Cai and Thissen (2014).

The computational cost of the MH-RM algorithm increases approximately linearly with dimensions, compared to the exponential increase for quadrature-based methods. This efficiency comes at the cost of introducing Monte Carlo error, but with sufficient burn-in iterations and sample size, the MH-RM algorithm yields maximum likelihood estimates of the parameters for any number of dimensions without a steep increase in computational time as a function of the number of dimensions (Cai, 2010c). As such, MH-RM has become the default estimation algorithm for high-dimensional correlated factor models in many statistical software programs (Chalmers, 2012).

Latent Trait Estimation Methods

Once item parameters have been estimated using EM or MH-RM, individual latent trait scores must be estimated. Latent trait scores can be estimated by treating the model parameters as if they were known, as is reasonable if model parameters have been accurately estimated (Brown & Croudace, 2015). When the model parameters are fixed to their estimated values, the probability of every item response is assumed to be dependent only on the latent traits, and the fundamental approach to estimating trait scores is to search for values that maximize the joint likelihood of each response pattern $\boldsymbol{u} = (u_1, u_2, \dots, u_m)$.

One approach to estimation is the maximum likelihood approach. The maximum likelihood approach assumes local independence, that is, the item responses are independent after conditioning on the latent trait. The maximum likelihood scores are found iteratively by maximizing the mode of the likelihood function (Brown & Croudace, 2015):

$$l(u|\theta) = \prod_{i=1}^{m} P_i(\theta)^{u_i}\big(1 - P_i(\theta)\big)^{1-u_i} \tag{14}$$

This score, however, is undefined for some response patterns, notably for "perfect" patterns (for instance, by answering "Strongly Agree" to all items). The likelihood of such a pattern increases

infinitely as the latent trait score increases (Brown & Croudace, 2015). The score is also undefined when non-keyed responses are given to all items (for instance, by answering "Strongly Disagree" on all items). Such response patterns can be quite common in screening tests.

An alternative approach is to use prior information about the trait distribution in addition to the actual item responses. With this Bayesian approach, the likelihood of a score given an observed response pattern (i.e., posterior distribution) is computed from both the known distribution of the score in the population (i.e., the prior distribution, usually assumed to be the standard normal distribution) and the likelihood of the observed response pattern given the score. Two alternative methods using this approach are expected a posteriori (EAP) estimation and maximum a posteriori (MAP) estimation. EAP estimation computes the mean of the posterior distribution, while MAP estimation maximizes the mode of the posterior distribution (Embretson & Reise, 2000). By adding information from the prior distribution, the problem of undefined trait scores for "perfect" patterns is overcome, and a score is always defined. EAP estimation has been shown to produce more accurate estimates than MAP estimations (LaHuis et al., 2025). Therefore, only EAP estimation is discussed in detail here.

EAP estimation is most often used when only one trait θ is measured (Bock & Mislevy, 1982). The EAP score is computed as the ratio of the integral of the posterior function weighted by the latent trait θ over the unweighted integral of the posterior function (Thissen & Wainer, 2001):

$$EAP[\theta] = \frac{\int \theta p(\theta|\boldsymbol{u}) d\theta}{\int p(\theta|\boldsymbol{u}\} d\theta}. \quad (15)$$

The posterior distribution is proportional to the likelihood of the response patterns times the prior distribution of θ, which is typically assumed to be multivariate normal with a mean vector of zero and estimated covariance matrix Σ.

The posterior EAP approach has been shown to achieve precise estimates with fewer items than the MLE method; however, they are known to shrink latent trait distributions toward the population mean. The amount of shrinkage depends on several factors, including the test length, and can be quite substantial in short tests (Brown & Croudace, 2015; Thissen & Wainer, 2001). This shrinkage reduces variance in estimated trait scores compared to true trait scores, and shrinkage is larger for individuals with more unusual response patterns or lower information (i.e., individuals whose trait levels are poorly measured by the items; C. Wang, 2015). For well-measured individuals with typical response patterns, EAP estimates closely approximate MLE and MAP estimates (C. Wang, 2015).

EAP Estimation via Quasi-Monte Carlo Integration. EAP estimation requires numerical integration over the posterior distribution of latent traits. For unidimensional and low-dimensional models, standard adaptive quadrature provides accurate integration with manageable computation cost. However, for higher-dimensional models, quasi-Monte Carlo (QMC) integration provides a more efficient alternative (Jank, 2005).

QMC methods approximate the integral by evaluating the integrand at a carefully selected set of points that are more evenly distributed across the integration domain than random Monte Carlo samples. These quasi-random sequences, such as Sobol or Halton sequences, achieve better coverage of the integration space than simple random sampling, leading to more accurate approximation with fewer evaluation points (Jank, 2005). The convergence rate of QMC integration is $O(1/n)$ compared to $O(1/\sqrt{n})$ for standard Monte Carlo integration, where n is the number of evaluation points, making QMC substantially more efficient while maintaining the same level of accuracy (Jank, 2005).

In the context of MIRT, QMC integration is particularly beneficial for computing EAP estimates from bifactor and high-dimensional correlated factor models. For these models, the posterior distribution may have a complex multimodal structure, and QMC's superior coverage properties ensure accurate representation of the posterior across its support. The *mirt* package implements QMC integration using Sobol sequences, with the number of quadrature points (typically 5,000 to 10,000) determining the accuracy-efficiency tradeoff (Chalmers, 2012).

QMC integration is recommended for bifactor models regardless of dimensionality because the effective dimensionality of the posterior distribution in bifactor models is equal to the number of specific factors plus one for the general factor. For correlated factor models with three or more dimensions, QMC similarly provides more accurate and efficient integration than adaptive quadrature (Jank, 2005). For unidimensional and two-dimensional correlated factor models, standard adaptive quadrature is typically sufficient.

The choice of integration method interacts with the choice of estimation algorithm. When EM is used for parameter estimation with adaptive quadrature, it is natural to use the same quadrature scheme for EAP estimation. However, when MH-RM is used for parameter estimation, QMC integration for EAP estimation provides better consistency because both approaches avoid the exponential growth in computational cost associated with traditional quadrature. For this reason, the combination of MH-RM estimation with QMC integration for EAP scoring has become standard practice for high-dimensional IRT models (Chalmers, 2012).

Differential Item Functioning (DIF)

Unbiased classification depends heavily on the assumption that psychological assessments measure constructs equivalently across different individuals and groups. When this assumption is not met, bias can emerge, leading to misclassification, unequal access to services,

and compromised scientific validity. One of the most well-known forms of such measurement bias is DIF.

An item is said to exhibit DIF if two respondents from distinct subgroups (e.g., gender, language, etc.) have equal levels of the trait being measured (θ) but do not have the same probability of endorsing each response category for that item (Edwards & Edelen, 2009). DIF indicates potential bias in items and threatens the validity of inferences drawn from test scores (Camilli & Shepard, 1994; Zumbo, 1999). There are two main types of DIF: uniform DIF and non-uniform DIF (Figure 5A and 5B, respectively). Uniform DIF occurs when the item consistently favors one group over another across all levels of the latent trait (e.g., a screening item that always appears more severe for men than women). This is reflected by a difference in difficulty parameters, but not in discrimination parameters ($a_{focal} = a_{ref}$ and $b_{focal} \neq b_{ref}$). Non-uniform DIF occurs when the discrimination parameters differ between groups, meaning that the strength or direction of the group difference changes across the trait continuum (i.e., $a_{focal} \neq a_{ref}$). Put simply, the direction of the bias depends on the level of the latent trait, such that an item might favor one group at low trait levels but the other group at higher levels (Figure 4). This can introduce complex bias into trait estimations and diagnostic decision-making. DIF is distinct from general item difficulty; rather, it refers specifically to construct-irrelevant variance associated with group membership (Penfield & Camilli, 2006). In this way, DIF is directly tied to issues of fairness and equity in assessment. If an item systematically favors one demographic group over another, even when individuals are otherwise equivalent in their symptom severity or functioning, this constitutes item bias and undermines the legitimacy of diagnostic conclusions.

Examples of DIF are plentiful in clinical, educational, and cross-cultural contexts. For instance, depression screening tools developed in Western contexts may show DIF when used

with non-Western populations due to cultural differences in symptom expression (Haroz et al., 2016; Kalibatseva et al., 2014; Uebelacker et al., 2009). In developmental settings, items related to language or social engagement may exhibit DIF between neurotypical and neurodiverse children (Ross et al., 2023; Schiltz & Magnus, 2021). It is well established that there is gender-based DIF in many autism scales (Schiltz & Magnus, 2020; Schwartzman et al., 2022). In cross-linguistic studies, translated items may differ in emotional connotation or idiomatic structure, introducing unintended bias (Krogsgaard et al., 2021; Nielsen et al., 2023; Orlando & Marshall, 2002; Terluin et al., 2016). These examples highlight that DIF is not limited to demographic comparisons but also applies to latent or overlapping diagnostic subgroups, especially when working with multidimensional assessments.

It is important to distinguish DIF from impact. Impact refers to true mean differences with respect to the distributions of latent traits between groups. Formally, if group A and group B have latent trait distributions with means μ_A and μ_B respectively, impact is present when $\mu_A \neq \mu_B$. DIF represents a threat to validity that should be detected and adjusted for, whereas impact represents valid construct variance that needs to be preserved. Misattributing impact to DIF leads to inappropriate test modifications that reduce sensitivity to real group differences. Conversely, misattributing DIF to impact leads to retention of biased items and unfair assessment of affected groups.

However, the presence of impact complicates the detection and interpretation of measurement bias. When both impact and DIF are present simultaneously, traditional DIF detection methods may produce misleading results (Shealy & Stout, 1993). If a focal group has both lower average trait levels (i.e., negative impact) and items that disadvantage them (DIF), standard approaches that condition on observed assessment scores rather than true latent traits

may fail to separate those effects. The observed group differences in item responses reflect both the true trait difference (impact) and the measurement bias (DIF), making it difficult to isolate the DIF component. Advanced DIF detection methods attempt to address this by estimating and conditioning on the latent trait distribution rather than observed scores, but this remains an active area of methodological development (De Ayala & Santiago, 2017; DeMars & Lau, 2011; Sajobi et al., 2022).

DIF Detection

In research settings, it is impossible to know if an item truly demonstrates DIF; however, it is possible to evaluate the evidence for DIF. A key difficulty is that many traditional DIF detection methods assume the grouping variable is known and explicitly coded (e.g., gender, ethnicity, diagnostic category; Battauz, 2019; Hambleton et al., 2011; Wald, 1947). Additionally, most DIF detection methods only assess evidence of DIF using two groups (i.e., the reference group and the focal group) created using a single, binary demographic indicator. The reference group typically refers to the majority group in the sample or the group that has historically been advantaged. The focal group typically refers to the minority group or the group that has historically been disadvantaged (Jodoin & Gierl, 2001). The primary benefit of defining a reference and focal group is to interpret the evidence of DIF in terms of how the items are advantaging or disadvantaging the groups. However, this marginal approach to group membership fails to account for the complex intersection of multiple social identities within an individual (Crenshaw, 1991).

If one were to try to account for intersectional identities by use of marginal DIF detection methods, multiple pairwise comparisons between all grouping categories would need to be performed (Albano et al., 2024; Russell & Kaplan, 2021). However, the number of comparisons required is a product of the number of levels in all examined demographic groups. For example,

both Albano et al. (2024) and Russell and Kaplan (2021) utilized the same empirical dataset, which contained item responses from 19,490 students. Russell and Kaplan (2021) created 16 groups based on gender (Male or Female), race (Black, Asian, Hispanic, and White), and economic status (Advantaged and Disadvantaged) and most groups contained between 1,000 and 5,000 individuals. Albano et al. (2024) created six intersectional groups between gender (Male or Female) and race (Black, Hispanic, or White), and most groups contained at least 1,500 individuals. Both studies concluded that adapting marginal DIF methods to perform intersectional DIF detection through this multiple comparison approach successfully identified DIF that would otherwise have gone undetected. However, the large sample size (N = 19,490) used in both studies was much larger than is reasonable to expect in social sciences research.[5] As such, use of these methods for DIF detection in social sciences research may be unreasonable.

Additionally, exclusively focusing on observed groups may obscure other factors contributing to DIF (Zumbo & Gelin, 2005). DIF can also result from unknown or latent subgroups, such as unmeasured cultural identities, treatment experiences, or symptom clusters not captured by observed variables (Hoover, 2022; Schwartzman et al., 2022; Zumbo & Gelin, 2005). Current methods developed to detect DIF caused by a latent variable are not readily applicable to the Likert-type scales commonly used in psychological research. Most approaches use mixture IRT models, which aim to detect latent classes of respondents with differing item parameter profiles without requiring a prespecified grouping variable (Frick et al., 2012; Rost, 1990). Although polytomous extensions of these models exist—including the mixture graded response model—their application is substantially more complex than their dichotomous

[5] It is worth noting that more recent methods have leveraged machine learning techniques to develop methods to test for DIF across more complex groups (Bauer et al., 2020; Classe & Kern, 2024; Quan & Wang, 2026), but these methods have not been widely adopted in the literature. Additionally, they still assume that the variables across which DIF occurs have been measured by the researcher.

counterparts. Simulation research has demonstrated that mixed polytomous IRT models applied to short scales with ordered response categories, conditions typical of survey and clinical research, require sample sizes of at least 2,500 to yield accurate parameter and standard error estimates (Kutscher et al., 2019). A further challenge lies in an interaction between two conceptually distinct phenomena: DIF and impact. Impact reflects differences in the true mean level of the latent trait. Because estimation is sensitive to the latent ability distribution, groups that differ in mean latent trait levels can produce spurious latent classes in the absence of any genuine DIF, while classes defined by DIF alone (and no mean differences) may be poorly recovered (De Ayala & Santiago, 2017; DeMars & Lau, 2011; Frick et al., 2015). Non-normal latent trait distributions, which are common in clinical samples, increase the risk of spurious classes (Sen et al., 2016). Model identification also requires class-invariant anchor items, the selection of which is itself an unsolved problem in the mixture IRT context (Cho et al., 2016; Choi & Cohen, 2019; Karadavut, 2021; Zou, 2025).[6] Finally, the number of latent classes must be specified or selected using information criteria that require a sample size of at least 1,000 for accurate estimation. Given these constraints, use of latent-class DIF detection methods may be unreasonable in many psychological research contexts where samples are often modest in size, scales are relatively short, and trait distributions may not be normal.

Most DIF detection procedures have been developed for unidimensional scales (Lord, 2012; Mantel & Haenszel, 1959; Shealy & Stout, 1993). Investigating DIF in multidimensional scales with these procedures may not be appropriate, leading to unintended consequences such as

[6] A noteworthy recent development that postdates the simulation work reported here is Wallin and Huang (2026), who propose a hybrid latent-class IRT model for ordinal data that simultaneously estimates class membership and identifies DIF items without pre-specified anchor items or known group labels. Because this preprint appeared after the present study was completed, it was not feasible to incorporate this approach, but it represents a promising direction for future work that may circumvent the anchor selection challenges documented in this literature review.

misidentification of DIF resulting from different conditional distributions of latent traits for different participant subgroups (Bulut & Suh, 2017). DIF detection approaches such as the multidimensional simultaneous item bias test (SIBTEST; Stout et al., 1997), IRT likelihood ratio (IRT-LR) test (Suh & Cho, 2014), and multiple indicators multiple cases (MIMIC) model (S. Lee et al., 2017) have all been shown to work well to detect DIF in multidimensional scales (Bulut & Suh, 2017). These approaches, however, still require the researcher to specify the grouping variable a priori. To my knowledge, no method currently exists that simultaneously accommodates unobserved group membership, a multidimensional latent structure, and ordinal responses in a unified DIF detection framework. Existing approaches that handle latent groups generally assume unidimensionality, while methods that accommodate multidimensionality or complex group structures generally require observed grouping variables.

Differential Test Functioning (DTF)

While DIF focuses on bias at the item level, it is important to recognize that multiple items with DIF can accumulate to produce differential test functioning (DTF; Chalmers et al., 2016; Yavuz Temel, 2023), a form of bias that impacts the overall scale (or test) score. In other words, even if individual items show only modest DIF, their combined effect can shift scale scores for members of particular groups, leading to over- or under-diagnosis (Chalmers et al., 2016; Yavuz Temel, 2023). DTF is especially critical in diagnostic contexts, where the scale score is used to make the classification decision. For instance, if a depression screener includes multiple items that exhibit DIF caused by age, older adults may systematically score lower, potentially resulting in false negative diagnoses (Gonzalez & Pelham III, 2021). Conversely, in educational assessments, test-level bias could lead to over-identification of learning disorders in linguistically diverse populations (Hambleton et al., 2011; Zumbo, 1999).

Given these risks, it is essential to evaluate not only individual item properties but also the aggregate functioning of a scale across subgroups. In some cases, DIF may cancel out across items, but in others it may compound and distort diagnostic thresholds. This is particularly relevant for clinical cut-points, where scale-level bias can directly influence whether a person receives a diagnosis or not. This underscores the importance of evaluating a scale for both DIF and DTF (Camilli & Shepard, 1994).

Chapter 3: RF-Based Diagnostic Classification

Machine learning (ML) models have gained popularity as an alternative set of methods for diagnostic assessment (Boness et al., 2020; Epistola, 2022; Giannouli & Kampakis, 2024; Gonzalez, 2021; Holden et al., 2011). These machine learning methods differ fundamentally from psychometric modeling, as ML methods do not assume a theoretical construct or dimensional structure. Put another way, ML methods are nonparametric. This chapter introduces random forest (RF; Breiman, 2001) as one nonparametric machine learning approach to diagnostic classification.

RF is a nonparametric, ensemble-based predictive algorithm. By constructing an ensemble of decision trees, each trained on a different bootstrapped sample and using a random subset of predictors at each split, RF captures complex interactions and non-linear effects without imposing a parametric model.

Decision Trees

Decision trees were developed to discriminate between classes of objects (Jacobucci et al., 2023). The purpose of decision tree algorithms is to determine underlying patterns in a given dataset between a set of inputs (e.g., scale items) and an output (e.g., diagnostic classes) using easily interpretable rules that resemble human reasoning (Figure 6).

Decision trees use recursive partitioning, a process that repeatedly splits a dataset into two distinct groups based on their relationship with the outcome. Using a scale meant to diagnose ADHD as an example, we can consider the tree in Figure 6 with three scale items as predictors. Splitting rules are used to identify the subsets of the individuals who are most similar. For example, the first split partitions the data into two groups based on how participants responded to the item "Forget things easily". The left side of the tree (L_1) contains all individuals who responded with less than a four on that item, while those who responded with a four or five are placed into the other group on the right side of the tree (R_1). This process is repeated until each group contains only individuals with the same classification. Note that the L_1 and R_1 are each further split into two groups based on responses to an item addressing attention. Each node in the tree specifies the class breakdown of the outcome within each group as well as the percentage of individuals in the full dataset that fall within that region. For example, the root node (the first in the tree) indicates that 53% of the sample falls into the "control" group, while 47% of the sample falls into the "ADHD" sample. The process of determining how the splits are made and evaluated differs depending upon the specific algorithm. This study focuses on the use of the classification and regression tree (CART) algorithm (Breiman et al., 1984).

The CART Algorithm

The CART algorithm is a greedy algorithm developed by Breiman et al. (1984) to conduct both classification and regression trees. The term "greedy" indicates that the CART algorithm searches for the best possible outcome at the current step, without considering any previous or future steps (Berk, 2008). The CART algorithm is driven by three main decisions: (1) whether the outcome is continuous or discrete, (2) the selection of the splitting variable, and

(3) the stopping criteria. As this dissertation is focused on classification, only classification trees, where the outcome is discrete, will be discussed.

Variable Splitting and Stopping Criteria. The CART algorithm selects the item and item threshold (partitioning value) that splits the data into two groups, minimizing heterogeneity within each group (Breiman et al., 1984). All possible response values of the items are considered decision points to partition the data. A classification tree is grown using the Gini index for variable splitting and predicting which class *k* each observation belongs to by determining the most common (i.e., modal) class for the observations within a given group (Jacobucci et al., 2023).[7] The Gini index (also called Gini impurity) is a measure of node heterogeneity used to assess how mixed the classes are within a node. The Gini index is calculated as

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}), \tag{16}$$

where $\hat{p}_{mk}$ is the proportion of the observations in the group *m* from class *k*. The Gini index equals zero when all observations in the node belong to a single class (perfect purity) and reaches a maximum value when classes are equally distributed (maximum impurity). For binary classification, the maximum Gini index is 0.5.

When evaluating a potential split, the CART algorithm calculates the weighted average Gini index of the resulting child nodes:

$$G(split) = \frac{n_{left}}{n} G_{left} + \frac{n_{right}}{n} G_{right}, \tag{17}$$

where n_{left} and n_{right} are the number of observations in the left and right child nodes, and *n* is the total number of observations in the parent node. The split that minimizes *G*(*split*) is selected.

[7] One could use entropy instead of the Gini index, but as random forest uses Gini, entropy is not discussed here.

This greedy approach ensures that each split maximally reduces heterogeneity, creating more homogenous subgroups at each step (Berk, 2008; Strobl, Boulesteix, & Augustin, 2007).

This process is then repeated on each subgroup to grow the classification tree until a stopping criterion is met. Possible stopping criteria include tree depth, sample size required, or a minimum improvement in prediction accuracy. As the random forest algorithm (discussed later) uses sample size as its stopping criteria, other approaches are not discussed here. Using sample size as a stopping criterion, the tree is grown until it reaches a specified minimum number of individuals in one or more of the final nodes.

However, trees grown with lenient stopping criteria often overfit the training data, capturing noise rather than signal. To address this, CART employs cost-complexity pruning, which systematically removes branches that provide minimal improvement to model fit (Breiman et al., 1984). The pruning process works backwards from the full tree, collapsing internal nodes to create progressively simpler trees. Each potential pruned tree is evaluated using a complexity parameter (*cp*), which penalizes tree size. For a given *cp* value, a split is retained only if it improves the model fit by at least *cp*. Cross-validation is used to identify the optimal *cp* value: the tree is evaluated on the hold-out folds, and the cross-validation error is calculated for each *cp* value. A common strategy to determine the optimal *cp* value is the one-standard-error rule, which selects the simplest tree (largest *cp*) whose cross-validation error is within one standard error of the minimum. Although one could choose the *cp* that produces the tree with the minimum cross-validation error, use of the one-standard-error rule allows for a balance of parsimony and accuracy (Breiman et al., 1984).

The pruning process addresses the bias-variance tradeoff inherent in tree-based models. Large, unpruned trees have low bias because they can fit complex patterns in the training data,

but they have high variance as they are sensitive to small changes in the training sample, leading to overfitting (Breiman et al., 1984; James et al., 2021). In contrast, small, heavily pruned trees have higher bias because they cannot capture all relevant patterns, but they have lower variance as they generalize to new samples better (Breiman et al., 1984; Hastie et al., 2009). The optimal tree size balances these competing concerns. Cost-complexity pruning with cross-validation provides a principled method for finding this balance: trees that are too large have poor cross-validation performance due to overfitting, while trees that are too small have poor performance resulting from underfitting.

Random Forests

Although RF can be used for both classification and regression tasks, the discussion here is focused on their use for classification tasks. RF for classification can either predict (1) probabilities (indicating the likelihood that an individual belongs to a given class) or (2) discrete class labels based on voting across decision trees. When probabilities are predicted, a decision threshold (i.e., probability cut-point) must be applied to convert the continuous probability into classes. This is not necessary when classes are predicted directly. An overview of the RF approach to classification can be seen in Figure 7.

The RF algorithm (Algorithm 1) is an extension of CART that improves prediction accuracy (Breiman, 2001). Each individual tree in a RF is created using a bootstrap sample of the data, and at each node, a random subset of predictors (i.e., items) is considered when determining the optimal split. Final classifications are made via majority vote or averaged predicted probabilities, the latter of which allows for threshold tuning to prioritize sensitivity or specificity depending on the classification goal. This combination of bootstrap sampling (i.e., bagging) and

item subsampling increases predictive accuracy and model stability (Fernandez-Delgado et al., 2014; Speiser et al., 2019).

Algorithm 1 Random Forests Algorithm (Breiman, 2001; Jacobucci et al., 2023)

1. Set the number of trees (*B*)
2. Set the number of variables to consider when creating each partition (*m*)
3. For b = 1 to *B*:
    a. Draw a bootstrap sample of size *N* from the data.
    b. Grow a decision tree on the bootstrap sample, using a random set of *m* variables for each partition.
    c. Use the out-of-bag (OOB) samples to create predictions from the decision tree.
4. Output the ensemble of trees $\{T_b\}_{b=1}^{B}$.
5. Calculate the final predicted values for each individual using:

    a. If a regression approach is taken, use the average prediction across all trees:

$$\hat{f}_{rf}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$$

    b. Else, for classification, use majority voting across all trees where $\hat{C}_{rf}^{B}(x)$ is a class prediction of the b^{th} tree:

$$\hat{C}_{rf}^{B}(x) = majority\ vote\ \{\hat{C}_b(x)\}_1^B$$

The two primary tuning parameters for RF are the number of trees (*B*) and the number of predictors considered at each split (*m*). Typically, *B* is large, falling between 500 and 2,000. Increasing B improves model stability but increases computational cost. The variable subset size, *m*, is recommended to be small (e.g., $\sqrt{p}$, where *p* is the number of predictor variables). Decreasing *m* reduces the correlation between trees and can improve model performance by increasing the diversity of the ensemble (Breiman, 2001). As the original RF algorithm assumes balanced data (i.e., equal proportions of classes), modifications need to be made when data are unbalanced. One such modification is the use of class-specific weighting. Class-specific weighting adjusts the sampling and splitting mechanisms to prioritize underrepresented classes, thereby improving model performance in scenarios where standard RF may exhibit poor minority-class recall.

RF-based classifications have several advantages. First, RF classifications have been shown to be very accurate (Strobl, 2013). Second, as a nonparametric method, RF does not invoke strict assumptions about the data. Similar to other ML methods, RF is algorithmic in nature; the algorithm learns from the data instead of fitting a model to the data. Third, RF is well suited for a complex data structure (e.g., non-linear relationships between items and classes). As such, RF may be able to capture complex, cross-dimensional interactions that may be present in diagnostic assessment data. However, it also means that RF could perpetuate or amplify biases present in the training data.

Interpretability and Variable Importance Measures

An important limitation of RF is model interpretability. RF lacks interpretability because a single decision tree is no longer available (Jacobucci et al., 2023). However, RF also provides measures of variable importance for each candidate predictor (Breiman, 2001; Liaw & Wiener, 2022; Svetnik et al., 2003). Liaw and Weiner (2022) implemented two algorithms for calculating variable importance measures in the *randomForest* R package. The first is based on the Gini criterion. Specifically, at each split of the tree, the decrease in the Gini node impurity is recorded for the variable x_j that was used to form the split. The average of all decreases in the Gini impurity in the forest where x_j forms the split yields the Gini variable importance measure.

The second calculates variable importance as mean decrease in accuracy (MDA) using the out-of-bag (OOB) observations. Because each tree in a RF is grown on a bootstrap sample of the training data, approximately 37%[8] of observations are excluded from any given bootstrap sample (Breiman, 2001). For each observation *i*, predictions are made using only those trees for which observation *i* was not included in the bootstrap sample (OOB trees for that observation;

[8] This proportion arises because the probability that a given observation is not selected in n draws with replacement from n observations is $(1-1/n)^n$, which converges to $1/e$, or about .37 as n increases (Efron & Tibshirani, 1994).

Breiman, 2001). These predictions are then aggregated (via majority vote for classification or averaging for regression) to produce an OOB prediction for observation *i*. The OOB error is then calculated by comparing these OOB predictions to the true outcomes across all observations. The OOB error provides an unbiased estimate of prediction error without requiring cross-validation or a hold-out test set, making it a computationally efficient alternative for model evaluation. Studies have shown that OOB error closely approximates leave-one-out cross-validation error (Efron, 1983).

Variable importance metrics provide a sense of how each variable contributes to classification decisions, but there are limitations of variable importance metrics. First, when multiple predictor variables are strongly correlated, variable importance measures may assign most, or all, of the importance of these variables to a single variable within the correlated set (Grimm et al., 2023). Second, variable importance metrics do not provide information about the nature of the association between the predictor variable and the outcome. That is, the variable importance metrics do not indicate if the association was linear or nonlinear, nor do they indicate the nature of the interaction effects (Grimm et al., 2023). Additionally, RF variable importance measures may yield misleading results when the predictor variables are different types (e.g., some are continuous and others are categorical) or when the number of levels among categorical predictors varies (Strobl, Boulesteix, Zeileis, et al., 2007). As such, while knowing which variables are important for classification is helpful to researchers, the variable importance metrics do not indicate *how* the variables combine to predict the classifications. This limitation cannot be understated, especially in the larger context of diagnostic assessment where classification decisions may need to be clearly interpretable.

Chapter 4: Monte Carlo Simulation Study

A Monte Carlo simulation was used to evaluate the use of both MIRT and RF for classification under increasingly complex conditions. All data generation and analyses were conducted in R (R Core Team, 2024), and R scripts used for the simulation can be found on GitHub (https://github.com/cbain1/difml_open). The study simulates psychological scale data under conditions that vary in number of scale dimensions, DIF strength, number of items that contain DIF, etc. Cross-validation methods were used to evaluate the generalizability of classification methods. A total of 152 conditions were evaluated, each designed to reflect realistic psychometric challenges. Each condition was replicated 1000 times. Each dataset was then analyzed with four IRT-based approaches and one RF-based approach to classification.

Simulation Conditions

A fully crossed design was used, manipulating the variables seen in Table 1. A detailed breakdown of these conditions can be found in Appendix A. The sample size of 1,000 was chosen to reflect a realistic clinical dataset while ensuring stable parameter estimation and convergence in both IRT and RF models, particularly in multidimensional conditions.

Latent Trait and Diagnostic Score Generation

Latent Trait Generation with Measurement Error. Two distinct latent trait structures are generated for each simulee, approximating realistic measurement conditions: a gold standard trait vector and an observed trait vector that includes measurement error. Each simulee's gold standard latent trait vector $\boldsymbol{\theta}_{gs} = (\theta_{gs,1}, \ldots, \theta_{gs,D})$ is sampled from a multivariate normal distribution with a mean of zero and a given covariance matrix Σ. For unidimensional scales (D = 1), this simplifies to a univariate normal distribution with mean zero and variance one. For multidimensional scales (D = 3), the correlations between dimensions, or the off-diagonal

elements of Σ, are fixed to be 0.5, reflecting one common structure of correlated symptom dimensions in psychological constructs (Bovin et al., 2016; Caspi et al., 2014; Henry & Crawford, 2005; Lahey et al., 2012; Zanon et al., 2021).

In practice, psychological assessments do not measure latent traits perfectly. To simulate realistic measurement conditions, a second set of latent trait scores is generated for each simulee to incorporate measurement error. This observed trait vector $\boldsymbol{\theta}_{\text{scale}} = (\theta_{\text{scale1}}, \ldots, \theta_{\text{scaleD}})$ is constructed by adding noise to the gold standard traits:

$$\theta_{\text{scale}} = \rho \times Z(\theta_{gs}) + \sqrt{(1-\rho^2)} \times Z(\epsilon), \tag{18}$$

where $Z(\cdot)$ denotes standardization, ϵ is a matrix of independent random normal deviates (same dimensions as $\boldsymbol{\theta}_{\mathbf{gs}}$), and ρ is the correlation parameter that controls the strength of the relationship between the gold standard and the observed traits. The parameter ρ takes values of either 0.5 or 1.0 across conditions such that when ρ is equal to 1.0, no measurement error is present.

This simulates the measurement error in trait estimation that occurs when assessments have limited reliability or when items do not perfectly capture the underlying construct (Cheng & Yuan, 2010; Gonzalez et al., 2021; C. Wang et al., 2019). The simulated matrix of scale traits, $\boldsymbol{\theta}_{\mathbf{scale}}$, is used as the assumed latent trait values when generating item responses, while the simulated matrix of gold standard traits, $\boldsymbol{\theta}_{\mathbf{gs}}$, is used to generate diagnostic classes. This simulates a realistic scenario where classification methods must recover true diagnoses from imperfectly measured data.

Impact: Group Differences in Latent Traits. To evaluate how classification methods perform when groups differ in their latent trait levels, impact is systematically manipulated by shifting the gold standard traits of focal group members. After generating the initial gold standard latent traits $\boldsymbol{\theta}_{\mathbf{gs}}$ for all individuals from the multivariate normal distribution, focal group

members are identified based on an interaction of demographic variables (described in the DIF manipulations section below). Impact is then simulated by shifting the latent traits of focal group members such that $\theta_{gs,\ focal} = \theta_{gs,\ focal} + \delta$, where $\delta \in \{-1, 0, 1\}$ represents the impact magnitude. Negative impact conditions ($\delta = -1$) simulate scenarios where the focal group has genuinely lower symptom severity, such as lower depression rates in populations with stronger support systems. Positive impact conditions ($\delta = 1$) simulate scenarios where the focal group has genuinely higher symptom severity, such as higher anxiety rates in populations experiencing chronic stress. The magnitude of δ represents a large effect size (Cohen, 2009), creating meaningfully different diagnostic rates across groups while remaining within plausible ranges observed in real group differences. These values are consistent with published simulation research on impact (Bolt, 2002; Kim & Oshima, 2013; W.-C. Wang & Su, 2004; Woods, 2009).

Impact is applied uniformly across all latent dimensions. For three-dimensional scales, all three dimensions are shifted by the same amount for focal group members. This assumes that group differences affect all symptom domains proportionally, which simplifies the simulation design while still capturing the essential features of impact.

Diagnostic Class Generation via Euclidean Distance

Generated diagnostic classes were derived using a Euclidean distance-based approach in multidimensional latent trait space, applied to the gold standard traits $\boldsymbol{\theta}_{\mathbf{gs}}$. This method treats diagnostic classification as a proximity problem, where simulees whose latent trait profiles are closest to a theoretical "diagnosed" prototype are assigned to the diagnosed class. This approach was chosen because it (1) naturally extends to multidimensional contexts without requiring dimension-specific thresholds, and (2) creates realistic classification boundaries that do not

perfectly align with any single classification method being tested, thereby providing a fair benchmark for comparison (Cover & Hart, 1967; Spence, 1983; Winsberg & De Soete, 1993).

For a simulee *i* with gold standard trait scores $\boldsymbol{\theta_i} = (\theta_{i1}, \theta_{i2}, ..., \theta_{iD})$ across *D* dimensions, the Euclidean distance to a reference point $\boldsymbol{\theta^*} = (\theta^*_1, \theta^*_2, ..., \theta^*_D)$ is calculated as:

$$d_i = \sqrt{\sum_{d=1}^{D} (\theta_{id} - \theta_d^*)^2}, \tag{19}$$

where d_i represents the distance from individual *i* to the reference point in *D*-dimensional space. Simulees with smaller distances are more similar to the prototype diagnosed case (i.e., the reference point). A diagnostic classification is generated by comparing this distance to a given threshold value *t*:

$$\text{Diagnosed if } d_i \leq t$$

$$\text{Not diagnosed if } d_i > t.$$

The threshold *t* is selected to achieve the desired prevalence rate (30% in this study). Specifically, individuals are rank-ordered by their distance d_i, and the 30% of individuals with the smallest distances are classified as diagnosed.

The reference point $\boldsymbol{\theta^*}$ represents the location in the latent space that characterizes a prototypical diagnosed case. In this study, $\boldsymbol{\theta^*}$ is defined as:

$$\theta_d^* = \max(\theta_{id}) \text{ for each dimension } D. \tag{20}$$

That is, the reference point is located at the maximum observed latent trait value on each dimension. This represents an individual with the highest possible symptom severity across all dimensions in our sample. This choice was made as it (1) aligns with the clinical intuition that diagnosed individuals exhibit elevated symptoms across multiple domains and (2) creates a non-compensatory structure to the generative classification boundary (Junker & Sijtsma, 2001; Rupp et al., 2010; Templin & Henson, 2006).

For an illustrative example of this process, consider a two-dimensional scale ($D = 2$) with three individuals:

Individual A: $\boldsymbol{\theta} = (2.5, 2.3)$

Individual B: $\boldsymbol{\theta} = (1.0, 2.8)$

Individual C: $\boldsymbol{\theta} = (0.5, 0.3)$.

In this example, the maximum observed trait values are $\theta^*_1 = 2.5$ and $\theta^*_2 = 2.8$. The reference point is therefore $\theta^* = (2.5, 2.8)$. The Euclidean distances are:

$$d_A = \sqrt{(2.5 - 2.5)^2 + (2.3 - 2.8)^2} = \sqrt{0 + .25} = 0.5$$

$$d_B = \sqrt{(1.0 - 2.5)^2 + (2.8 - 2.8)^2} = \sqrt{2.25 + 0} = 1.5$$

$$d_C = \sqrt{(0.5 - 2.5)^2 + (0.3 - 2.8)^2} = \sqrt{4 + 6.25} = 3.2.$$

Individual A has the smallest distance and would be most likely to be classified as diagnosed. Despite individual B having a higher score on the second dimension (2.8 vs. 2.3), the low score on the first dimension (1.0) increases the overall distance. This illustrates the non-compensatory nature of the Euclidean distance approach: high scores on one dimension cannot fully compensate for low scores on another. An example of this process for both one and three dimensions can be found in Figure 8.

Several alternative approaches for generating diagnostic classifications were considered. Latent class mixture models could assign individuals to diagnosed versus non-diagnosed classes based on class membership probabilities; however, this approach would conflate the data generation mechanism with IRT-based classifications if a mixture structure was assumed, potentially biasing comparisons (Paxton et al., 2001). Threshold-based classification on a higher-order factor would involve fitting a second-order model and thresholding on the general factor score. This approach would give an unfair advantage to the second-order IRT-based approach

discussed later. Non-compensatory rules with independent dimension-specific thresholds would require all dimensions to exceed predetermined cut-points. While this aligns with some diagnostic criteria, it would give an unfair advantage to the MGRM-based approach discussed later.

As such, the Euclidean distance approach was selected because it does not favor any of the classification methods being compared. It creates a classification boundary in latent space that must be approximated by all methods, providing a neutral benchmark for evaluation. Additionally, this approach is conceptually related to prototype-based classification models used in cognitive psychology and machine learning (e.g., *K*-nearest neighbors with $K = 1$), lending it theoretical support (Cover & Hart, 1967; Rosch, 1975; J. D. Smith, 2014).

This approach, however, does make some assumptions that should be acknowledged. The standard Euclidean distance formula weights all dimensions equally. In reality, some symptom domains may be more diagnostically relevant or important than others. A weighted distance metric (e.g., Mahalanobis distance; Mahalanobis, 1936) could accommodate differential weighting, but this would require specifying the relative importance of each dimension a priori. In this simulation study, equal weighting was used to maintain simplicity and avoid introducing additional arbitrary parameters.

The Euclidean distance creates a spherical (or hyperspherical in higher dimensions) decision boundary around the reference point, meaning that individuals at the same distance from the reference point are equally likely to be classified as diagnosed, regardless of their specific symptom profile. In contrast, real diagnostic criteria often have irregular boundaries (e.g., requiring minimum thresholds on some dimensions but not others). This limitation is acceptable

for the purposes of this simulation study, as the goal is to evaluate how well different methods can recover a consistent ground truth, not to perfectly replicate clinical decision-making.

The reference point is defined as the maximum observed trait values, which means it could theoretically vary if extreme outliers were present. However, with the large sample size (N=1000) and controlled latent trait distributions used in this study, the reference point is expected to be highly stable across replications. Sensitivity analyses confirmed that the reference point varied by less than 0.4 standard deviations across replications within each condition.

Item Parameter Generation

Either 15 or 30 Likert-type items were simulated, each with five response categories. Based on the *a*-parameter distributions in previous research, *a*-parameters were drawn from a normal distribution with $\mu = 1.7$ and $\sigma = 0.3$, and *b*-parameters were generated by first drawing b_1 from a normal distribution with $\mu = -1.5$ and $\sigma = 0.5$ (Hill, 2004; Manapat & Edwards, 2022). Then, a difference parameter was drawn from a normal distribution with $\mu = 1$ and $\sigma = 0.2$, and added to b_1 to create b_2 (see Manapat & Edwards, 2022; Woods, 2007). Since all data sets will have items containing five response options, this process was repeated until a total of four *b*-parameters were created ($b_2 = b_1$ + difference parameter 1; $b_3 = b_2$ + difference parameter 2; $b_4 = b_3$ + difference parameter 3).

Item Response Generation

Items were evenly distributed across dimensions. With one dimension, all items load on the single factor. With three dimensions and 15 items, each dimension has 5 items. With three dimensions and 30 items, each dimension has 10 items. Item responses were generated using the MGRM (Samejima, 1997), and each item was fixed to load onto only one latent dimension, estimating a simple structure.

DIF Manipulations

DIF was introduced based on intersectional group membership (sex × race). Three focal groups (Female × Black, Female × LatinX, and Male × Black) were simulated to exhibit DIF relative to the reference group (Female × White). For affected participants, the slope and threshold parameters were adjusted such that $b'_j = b + s,\ s \in \{0,1,3\}$. Note that all DIF adjustments increase threshold parameters (uniform DIF favoring the reference group). This creates unbalanced DIF where focal group members consistently have higher severity parameters for items.

The proportion of individuals in the focal group varied stochastically across replications due to the random sampling of demographic characteristics. To assess the stability of focal group composition, a sensitivity analysis was conducted across 1,000 replications for each condition. For conditions with DIF present, the mean focal group percentage was 23.0% (SD = 0.013), with a maximum standard deviation of 0.014 across conditions. The maximum observed range in focal group percentage within any single condition was 0.103. This variability reflects the natural sampling variation inherent in randomly assigning demographic characteristics based on population probabilities. The observed variation in focal group composition is sufficiently small to ensure that differences in classification performance across conditions reflect the manipulated factors (dimensionality, DIF strength, impact) rather than substantial fluctuations in sample composition. Conditions without DIF had no focal group designation, as all individuals were treated identically regardless of demographic characteristics.

Classification Methods

IRT-Based Classification

Three IRT-based classification approaches are evaluated, each representing a different scoring rule for how multidimensional latent trait information should be combined to produce diagnostic decisions. These approaches differ in both the psychometric model estimated and the classification rule applied, reflecting different theoretical assumptions about the structure of psychopathology. A fourth IRT-based classification approach was included, not for evaluation, but to determine the best possible classification performance that accounts for estimation error. All models are estimated using the *mirt* package in R (Chalmers, 2012).

Compensatory Second-Order MGRM (Second-Order IRT). A second-order MGRM, as described in Chapter 2, is estimated wherein specific factors representing distinct symptom dimensions load on a higher-order general factor representing overall disorder severity. This model is implemented as a constrained bifactor model, with constraints ensuring mathematical equivalence to the second-order structure. Model parameters are estimated using the EM algorithm, and latent trait scores are computed using EAP estimation with QMC integration with 10,000 quadrature points. As described in Chapter 2, QMC integration is used for all bifactor models regardless of dimensionality because it provides a more accurate approximation of the posterior distribution in the multidimensional latent space created by the general and specific factors (Jank, 2005).

The general factor EAP estimate (θ_G) is extracted for each individual and used as a unidimensional severity index. A classification threshold t_G is determined by applying Youden's index to ROC curve analysis on the training data. Youden's index, calculated as sensitivity plus specificity minus one, identifies the threshold that maximizes classification accuracy while

balancing sensitivity and specificity (Youden, 1950). An individual is classified as diagnosed if $\hat{\theta}_G \geq t_G$. This approach reflects a compensatory classification model wherein high standing on one specific dimension can compensate for lower standing on another dimension through their shared contribution to the general factor (Stucky & Edelen, 2014; C. Wang, 2015). This compensatory structure mirrors diagnostic conceptualizations wherein overall symptom severity, rather than domain-specific severity, determines clinical significance. The use of a single threshold on the general factor simplifies the classification decision and aligns with dimensional approaches to psychopathology that emphasize transdiagnostic factors.

Non-compensatory MGRM (IRT_{multi}). The second approach implements a correlated factors MGRM as described in Chapter 2, with each item loading on a single factor corresponding to its symptom domain. For unidimensional scales (D = 1), a standard unidimensional GRM is estimated using the EM algorithm with adaptive quadrature for EAP estimation. For multidimensional scales (D = 3), the MH-RM algorithm is used for parameter estimation, and QMC integration with 10,000 quadrature points is used for EAP estimation, as is recommended for high-dimensional models (Jank, 2005).

Separate EAP scores are estimated for each dimension ($\hat{\theta}_1, \ldots, \hat{\theta}_D$), and dimension-specific thresholds ($t_1, \ldots, t_D$) are determined independently via ROC analysis with Youden's index applied to training data (Youden, 1950). An individual is classified as diagnosed only if they exceed the threshold on all dimensions simultaneously. In other words, $\hat{\theta}_{i1} \geq t_1$ and $\hat{\theta}_{i2} \geq t_2$ and $\hat{\theta}_{i3} \geq t$. This approach reflects a non-compensatory (conjunctive) classification model wherein high severity on one dimension cannot compensate for low severity on another (Chalmers, 2020; Douissa & Jabeur, 2020). All criteria must be met for diagnosis, making this a more stringent classification rule than the second-order IRT-based approach. This

operationalization is consistent with diagnostic criteria requiring symptoms across multiple domains. This approach assumes that diagnostic constructs are fundamentally multidimensional and that severity of symptoms across all relevant domains is necessary for clinical significance.

Euclidean Distance (IRT_{ED}). The third approach uses the same estimated MGRM from the IRT_{multi} approach, ensuring that any performance differences between IRT_{multi} and IRT_{ED} reflect differences in the classification rule rather than differences in the underlying psychometric model. The same estimation procedures are used: EM with adaptive quadrature for unidimensional models and MH-RM with QMC integration for multidimensional models.

In this approach, however, rather than applying dimension-specific thresholds, a single severity index is calculated based on each simulee's Euclidean distance from the point of maximum trait levels in the *D*-dimensional latent space. The reference point representing maximum symptom severity is defined in Equation 18 and the calculation for Euclidean distance can be found in Equation 19. The negative of this distance ($-ED_i$) is used as a severity score, with larger values (smaller distances from the maximum) indicating greater disorder severity across all domains. A classification threshold t_{ED} is determined via Youden's index applied to the training data, and a simulee is classified as diagnosed if $-ED_i \geq t_{ED}$.

This method was included so as to evaluate if superior performance of one IRT-based approach over the other resulted from its relative match to the generating classification rule. The method uses the same distance-based criterion for classification that was used to generate diagnostic classes, and as such, if classification performance is fully dependent upon classification rule, this approach would perform best in all conditions.

Baseline Euclidean Distance (Baseline IRT). To determine the best possible classification performance, separate models for the focal group and reference group data were fit.

For unidimensional scales ($D = 1$), unidimensional GRMs were estimated using the EM algorithm with adaptive quadrature for EAP estimation. For multidimensional scales ($D = 3$), MGRMs were estimated via the MH-RM algorithm, and QMC integration with 10,000 quadrature points was used for EAP estimation, as is recommended for high-dimensional models (Jank, 2005). For conditions without DIF, a single model was fit to the full sample. EAP estimates for the focal and reference groups were then combined into a single matrix. Classes were predicted using the Euclidean distance-based classification approach described in the Diagnostic Class Generation via Euclidean Distance section above to the combined EAP estimates, with two modifications. First, the reference point was derived from the estimated rather than gold standard trait scores (as would be done in an applied scenario), and the threshold was determined by maximizing Youden's index rather than by a fixed prevalence rate.

Random Forest Classification

RF is used as a nonparametric machine learning approach for classification. RF classifiers were trained on the item-level information, treating each item as a predictor and the binary class label as the outcome (diagnosed or not diagnosed). This approach makes no assumptions about the dimensionality of the scale, functional form of item responses, or distributional properties of latent traits, allowing the algorithm to discover complex, nonlinear patterns in the data through recursive portioning (Breiman, 2001). The RF approach represents a fundamentally different paradigm from IRT-based classification: rather than estimating latent traits and applying thresholds, RF uses an ensemble of decision rules to illustrate the pattern of responses for each diagnostic class.

Each RF was set to have 1,000 decision trees, with each tree grown on a bootstrap sample of the training data. The number of predictors considered at each split (*mtry*) was set to the

square root of the number of predictors ($\sqrt{p}$), following standard recommendations for classification tasks (Breiman, 2001). This random predictor selection at each node promotes diversity among trees by ensuring that different trees partition the predictor space in different ways. This diversity improves ensemble performance by reducing correlation between individual tree predictions. Trees are grown to their maximum depth with terminal nodes containing cases from a single class when possible, or when nodes contain fewer observations than can be further split.

To address potential class imbalance, stratified bootstrap sampling is implemented within each tree to equalize class representation during tree growth, as is recommended (Jacobucci et al., 2023). Specifically, the *sampsize* parameter is set equal to the size of the minority class for both classes, ensuring that each tree is trained on equal numbers of diagnosed and non-diagnosed cases. This balanced sampling approach prevents the algorithm from developing a bias toward predicting the majority class, which would result in high specificity but more sensitivity (Jacobucci et al., 2023). Without this adjustment, RF would tend to classify most individuals as non-diagnosed because this maximizes overall accuracy when the base rate is 30%. The stratified sampling ensures that sensitivity and specificity receive equal consideration during tree growth. RF models were implemented using the *randomForest* and *caret* packages in R, allowing for efficient training and integration with cross-validation procedures (Kuhn, 2008; Liaw & Wiener, 2022).

Evaluation Metrics

For each condition, model accuracy, sensitivity (true positive rate), specificity (true negative rate), precision, NPV, and F1 were calculated to assess classification performance.

Cross Validation

To ensure robust and generalizable estimates of classification performance, all methods were evaluated using a hold-out test set approach. RF also implemented OOB estimation. This strategy balances computational efficiency with rigorous assessment of model generalization, following recommendations for Monte Carlo simulation studies in psychometrics (Jacobucci et al., 2023; Paxton et al., 2001; Schroeders & Gnambs, 2025).

Train-Test Split Design

For each simulation replication, two independent datasets were generated: a training set for model development and a holdout set for performance evaluation. The training dataset consists of 1,000 simulees and is used to train all classification models. The sample size is sufficient for stable parameter estimation in multidimensional IRT models with up to three dimensions (Jiang et al., 2016) while remaining computationally feasible. For IRT methods, the training data is used to estimate item parameters and to compute latent trait scores via expected a posteriori (EAP) estimation. For random forests, the training data is used to grow the ensemble of 1,000 decision trees through recursive partitioning with bootstrap sampling.

The test dataset contains 10,000 completely independent simulees generated using the same item parameters and data generation process as the training data, including identical DIF items, impact levels, and measurement error structures, but with different random seeds to ensure statistical independence. This large test set provides precise estimates of classification performance by minimizing sampling variability in performance metrics. The 10:1 ratio of test to training sample size follows recommendations for obtaining stable estimates of model generalization error (Beleites et al., 2013). The test set is used exclusively for prediction and performance evaluation, ensuring unbiased evaluation of out-of-sample generalization performance.

The use of independent train and test datasets, rather than resampling methods such as *k*-fold cross-validation, offers several advantages in this simulation context. Generating two independent datasets is substantially faster, which is critical given the computational burden of 152,000 unique datasets. The large test set of 10,000 observations provides more precise performance estimates than would be obtained from smaller cross-validation folds, where each fold might contain only 100 observations in 10-fold cross-validation applied to 1,000 cases. The test set is completely independent of the training data, providing the strongest possible test of generalization without any potential information leakage between training and testing phases. All methods are evaluated on identical test sets within each replication, ensuring fair comparison across classification approaches.

Classification Performance Analysis Plan

Simulation results from this study were analyzed using the classification and regression tree algorithm (CART; Breiman et al., 1984). The random forests algorithm was also used to assess variable importance (Breiman, 2001). Use of CART and RF variable importance follows the approach developed by Gonzalez et al. (2018) for analyzing complex simulation studies. This approach is outlined in Algorithm 2. CART is advantageous as it (1) accommodates the nested structure of simulation data, (2) automatically detects higher-order interactions without the need for pre-specification, (3) handles both continuous and categorical predictors well without transformation, and (4) provides interpretable decision rules that directly inform practice (Gonzalez et al., 2018).

Algorithm 2 CART and Ensemble Tree Algorithms for Monte Carlo Simulation Analysis

Input: Monte Carlo simulation results with $N = R \times C$ replications, performance metric Y, and simulation factors $\{X_1, \ldots, X_p\}$
Output: Pruned CART model, ensemble model performance, and variable importance measures

For each performance metric *Y*:

1. Convert categorical predictors to factors;
2. Stratify data by simulation condition;
3. Randomly split data into training (50%) and testing (50%) sets;
4. Fit an overgrown regression tree on the training data using lenient stopping criteria;
5. At each node, select the split that minimizes the residual sum of squares (RSS);
6. Apply cost-complexity pruning using cross-validation;
7. Select the optimal complexity parameter;
8. Prune the tree to obtain the final CART model;
9. Evaluate the pruned tree on the testing data using mean squared error (MSE) and $R^{2;}$

CART is a nonparametric tree-based method that recursively partitions data to identify homogenous subgroups (Breiman et al., 1984). To analyze results of a Monte Carlo simulation, each simulation replication is treated as an observation, with the performance metric (e.g., accuracy, sensitivity) as the outcome and all simulation factors (classification method, scale dimensionality, measurement error, number of items, percent DIF, where the DIF is located, the strength of the DIF, and impact) as predictors. The algorithm identifies which combinations of factor levels lead to the highest or lowest performance, producing an interpretable tree diagram. For example, a terminal node might indicate that "when dimensionality = 3 AND DIF strength > 1 AND method = S-O, mean accuracy = 0.72 (representing 8.5% of all replications)." This provides prescriptive guidance for applied researchers with similar data conditions.

Classification method and whether DIF was in a single factor or multiple factors were treated as unordered categorical variables, while the remaining six predictors were treated as ordered categorical variables to avoid making inferences by interpolating between values. Six different classification trees were used, one to predict each classification metric (e.g., accuracy,

sensitivity, etc.). As separate trees are built for each performance outcome, results are presented separately for each performance metric. All CART analyses were conducted in R using the *rpart* package (Therneau & Atkinson, 2025). The *rattle* package was used for enhanced tree visualizations from the CART approach (William, 2011).

To further examine interaction effects included in the CART models, additional visualizations were created using the *ggplot2* package (Wickham, 2016). Sobol MDA variable importance scores were used to identify which simulation factors were important predictors of classification performance across the classification approaches. The *ranger* package was used to calculate the Sobol MDA scores (Bénard et al., 2022).

Results

Figure 9 presents marginal means and standard deviations for classification performance (i.e., accuracy, sensitivity, specificity, precision, NPV, and F1) across all methods and sample groups, with specific values provided in Table S1 (Appendix A). As expected, the baseline IRT-based approach achieved the most balanced classification performance across metrics. By optimizing Youden's index rather than overall accuracy, this approach maximized the sum of sensitivity and specificity jointly, yielding strong sensitivity (M = 0.78, SD = 0.12) alongside reasonable specificity (M = 0.74, SD = 0.11). This balance came at the cost of precision relative to some other approaches—the RF approach, for example, achieved higher accuracy (M_{RF} = 0.77, M_{Base} = 0.75), specificity (M_{RF} = 0.83, M_{Base} = 0.74), and precision (M_{RF} = 0.62, M_{Base} = 0.58) in the full sample, though with notably lower sensitivity (M_{RF} = 0.65, M_{Base} = 0.78). Taken together, these results are consistent with the intended role of this approach as a performance ceiling for IRT-based classification after accounting for measurement error: it does not represent

the highest achievable value on any single metric, but rather the best attainable balance between sensitivity and specificity when the correctly specified model is fit separately to each group.

IRT_{ED} was included not because it is a reasonable approach researchers might take, but rather because it uses the same Euclidean distance-based decision rule as the baseline approach but estimates the IRT model using pooled data. As such, inclusion allows for a direct comparison between the separate group calibration approach and a method sharing the same decision rule but using a misspecified model, thereby isolating the contribution of correct model specification—rather than the decision rule itself—to classification performance. The IRT_{ED} approach was not the best performing IRT-based classification approach across all metrics, illustrating that the Euclidean distance-based decision rule alone is insufficient for strong classification performance.

Note that when the number of dimensions was one, IRT_{multi} and second-order IRT models were mathematically equivalent, as both models reduce to a unidimensional GRM. Thus, an additional comparison was done examining their relative average performance in multidimensional conditions only. These results (Table S2) show that second-order IRT-based classifications had consistently as high (in terms of specificity and NPV) or higher (all other metrics) average performance than IRT_{multi} classifications. As such, the IRT_{multi} approach was not considered in further analyses.

CART was used to analyze simulation results following the approach in Algorithm 2. An initial lenient complexity parameter (*cp*) value of zero was chosen to grow the full tree. Then the one standard error rule was applied to obtain a pruned tree. As the standard error was incredibly small during the cross-validation process (ranging from 0.002 to 0.007 across all models), applying the one standard error rule resulted in trees that remained too complex for interpretation ($min_{splits} = 70$, $max_{splits} = 110$).

To identify a more parsimonious tree, *cp* values were selected using a distance-based elbow detection method. Following the geometric approach described by Satopaa et al. (2011) and recently validated by Onumayi et al. (2022), the elbow point was identified by computing the perpendicular distance from each point on the normalized cross-validated error curve to the line connecting the curve's endpoints and then selecting the point with the maximum distance. This method identifies the point of maximum curvature where the tradeoff between model complexity and cross-validation error is optimized (Morgado et al., 2023; Onumanyi et al., 2022; Satopaa et al., 2011). The distance-based approach has been shown to be effective and stable across diverse error curves, including those with smooth curvature and noisy perturbations (Onumanyi et al., 2022), making it well-suited for model selection in Monte Carlo simulation analysis. An example of this approach illustrating the process of pruning the tree predicting classification accuracy can be found in Figure 10. Comparison of complexity parameters, tree parsimony, and model fit (both MSE and R^2) across all three parameter tuning approaches (overgrown tree, the one standard error rule, and this distance-based method) can be found in Table 4. Overall, across performance metrics, the distance-based CART decreased hold-out sample model fit ($M_{\Delta MSE} = 0.0024$, $M_{\Delta R2} = 0.0676$) as compared to the one standard error rule, but increased model parsimony ($M_{\Delta Splits} = 77.67$). As such, the CART models used to interpret Monte Carlo simulation results were fit using the distance-based selected *cp* value.

Sobol MDA variable importance metrics were calculated for each outcome separately for the RF and second-order IRT-based classification approaches to determine the relative importance of each manipulated data variable on classification performance within classification approach. All Sobol MDA values can be found in Table S3.

Accuracy. CART analysis determined that measurement error was the strongest predictor of accuracy (Figure 11). When measurement error was present ($\rho = 0.5$), accuracy was worse than in conditions without measurement error ($\rho = 1.0$). This result was expected, as it has been well documented that higher measurement error leads to higher prediction error (Blake & Gangestad, 2020; de Ron et al., 2022; Gilbert, 2025; Hasan & Chu, 2022; Holsclaw et al., 2015; Jacobucci & Grimm, 2020; Pankowska et al., 2025; Vowels, 2023).

In the measurement error conditions, second-order IRT-based classifications from a unidimensional assessment ($D = 1$) were the least accurate (M = 0.448, SD = 0.129), followed by second-order IRT-based classifications from a multidimensional assessment ($D = 3$; M = 0.665, SD = 0.039). The most accurate classifications were produced by the RF approach (M = 0.697, SD = 0.020); however, these classifications are worse than the classifications of a naïve model (i.e., every simulee was predicted into the majority, or null, class). In conditions with no measurement error ($\rho = 1$), RF formed a terminal node (M = 0.851, SD = 0.029). As this node did not have an average accuracy value that was clearly higher than all terminal nodes resulting from the second-order IRT branch, additional data visualizations were created to further examine relative performance of RF and second-order IRT-based classifications across different data conditions.

In conditions where RF and second-order IRT-based classification accuracy fell below the base rate accuracy of 0.700 (blue terminal node of Figure 11), second-order IRT-based classification performance was found to be dependent upon both measurement error and dimensionality. RF classifications were more accurate than second-order IRT-based classifications, regardless of measurement error and dimensionality, with the largest difference in performance occurring in unidimensional scales (Figure 12). When scales were

multidimensional, RF classifications were only slightly more accurate than second-order IRT-based classification

Variable importance results align with CART analyses, in that the only manipulated simulation factor that had an effective impact on RF classification accuracy was measurement error (Figure 13). Measurement error decreased RF classification accuracy by 15.4%, on average. The most important manipulated factor for second-order IRT-based classification accuracy was dimensionality, followed by measurement error, number of items, impact, *b* strength, and percent of items that contained DIF. DIF in one factor versus DIF in multiple factors was not found to be important to either RF or second-order IRT-based classifications (i.e., a negative Sobol MDA value; Figure 13).

Sensitivity. CART analysis determined that measurement error was the strongest predictor of sensitivity (Figure 14). Measurement error ($\rho = 0.5$) decreased classification sensitivity as compared to no measurement error ($\rho = 1.0$). This result was once again expected.

When measurement error was not present, RF classifications from longer scales were the most sensitive ($M = 0.815$, $SD = 0.040$), but RF classifications were not always more sensitive than second-order IRT-based classifications. The sensitivity of second-order IRT-based classifications was dependent on a variety of simulation variables. Additional analyses of short, unidimensional scales revealed that RF classifications were always more sensitive than second-order IRT-based classifications. However, in short multidimensional scales, RF classifications were found to be less sensitive than second-order IRT-based classifications when the focal group had either the same or higher average θ levels and DIF was either absent or small ($b_{\text{Focal}} - b_{\text{Ref}} = \{0, 1\}$; Figure 15A).

When measurement error was present and the scale was unidimensional, RF classifications were more sensitive (M = 0.505, SD = 0.04) than second-order IRT-based classifications (M = 0.375, SD = 0.04). When the scale was multidimensional, the sensitivity of second-order IRT-based classifications was dependent on a variety of simulation variables. Further analyses determined that RF classifications were less sensitive than second-order IRT-based classifications when the focal group had either the same or higher average θ levels and DIF was either absent or weak ($b_{Focal} - b_{Ref} = \{0, 1\}$; Figure 15B). Notably, however, this effect did not depend on scale length (as was the case in conditions without measurement error).

Variable importance results indicate that measurement error is the most important predictor in RF approaches, and scale dimensionality is most important in second-order IRT-based classification approaches (Figure 16). Measurement error decreased the sensitivity of RF classifications by 27.5%, on average; however, this varied across the number of items in the scale. Whether DIF was in one factor or multiple factors was not found to be important for either RF or second-order IRT-based classification sensitivity (i.e., a negative Sobol MDA value; Figure 16).

Specificity. CART analyses determined that classification method was the strongest predictor of specificity (Figure 17). When no measurement error was present, RF classifications had the highest average specificity (M = 0.878, SD = 0.023), and all RF terminal nodes had acceptable average specificity levels ($M_{min} = 0.777$, $M_{max} = 0.878$). However, RF classifications were not always more specific than second-order IRT-based classifications (far left yellow node of Figure 17).

As such, additional analyses were needed. Second-order IRT-based classification performance was found to be dependent upon all simulation variables except for if DIF items

were in one factor versus all factors. The only terminal node where average second-order IRT-based classification specificity exceeded the minimum RF average specificity occurred when a scale was unidimensional and the majority of items contained DIF (far left yellow terminal node in Figure 17). As such, the influence of other simulation factors (measurement error, impact, number of items, and DIF strength) on the specificity of classifications from both methods was investigated in only these conditions (D = 1 and percent DIF = 0.66; Figure 18).

Variable importance results indicate that measurement error is the most important predictor in RF approaches, and scale dimensionality is most important in second-order IRT-based classification approaches (Figure 19). Measurement error decreased the specificity of RF classifications by 10.1%, on average. Whether DIF was in one factor or multiple factors was not found to be important to either method (i.e., a negative Sobol MDA value; Figure 19).

Precision. CART analyses determined that measurement error was the strongest predictor of precision (Figure 20). In conditions without measurement error, RF classifications had acceptable precision (M = 0.735, SD = 0.047) and were more precise than second-order IRT-based classifications. In conditions with measurement error, both second-order IRT-based classifications and RF classifications had low precision (i.e., M < 0.70); however, RF classifications were more precise, on average, (M = 0.495, SD = 0.033) than second-order IRT-based classifications ($M_{D=1}$ = 0.256, $M_{D=3}$ = 0.459). CART determined that second-order IRT-based classification precision was dependent on dimensionality. As such, marginal means were examined and found that RF classifications were more precise than second-order IRT-based classifications in both unidimensional (M_{RF} = 0.482, SD_{RF} = 0.025, M_{IRT} = 0.257, SD_{IRT} = 0.111) and multidimensional (M_{RF} = 0.509, SD_{RF} = 0.034, M_{IRT} = 0.459, SD_{IRT} = 0.042) conditions. RF

classifications were slightly better in multidimensional conditions than unidimensional conditions, however, the CART did not split on dimensionality for RF classifications.

Variable importance results indicate that measurement error is the most important predictor for RF precision, although scale dimensionality was almost equally important for second-order IRT-based classification precision. Measurement error decreased the precision of RF classifications by 23.7%, on average. Length of scale was also found to be important for RF precision, such that 30-item scales had 5.1% more precise classifications than 15-item scales. The importance of whether DIF was in one factor or multiple factors was negligible for both RF and second-order IRT-based classification precision (i.e., a negative Sobol MDA value; Figure 21).

NPV. CART analyses determined that measurement error was the strongest predictor of NPV (Figure 22). The highest NPV was observed from RF classifications when no measurement error was present (M = 0.906, SD = 0.022), exceeding the average NPV of all second-order IRT child nodes. When scales were unidimensional and contained measurement error, RF classifications had higher average NPV (M = 0.783, SD = 0.013) than second-order IRT-based classifications (M = 0.626, SD = 0.087). However, when the scale was multidimensional and had measurement error, CART did not distinguish between methods. As such, additional mean analyses were needed, which determined that there was no meaningful difference between the NPV of second-order IRT-based classifications (M = 0.808, SD = 0.040) and RF classifications (M = 0.792, SD = 0.018), providing evidence that the lack of this effect in the CART model was not the result of over pruning the tree.

Variable importance results indicate that measurement error was the most important predictor for the NPV of RF classifications, while scale dimensionality was the most important

predictor for the NPV of second-order IRT-based classifications (Figure 23). Measurement error decreased the NPV of RF classifications by 11.8%, on average. Length of scale was also found to be important for RF precision, such that 30-item scales had an average NPV that was 2.2% higher than 15-item scales. Whether DIF was in a single factor or multiple factors was not important for either RF or second-order IRT-based classifications (i.e., a negative Sobol MDA value; Figure 23).

F1. A standard base rate for the F1 of 0.46 was set by calculating the F1 if all simulees were predicted to be in class 1 (as predicting all simulees to be in the null class would result in an undefined F1 score; Equation 6). CART analyses determined that measurement error was the strongest predictor of precision (Figure 24), decreasing the F1s of both RF and second-order IRT-based classifications. In conditions with no measurement error, F1 of RF classifications far exceeded the base rate (M = 0.760, SD = 0.047) and exceeded the average F1 of second-order IRT-based classifications (M_{max} = 0.726, SD = 0.179). In conditions with measurement error and unidimensional scales, RF classification F1 exceeded the base rate (M = 0.492, SD = 0.028) and average second-order IRT-based classification F1 (M = 0.296, SD = 0.098). CART analyses did not find differences in second-order IRT-based classification and RF classification F1s when scales were multidimensional. Marginal mean analysis supports this finding (M_{IRT} = 0.512, SD_{IRT} = 0.073; M_{RF} = 0.512, SD_{RF} = 0.030).

Variable importance results indicate that measurement error was the most important predictor for RF F1s, while measurement error and dimensionality were almost equally important for second-order IRT-based classification F1s (Figure 25). Measurement error decreased the F1 of RF classifications by 25.8%, on average. Length of scale was also found to be important for RF F1, such that increasing the length of the scale from 15 to 30 items increased

the F1 by 5%. Whether DIF occurred in a single factor or multiple factors was not important to either RF or second-order IRT-based classifications (i.e., a negative Sobol MDA value; Figure 25).

Discussion

The ability to accurately place research participants into diagnostic groups is of vital importance. Researchers typically create these groups using a two-step approach: a scale score is estimated via a psychometric model such as IRT, and then a threshold is applied to assign participants to diagnostic groups based on that estimate. When this approach is used, it invokes an assumption of item invariance. Systematic differences across groups (i.e., DIF) have been shown to bias diagnostic classifications, resulting in the systematic over- or under-diagnosis of individuals in some groups (Gonzalez & Pelham III, 2021; Lai et al., 2017, 2019). When the cause of DIF is simple, such as a single demographic grouping variable, researchers can use multi-group IRT to identify the bias and adjust scores post-hoc. However, when DIF arises from complex group membership (e.g., at the intersection of multiple demographic variables), traditional DIF detection methods may fail, and it may be unreasonable to expect that the IRT-based model needed to account for the DIF can be correctly specified and estimated. This dissertation sought to evaluate whether random forest (RF) is a viable alternative to IRT-based classification in these scenarios. Results suggest that RF classifications are robust to the presence of DIF and therefore may be a viable approach.

A Case for Compensatory Classification Rules

When scales were multidimensional, second-order IRT-based classifications consistently matched or outperformed IRT_{multi} across all classification metrics. IRT_{multi} applies a non-

compensatory classification rule requiring that individuals exceed set thresholds on each scale dimension separately. This rule led to poor classification performance in the present simulations.

Given these findings, it is recommended that researchers use a second-order approach over an IRT_{multi} approach when using IRT to classify participants from a multidimensional diagnostic assessment. One might question whether this result is an artifact of the match between the second-order IRT-based approach and the Euclidean distance-based data-generating mechanism. However, IRT_{ED}, which uses the Euclidean distance-based rule, produced highly imbalanced classifications with near-zero sensitivity and near-perfect specificity, demonstrating that the decision rule is not the only predictor of classification performance. This suggests that the superior performance of the second-order IRT-based classification approach is not a simulation design artifact.

However, by choosing a compensatory classifier over a non-compensatory classifier, an assumption is made about the nature of the construct. In this dissertation, the assumption is that psychopathology is best understood as a composite severity index rather than a profile of distinct facets that need to be accounted for separately when making classification decisions. Compensatory models classify based on the greater construct of interest, while non-compensatory models classify assuming each dimension needs be accounted for separately. Researchers should therefore carefully consider whether the domains of their disorder of interest are all distinctly important to a diagnosis or if a compensatory model better reflects the diagnostic criteria of the disorder (Kroc & Olvera Astivia, 2022; Tamano et al., 2025). It is also worth noting that this recommendation applies specifically to the between-item multidimensionality examined here. When constructs require elevated scores across distinct symptom domains—as in ADHD, which requires clinically significant symptoms in both

inattention and hyperactivity-impulsivity—adopting a compensatory scoring strategy would not align with the theoretical nature of the disorder.

Dimensionality Effects on IRT and RF Classification

Scale dimensionality was the most important predictor of second-order IRT-based classification accuracy, sensitivity, specificity, NPV and F1 (and the second most important for precision), yet had negligible effects on RF classification performance. This finding aligns with the parametric nature of IRT; IRT's estimate of a simulee's latent trait is dependent upon a researcher-specified structure. If the dimensionality of the structure is misspecified, the resulting estimate, and the subsequent classification threshold, is likely to be biased. As such, a classification threshold calibrated under one factor structure or one sample will not necessarily generalize (e.g., separate diagnostic groups accurately) to another structure or sample. The psychometric literature has documented substantial instability in factor analytic solutions across samples and methodological contexts (Goretzko et al., 2025; Goretzko & Bühner, 2022; Manapat et al., 2025), suggesting that the factor structure assumed when calibrating an IRT classification threshold may not hold in the population to which classifications are ultimately applied (Flake et al., 2022; Flake & Fried, 2020).

Second-order IRT-based classifications performed worse and exhibited greater variability across replications in unidimensional conditions relative to multidimensional conditions. This pattern was unexpected given that multidimensional classification is an inherently more complex problem. The variability was largest in short, 15-item unidimensional scales, where a single-group unidimensional GRM was used to estimate scale scores.

One possible explanation for this finding has to do with how the second-order factor structure constrains estimation differently depending on dimensionality. In multidimensional

conditions, the second-order MGRM anchors the general factor estimates across multiple specific factors. This anchoring shrinks the general factor estimate and the corresponding classification threshold, toward the mean. This can decrease the influence of sampling variability on parameter estimation and classification prediction (Huang et al., 2012). In unidimensional conditions, no such anchoring mechanism is available. With only a single factor to estimate, the classification threshold is estimated without the shrinkage effect of the second-order structure, making it more susceptible to idiosyncratic features of individual samples. In short, the very complexity of the multidimensional model provides a regularizing effect that is not present in the simpler unidimensional model.

In applied settings, researchers using an IRT-based classification approach with a short, unidimensional scale, such as a screening instrument, should anticipate variability in classification thresholds across samples. Rather than treating an estimated classification threshold as a fixed value, researchers are encouraged to assess the impact of sampling variability on this value. One practical approach to assessing the impact of sampling variability is to report bootstrap confidence intervals around the estimated classification threshold. These intervals provide a range of θ values within which classifications should be interpreted with caution. Individuals whose trait estimates fall within this range are neither clearly above nor clearly below the threshold, and their classifications may be unreliable if the study were to be replicated. This approach is consistent with more recent recommendations in the diagnostic classification literature to report uncertainty around cut-point estimates rather than treating them as fixed quantiles (Thiele & Hirschfeld, 2023).

Measurement error, indexed by the correlation between the generating gold-standard latent trait and the generating assessment latent trait, was the dominant predictor of RF

classification performance across all outcomes. When measurement error was present, RF classification accuracy, sensitivity, specificity, precision, NPV, and F1 decreased. This finding is consistent with the broader machine learning literature illustrating that classifiers trained on features measured with error inherit that error into their predictions (de Ron et al., 2022; Holsclaw et al., 2015). RF is not an exception to this, and its advantage over IRT in the presence of DIF does not extend to situations where there is a large degree of item-level measurement error. Rather, RF classifications are only accurate when items are reasonably reliable indicators of the latent construct, and complex DIF prevents parametric approaches from correctly modeling the latent structure. This is an important distinction, as psychometric models (e.g., IRT) are the current gold standard for determining how reliable items are relative to the latent construct.

Effects of DIF on IRT- and RF-Based Classifications

Neither DIF percentage nor DIF strength were among the most important predictors of RF classification performance across any outcome metric, and the location of DIF (i.e., whether DIF items were all part of a single factor or spread across multiple factors) was not found to be an important predictor of classification performance for either approach (Sobol MDA values were negative). This pattern is consistent with previous work demonstrating that RF is robust to the presence of unmodeled DIF (Bain et al., 2026) and extends those findings to the more complex multidimensional case. Additionally, this study utilized shorter scales than Bain et al. (2026) that are more representative of commonly used screening measures.

The insensitivity of RF to DIF location is perhaps unsurprising given that RF makes no assumptions about the factor structure of a scale. More surprising, however, is that DIF location was unimportant to second-order IRT-based classification performance. Because second-order

IRT models the factor structure of a multidimensional instrument, one might expect that DIF concentrated within a single factor would behave differently than DIF distributed across multiple factors. Since the former only affects a single first-order factor, it is possible it would contribute less bias to the general factor estimate than if DIF was present in all first-order factors.

Supplemental analyses indicate that IRT_{multi} classification performance was also similar in conditions with DIF in a single factor compared to DIF in multiple factors (Table S3). Unlike second-order IRT-based classification, IRT_{multi} applies independent threshold criteria per dimension with no mechanism for pooling information across subscales. However, since IRT_{multi} requires a simulee to exceed all dimension-specific cut-points to be placed into the diagnosed class, it is possible that the single-factor DIF location conditions led to incorrect classifications based on that single factor alone. As such, whether DIF is in a single factor versus all factors may not impact classification performance when compensatory scoring is used or when very strict non-compensatory classification rules are used. It is possible that if a less strict non-compensatory classification rule was used (e.g., a simulee only need exceed the diagnostic threshold on two of the three dimensions), whether DIF was in a single factor versus all factors may have more of an impact on classification performance.

DIF did moderate the relative performance of RF versus second-order IRT-based classifications in important ways, particularly with respect to sensitivity. When scales were multidimensional and the generating distribution of the focal group θ had a higher mean than the generating distribution for the reference group, second-order IRT-based classifications were more sensitive than RF, regardless of measurement error. RF's classification sensitivity in multidimensional conditions was lower when the average focal group θ was higher. This is likely because RF learned the base rate classification pattern from the full sample without

distinguishing between focal and reference group construct levels which could result in systematic under-detection of true positives in the focal group when that group has higher disorder prevalence. Researchers who suspect that individuals in the focal group have truly elevated symptom levels (or disorder prevalence) compared to those in the reference group should therefore exercise caution before using RF for classification.

Note that all DIF manipulations in this study increased item severity parameters for the focal group, meaning focal group members required a higher trait level than reference group members to endorse an item at the same level. As a result, IRT systematically underestimated focal group θ when DIF was present. This estimation error was compounded when the focal group's gold standard θ values was higher than the reference group's (i.e., positive impact conditions). The sensitivity advantages of a second-order IRT approach to classification over RF approach in these conditions could therefore partly be an artifact of DIF direction. Consistent with this interpretation, second-order IRT-based classifications were not more sensitive than RF classifications in negative impact conditions, where the focal group had lower gold standard θ values than the reference group. In this scenario, although DIF likely caused IRT to underestimate focal group θ values, as the focal group truly had lower θ values than the reference group, the underestimation was not observed at the classification level. Had DIF instead decreased severity parameters for the focal group, IRT would have overestimated focal group θ, and the pattern would reverse. In positive impact conditions, RF classifications would be more sensitive, but less specific, than second-order IRT-based classifications. In negative impact conditions, the overestimation from negative DIF would partially cancel out with the lower prevalence of the focal group, again attenuating differences between RF and second-order IRT-based classification performance.

Potential Limitations of the Use of IRT for Diagnostic Classification

Beyond the DIF and dimensionality findings, simulation results highlight an important pattern regarding the relationship between model specification and classification stability. The number of dimensions in the scale was the most important predictor of second-order IRT-based classification performance, while measurement error most strongly predicted RF classification performance. This reflects a fundamental difference in how these methods express uncertainty: for IRT, classification performance is closely related to correct specification of the underlying latent structure while RF classification performance is most strongly related to the strength of the relationship between the items and the underlying construct.

Additionally, the present results illustrate that classifications based on traits estimated via the GRM were highly variable across replications, even in conditions that did not contain measurement error or DIF. This suggests that the process of estimating a classification threshold based on a single continuous estimated θ distribution is sensitive to sampling variability in ways to which the constrained bifactor structure of the second-order IRT model is less sensitive. Practitioners who rely on the GRM for unidimensional scales should be aware of this possibility, particularly in shorter instruments. It is also likely a similar result would occur in smaller samples, as both number of items and sample size contribute to the overall amount of information that is used when estimating a GRM.

Potential Limitations of the Use of RF for Diagnostic Classification

The present results indicate that RF can produce accurate classifications under conditions of complex, unmodeled DIF in both unidimensional and multidimensional scales. These findings, however, should be considered alongside a practical limitation. Because RF predictions are the result of many individual decision trees, it is as possible to articulate why a given

individual was classified into a particular group as it is when a IRT-based approach to classification is used (Fife & D'Onofrio, 2023; Jacobucci & Grimm, 2020). Post-hoc explanation tools such as Shapley additive explanation (Lundberg & Lee, 2017) can describe which items contributed most to a given prediction, but these approximations are local and are fundamentally different from explanations of the latent structure of the construct. As Rudin (2019) has noted, explaining a black-box model after the fact is a fundamentally different activity from building an interpretable model in the first place. When researchers need to clearly justify group assignments and link these assignments with psychological theory, IRT-based classification is the preferable approach.

Two additional limitations follow from this. First, RF does not produce individual-level standard errors, so there is no direct means of quantifying classification uncertainty for a given individual. Second, application of a RF classifier trained in one sample to another sample is not straightforward. The patterns driving the classification decisions may not generalize, and RFs cannot be compared across studies in the way that IRT parameters can. This limits the contribution of RF-based classifications to cumulative theoretical knowledge about the diagnostic construct. Together, these considerations suggest that RF is best understood as a practical solution for a specific problem, namely achieving accurate classification when complex DIF renders parametric approaches such as IRT-based classification unreasonable, rather than a general-purpose alternative to IRT in all cases (Bain et al., 2026; Doshi-Velez & Kim, 2017; Grimm et al., 2023; Jacobucci et al., 2023; Rudin, 2019; Yarkoni & Westfall, 2017).

Practical Implications

Taken together, the findings from this dissertation offer several recommendations for researchers using psychological scales for diagnostic classification. First, when scales are

multidimensional, researchers should use second-order IRT-based scoring over dimension-specific (non-compensatory) IRT scoring, unless they have strong theoretical justification for the disorder of interest needing non-compensatory scoring. The consistently poor performance of IRT_{multi} classifications relative to second-order IRT-based classifications makes dimension-specific threshold application a questionable default choice.

Second, when complex DIF is present, RF is a viable alternative to IRT-based classification for overall accuracy, precision, NPV, and F1. RF's robustness to DIF makes it well-suited to situations where complex violations of psychometric assumptions make it unreasonable to use a psychometric approach. As participants in psychological research hold complex, intersectional identities, it is likely that complex DIF may be present in the data (Crenshaw, 1991). Previous research has found that adapting common DIF methods to perform intersectional DIF detection may be infeasible in social sciences research, due to sample size limitations. As such, intersectional DIF detection may be infeasible. As such, this study is just one example in a growing body of literature on how machine learning models may be useful when complex DIF is present in data (Bain et al., 2026; Kraus et al., 2022, 2024; Quan & Wang, 2026).

Third, measurement error was the primary predictor of RF performance. When scale scores are not strongly correlated with gold standard assessments, RF performance decreases substantially. This implies that researchers should not view RF as a universal solution to classification problems; rather, RF is most beneficial when items are reasonably reliable and valid indicators of the underlying construct, but DIF increases bias in parametric scoring.

Fourth, when researchers are most concerned with maximizing the sensitivity of classifications (e.g., in screening applications where false negatives carry significant costs), and

when the focal group is expected to have elevated symptom levels relative to the reference group, the second-order IRT approach to classification may produce more sensitive classifications than RF. Researchers should consider the relative costs of false positives and false negatives in their specific context when selecting between approaches (Ryan & Oquendo, 2020; Sheldrick et al., 2015), and should not assume that RF's advantages on other classification metrics extend to sensitivity under these demographic conditions.

Fifth, the finding that DIF location had negligible impact on classification for either method suggests that the distinction between unidimensional and cross-dimensional DIF is less important for classification accuracy than the overall magnitude of DIF. If classification is the ultimate goal, researchers do not need to know if DIF occurs in a single dimension or in all dimensions. Rather, the amount of measurement error introduced into the latent trait estimates, whether through the number of DIF items, strength of the DIF, or the scale being a relatively poor approximation of the gold standard diagnosis (e.g., weak to moderate correlation between the scale and gold standard) is most important.

Finally, researchers using IRT for classification on unidimensional scales should be aware that IRT-based classification thresholds may be more strongly affected by sampling variability than is commonly discussed, and this variability should be explicitly quantified. One practical approach involves constructing bootstrap confidence intervals around the estimated classification threshold (Thiele & Hirschfeld, 2023).

Methodological Considerations: Notes on the Use of CART for Simulation Analyses

The use of CART to analyze Monte Carlo simulation results can offer advantages over descriptive statistics or ANOVA-based analysis methods, including the ability to model higher-order interactions and produce interpretable decision rules (Gonzalez et al., 2018). However,

current use presented challenges that have not been discussed in the literature. When simulation designs involve a large number of factors and a large number of replications, the standard error of the cross-validation error estimates becomes exceedingly small. As a result, implementing the 1-SE rule with the suggested lenient tuning parameters (i.e., *cp* = 0.01 or *cp* = 0.001) provided virtually no pruning relative to the initial overgrown tree. Additionally, trees selected using the 1-SE rule were incredibly large (e.g., 70 to 110 splits), rendering them uninterpretable in the context of identifying meaningful simulation patterns. This could be a consequence of the size of the simulation: with millions of observations in the results data frame, the cross-validation error is estimated very precisely, so even a marginal gain from adding a split is consistently distinguishable from noise. The 1-SE rule, which was designed to balance precision and parsimony, loses its usefulness when cross-validation error approaches zero.

Several features of the present simulation design differed from the studies examined by Gonzalez et al. (2018). First, the two simulations presented by Gonzalez et al. (2018) each evaluated a single statistical method under varying data conditions—power of the mediated effect in one case, and the accuracy of coefficient in the other. Neither simulation compared multiple methods. The present dissertation, by contrast, directly compared the classification performance of two approaches (second-order IRT-based and RF-based) across simulation conditions, meaning that method was itself a predictor in the CART. This additional source of systematic variance, and its interactions with the remaining seven data condition manipulations, increased the complexity of the predictor space. Second, Gonzalez et al. (2018) explicitly noted in their limitations that the CART approach should be tested in simulations with more than the six manipulated features used in their largest demonstration. The present simulation included eight design factors. With more predictors, the algorithm has more candidate splits to evaluate at

each node, which contributes to larger overgrown trees and, in combination with the small cross-validation error estimates, further affects the 1-SE rule's performance. Researchers applying CART to analyze similarly complex simulation designs may also face this challenge and may need to perform *a priori* analyses to reduce the dimensionality of the results matrix (as was done here removing IRT_{multi} and IRT_{ED}), explore alternative pruning strategies, or apply an explicit parsimony criterion (e.g., a fixed maximum tree depth) beyond what cross-validation standard errors alone can provide.

To the author's knowledge, this specific application of elbow detection implemented within this dissertation has never been applied to CART. The elbow method has been extensively utilized in cluster analysis, particularly for selecting the number of clusters in k-means, where it identifies the point at which adding more clusters produces diminishing reductions in within cluster variance (see, e.g., Ketchen & Shook, 1996; Thorndike, 1953). It has also been applied to determine the best number of factors in an exploratory factor analysis (i.e., the elbow rule for scree plots; Cattell, 1966). Extension of this method to tree-based model selection was found to be useful. The method performed well in the present context, yielding trees that were interpretable (averaging 10 to 15 splits across models) while maintaining acceptable holdout-sample fit.

Limitations and Future Directions

As limitations, simulation results are only generalizable to the conditions studied, and results may not necessarily be duplicated in other conditions. The data-generating model used a compensatory model based on Euclidean distance to define diagnostic class membership. This choice was theoretically motivated by dimensional and transdiagnostic models of psychopathology, but it may have contributed to second-order IRT-based classifications

outperforming IRT_{multi} classifications. The recommendation to prefer the second-order IRT approach over the IRT_{multi} approach may, therefore, not generalize to assessments designed with a non-compensatory diagnostic structure. It is also possible that RF would perform poorly with non-compensatory scales, as it does not account for distinct dimensions within a scale. Future research should evaluate classification performance under non-compensatory data-generating mechanisms to determine whether the present findings hold across different theoretical models of a disorder.

The simulation study also examined only two levels of measurement error ($\rho = 0.5$ or 1). While these levels were chosen based on a range of published values (Antons et al., 2026; Bain et al., 2025; Dahlgren et al., 2025; C. R. G. Jones et al., 2024; Kim et al., 2024; Rosenthal et al., 2021), future work should examine more measurement error conditions to identify the correlation at which RF is no longer a good classifier.

Simple summed scores were not included as a comparison approach. Although previous work has shown that IRT and RF outperform summed scores when measurement error is low, samples are large, and scales have many items (Gonzalez, 2021), including a summed score benchmark in the presence of DIF would allow researchers to better evaluate the costs and benefits of using more sophisticated methods. Additionally, all IRT-based thresholds were derived using Youden's index, which treats false positives and false negatives as equally costly and assumes equal group prevalence. However, other strategies exist in the literature. For example, Hassanzad and Hajian-Tilaki (2024) recommend the Index of Union (IU) as a preferable alternative, noting that Youden's index can diverge meaningfully from other methods when score distributions are skewed or group variances differ. Whether the choice of threshold

method affects classification performance and stability, and whether the IU produces more consistent cut-points across replications, are questions that require future research.

A related question concerns whether the threshold instability observed for IRT approaches would be reduced by using a summed score cut-point rather than a θ-based one. A θ-based threshold depends on item parameter estimates that vary across samples, meaning both the score metric and the threshold placement vary from replication to replication (Fink et al., 2025; Sun et al., 2020; Xu & Liu, 2022; Yang et al., 2012). Summed score thresholds do not have this problem because the score metric is fixed by the response format. At the same time, IRT's differential weighting of items is known to improve score precision (Chen et al., 2025; Edwards & Edelen, 2009; O'Connor, 2018), classification accuracy (Kawilapat et al., 2022), and sensitivity to individual-level change (Tang et al., 2023), which suggests that IRT-based thresholds may better separate diagnostic groups when the model is correctly specified. Directly comparing threshold stability across scoring approaches is an important direction for future work.

Additional DIF conditions could also be examined. In this simulation study, the focal group was defined by intersections of simulated race and gender variables, with DIF affecting all focal groups uniformly. This was chosen as it is a simple possible case of intersectional DIF. In practice, DIF tends to vary in direction, magnitude, and which items are affected across group membership configurations. Future work should examine more complex intersectional DIF patterns, including non-uniform and unbalanced DIF. Additionally, all DIF manipulations in the present simulation increased item severity parameters for the focal group, meaning focal group members needed a higher level of the trait than reference group members to endorse items in the same way. This leads to underestimated θ levels, and as such, is likely to increase the rate of

false negatives in the focal group. Future research is needed to evaluate the relative effect of DIF in a single factor versus DIF in multiple factors on classification performance when DIF decreases severity parameters rather than increases them.

The present work does not address the interpretability of the classification methods compared here. RF's ensemble structure makes it difficult to understand why or how particular classification decisions are made, and it lacks a connection to the theoretical latent structure of the construct (Fife & D'Onofrio, 2023; Jacobucci et al., 2023). As RF is increasingly adopted for psychological assessment, developing interpretable extensions (e.g., through SHAP values or related tools; Lundberg & Lee, 2017) or examining more interpretable, but less accurate methods such as decision trees, remains vital. As it stands, if a researcher requires a transparent way to connect observed data with predicted classifications, IRT-based classification is still the best approach.

Additionally, the present work used a stratified bootstrap sampling approach to address class imbalances in the training data. This approach was recommended by Jacobucci et al. (2023), however, other approaches to handling class imbalance could be investigated. For example, one popular approach is the use of class-specific weighting (e.g., Bain et al., 2026; Haixang et al., 2017). Additionally, use of more sophisticated approaches like the synthetic minority oversampling technique (SMOTE; Chawla et al., 2002) could be investigated. Future research could examine the relative performance of these approaches within the context of diagnostic classification with DIF.

Finally, all simulations used a fixed prevalence of 30% for the diagnosed class. Classification performance metrics depend on base rate in complex ways, and the relative advantages of RF and IRT may change as prevalence changes. This matters most for low base

rate disorders such as schizophrenia or rare neurodevelopmental conditions, where even small classification errors can produce large errors in estimated prevalence and where the costs of misclassification tend to be high (Lavigne et al., 2016; Zimmerman, 2022; Zimmerman & Holst, 2018). Previous work has shown that higher prevalence rates (30% compared to 12%), or more balanced class sizes, led to lower classification precision of classifications for those in the focal group (Bain et al., 2026), suggesting that base rate is a meaningful moderator for the patterns observed here.

Conclusion

The purpose of this dissertation was to determine if RF is a viable approach to diagnostic classification from self-report scales that contain complex DIF, contributing to the growing body of literature on the use of nonparametric methods in psychometric contexts (Bain et al., 2026; Classe & Kern, 2024; Gonzalez, 2021, 2025; Gonzalez et al., 2024; Kraus et al., 2022, 2024; Orrù et al., 2020, 2022; Quan & Wang, 2025). Additionally, this study provides prescriptive guidance for researchers when predicting classifications from complex data. Findings suggest that RF is a viable alternative to IRT-based classification in most, but not all, cases.

As psychological assessment continues to become increasingly complex, working with increasingly heterogeneous populations, aligning methodological approaches with both theoretical and empirical demands will be crucial to ensuring validity, fairness, and replicability of diagnostic classifications. By providing evidence-based guidance on the tradeoffs between the use of IRT and RF for complex psychological assessment classification problems, this dissertation contributes to more informed and equitable diagnostic classification in psychological research.

References

Aghvinian, M., & Sergi, M. J. (2018). Social functioning impairments in schizotypy when social cognition and neurocognition are not impaired. *Schizophrenia Research. Cognition*, *14*, 7–13. https://doi.org/10.1016/j.scog.2018.07.001

Albano, T., French, B. F., & Vo, T. T. (2024). Traditional vs Intersectional DIF Analysis: Considerations and a Comparison Using State Testing Data. *Applied Measurement in Education*, *37*(1), 57–70. https://doi.org/10.1080/08957347.2024.2311935

Antons, S., Brandtner, A., Oelker, A., Trotzke, P., Wegmann, E., Brand, M., & Müller, S. M. (2026). Psychometric evaluation of a trans-addiction craving questionnaire: The Craving Assessment Scale for Behavioral Addictions and Substance-use Disorders (CASBAS). *Addiction*, *121*(2), 400–412. https://doi.org/10.1111/add.70197

Avila, M. L., Stinson, J., Kiss, A., Brandão, L. R., Uleryk, E., & Feldman, B. M. (2015). A critical review of scoring options for clinical measurement tools. *BMC Research Notes*, *8*, 612. https://doi.org/10.1186/s13104-015-1561-6

Ayearst, L. E., & Bagby, R. M. (2010). Evaluating the psychometric properties of psychological measures. In *Handbook of assessment and treatment planning for psychological disorders, 2nd ed* (pp. 23–61). The Guilford Press.

Bain, C. M., Manapat, P. D., Manapat, D. M., Brenna, K. B., & Grimm, K. J. (2026). When DIF Goes Unmodeled: Assessing the Viability of Random Forest for Diagnostic Classification. *Journal of Behavioral Data Science*, *6*(2), 1–39. https://doi.org/10.35566/jbds/bainmmbg

Bain, C. M., Norris, J. E., Conley, A., Manapat, P. D., & Ethridge, L. E. (2025). A Psychometric Analysis of the Duke Misophonia Questionnaire. *Journal of Clinical Psychology*, *81*(12), 1195–1212. https://doi.org/10.1002/jclp.70028

Baker, A., Simon, N., Keshaviah, A., Farabaugh, A., Deckersbach, T., Worthington, J. J., Hoge, E., Fava, M., & Pollack, M. P. (2019). Anxiety Symptoms Questionnaire (ASQ): Development and validation. *General Psychiatry*, *32*(6), e100144. https://doi.org/10.1136/gpsych-2019-100144

Bartlett, M. S. (1937). The Statistical Conception of Mental Factors. *British Journal of Psychology. General Section*, *28*(1), 97–104. https://doi.org/10.1111/j.2044-8295.1937.tb00863.x

Battauz, M. (2019). On Wald tests for differential item functioning detection. *Statistical Methods & Applications*, *28*(1), 103–118. https://doi.org/10.1007/s10260-018-00442-w

Bauer, D. J., Belzak, W. C. M., & 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

Beleites, C., Neugebauer, U., Bocklitz, T., Krafft, C., & Popp, J. (2013). Sample size planning for classification models. *Analytica Chimica Acta*, *760*, 25–33. https://doi.org/10.1016/j.aca.2012.11.007

Bénard, C., Da Veiga, S., & Scornet, E. (2022). Mean decrease accuracy for random forests: Inconsistency, and a practical solution via the Sobol-MDA. *Biometrika*, *109*(4), 881–900. https://doi.org/10.1093/biomet/asac017

Berk, R. A. (2008). Classification and Regression Trees (CART). In *Statistical Learning from a Regression Perspective* (pp. 1–65). Springer New York. https://doi.org/10.1007/978-0-387-77501-2_3

Bianchini, K. J., Etherton, J. L., Greve, K. W., Heinly, M. T., & Meyers, J. E. (2008). Classification Accuracy of MMPI-2 Validity Scales in the Detection of Pain-Related Malingering A Known-Groups Study. *Assessment*, *15*, 435–449. https://doi.org/10.1177/1073191108317341

Black, L., Humphrey, N., Panayiotou, M., & Marquez, J. (2024). Mental Health and Well-being Measures for Mean Comparison and Screening in Adolescents: An Assessment of Unidimensionality and Sex and Age Measurement Invariance. *Assessment*, *31*(2), 219–236. https://doi.org/10.1177/10731911231158623

Blake, K. R., & Gangestad, S. (2020). On Attenuated Interactions, Measurement Error, and Statistical Power: Guidelines for Social and Personality Psychologists. *Personality and Social Psychology Bulletin*, *46*(12), 1702–1711. https://doi.org/10.1177/0146167220913363

Blakely, T., McKenzie, S., & Carter, K. (2013). Misclassification of the mediator matters when estimating indirect effects. *Journal of Epidemiology and Community Health*, *67*(5), 458–466. https://doi.org/10.1136/jech-2012-201813

Blanton, H., & Jaccard, J. (2018). From Principles to Measurement: Theory-Based Tips on Writing Better Questions. In *Measurement in Social Psychology*. Routledge.

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

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP Estimation of Ability in a Microcomputer Environment. *Applied Psychological Measurement*, *6*(4), 431–444. https://doi.org/10.1177/014662168200600405

Bolt, D. M. (2002). A Monte Carlo Comparison of Parametric and Nonparametric Polytomous DIF Detection Methods. *Applied Measurement in Education*, *15*(2), 113–141. https://doi.org/10.1207/S15324818AME1502_01

Boness, C. L., Stevens, J. E., Steinley, D., Trull, T., & Sher, K. J. (2020). Using Complete Enumeration to Derive "One-Size-Fits-All" versus "Subgroup-Specific" Diagnostic Rules for Substance Use Disorder HHS Public Access. *Assessment*, *27*(6), 1075–1088. https://doi.org/10.1177/1073191120903092

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*(3), 425–440. https://doi.org/10.1007/s11336-006-1447-6

Borsboom, D., & Cramer, A. O. J. (2013). Network Analysis: An Integrative Approach to the Structure of Psychopathology. *Annual Review of Clinical Psychology*, *9*(Volume 9, 2013), 91–121. https://doi.org/10.1146/annurev-clinpsy-050212-185608

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*(2), 203–219. https://doi.org/10.1037/0033-295X.110.2.203

Bovin, M. J., Marx, B. P., Weathers, F. W., Gallagher, M. W., Rodriguez, P., Schnurr, P. P., & Keane, T. M. (2016). Psychometric properties of the PTSD Checklist for Diagnostic and Statistical Manual of Mental Disorders–Fifth Edition (PCL-5) in veterans. *Psychological Assessment*, *28*(11), 1379–1391. https://doi.org/10.1037/pas0000254

Bradford, A., Meyer, A. N. D., Khan, S., Giardina, T. D., & Singh, H. (2024). Diagnostic error in mental health: A review. *BMJ Quality & Safety*, *33*(10), 663–672. https://doi.org/10.1136/bmjqs-2023-016996

Bradshaw, L., & Templin, J. (2014). Combining Item Response Theory and Diagnostic Classification Models: A Psychometric Model for Scaling Ability and Diagnosing Misconceptions. *Psychometrika*, *79*(3), 403–425. https://doi.org/10.1007/s11336-013-9350-4

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5–32. https://doi.org/10.1007/978-3-030-62008-0_35

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and Regression Trees*. Taylor and Francis.

Brennan, A. T., Getz, K. D., Brooks, D. R., & Fox, M. P. (2021). An underappreciated misclassification mechanism: Implications of nondifferential dependent misclassification of covariate and exposure. *Annals of Epidemiology*, *58*, 104–123. https://doi.org/10.1016/j.annepidem.2021.02.007

Brooks, J. L., Adams, L. B., Woods-Giscombé, C. L., Currin, E. G., & Corbie-Smith, G. M. (2022). Factor analysis of the Center for Epidemiological Studies Depression Scale in American Indian women. *Research in Nursing & Health*, *45*(6), 733–741. https://doi.org/10.1002/nur.22267

Brown, A., & Croudace, T. J. (Eds.). (2015). Scoring and estimating score precision using multidimensional IRT. In *Handbook of Item Response Theory Modeling: Applications of Typical Performance Assessment* (pp. 307–333). Routledge, Taylor & Francis Group.

Bulut, O., & Suh, Y. (2017). Detecting Multidimensional Differential Item Functioning with the Multiple Indicators Multiple Causes Model, the Item Response Theory Likelihood Ratio Test, and Logistic Regression. *Frontiers in Education*, *2*. https://doi.org/10.3389/feduc.2017.00051

Buss, J. F., Watts, A. L., & Lorenzo-Luaces, L. (2023). Methods for quantifying the heterogeneity of psychopathology. *BMC Psychiatry*, *23*, 897. https://doi.org/10.1186/s12888-023-05377-5

Byllesby, B. M., & Palmieri, P. A. (2023). A Bifactor Model of General and Specific PTSD Symptom Change During Treatment. *Assessment*, *30*(8), 2595–2604. https://doi.org/10.1177/10731911231156646

Cai, L. (2010a). A Two-Tier Full-Information Item Factor Analysis Model with Applications. *Psychometrika*, *75*(4), 581–612. https://doi.org/10.1007/s11336-010-9178-0

Cai, L. (2010b). High-Dimensional Exploratory Item Factor Analysis by A Metropolis–Hastings Robbins–Monro Algorithm. *Psychometrika*, *75*(1), 33–57. https://doi.org/10.1007/s11336-009-9136-x

Cai, L. (2010c). Metropolis-Hastings Robbins-Monro Algorithm for Confirmatory Item Factor Analysis. *Journal of Educational and Behavioral Statistics*, *35*(3), 307–335. https://doi.org/10.3102/1076998609353115

Cai, L., & Thissen, D. (2014). Modern Approaches to Parameter Estimation in Item Response Theory. In *Handbook of Item Response Theory Modeling*. Routledge.

Camilli, G., & Shepard, L. A. (1994). *Methods for Identifying Biased Test Items*. Sage.

Carvalho, L. F., Costa, A. R. L., Otoni, F., & Junqueira, P. (2019). Obsessive–compulsive personality disorder screening cut-off for the Conscientiousness dimension of the

Dimensional Clinical Personality Inventory 2. *The European Journal of Psychiatry*, *33*(3), 112–119. https://doi.org/10.1016/j.ejpsy.2019.05.002

Caspi, A., Houts, R. M., Belsky, D. W., Goldman-Mellor, S. J., Harrington, H., Israel, S., Meier, M. H., Ramrakha, S., Shalev, I., Poulton, R., & Moffitt, T. E. (2014). The p Factor: One General Psychopathology Factor in the Structure of Psychiatric Disorders? *Clinical Psychological Science : A Journal of the Association for Psychological Science*, *2*(2), 119–137. https://doi.org/10.1177/2167702613497473

Cattell, R. B. (1966). The Scree Test For The Number Of Factors. *Multivariate Behavioral Research*, *1*(2), 245–276. https://doi.org/10.1207/s15327906mbr0102_10

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. (2020). Partially and Fully Noncompensatory Response Models for Dichotomous and Polytomous Items. *Applied Psychological Measurement*, *44*(6), 415–430. https://doi.org/10.1177/0146621620909898

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

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Chen, Y., Li, X., Liu, J., & Ying, Z. (2025). Item Response Theory—A Statistical Framework for Educational and Psychological Measurement. *Statistical Science*, *40*(2), 167–194. https://doi.org/10.1214/23-STS896

Chen, Y., Li, X., & Zhang, S. (2019). Joint Maximum Likelihood Estimation for High-Dimensional Exploratory Item Factor Analysis. *Psychometrika*, *84*(1), 124–146. https://doi.org/10.1007/s11336-018-9646-5

Cheng, Y., & Yuan, K.-H. (2010). The Impact of Fallible Item Parameter Estimates on Latent Trait Recovery. *Psychometrika*, *75*(2), 280–291. https://doi.org/10.1007/s11336-009-9144-x

Chmielewski, M., & Watson, D. (2008). The heterogeneous structure of schizotypal personality disorder: Item-level factors of the schizotypal personality questionnaire and their associations with obsessive-compulsive disorder symptoms, dissociative tendencies, and normal personality. *Journal of Abnormal Psychology*, *117*(2), 364–376. https://doi.org/10.1037/0021-843X.117.2.364

Cho, S., Suh, Y., & Lee, W. (2016). An NCME instructional module on latent DIF analysis using mixture item response models. *Educational Measurement: Issues and Practice*, *35*(1), 48–61. https://doi.org/10.1111/emip.12093

Choi, Y.-J., & Cohen, A. S. (2019). Comparison of Scale Identification Methods in Mixture IRT Models. *Journal of Modern Applied Statistical Methods*, *18*. https://doi.org/10.22237/jmasm/1556669700

Chyou, P.-H. (2007). Patterns of bias due to differential misclassification by case-control status in a case-control study. *European Journal of Epidemiology*, *22*(1), 7–17. https://doi.org/10.1007/s10654-006-9078-x

Classe, F., & Kern, C. (2024). Detecting Differential Item Functioning in Multidimensional Graded Response Models With Recursive Partitioning. *Applied Psychological Measurement*, *48*(3), 83–103. https://doi.org/10.1177/01466216241238743

Cohen, J. (2009). *Statistical power analysis for the behavioral sciences* (2. ed., reprint). Psychology Press.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*(1), 21–27. https://doi.org/10.1109/TIT.1967.1053964

Crenshaw, K. (1991). Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color. *Stanford Law Review*, *43*(6), 1241–1299. https://doi.org/10.2307/1229039

Dahlgren, C. L., Bang, L., & Degobi, E. B. (2025). Screening for eating disorders in adolescents: Psychometric evaluation of the eating disorder examination questionnaire short version (EDE-QS) in a community sample. *BMC Psychology*, *13*(1), 1042. https://doi.org/10.1186/s40359-025-03400-w

Dawson, D. A., Saha, T. D., & Grant, B. F. (2010). A Multidimensional Assessment of the Validity and Utility of Alcohol Use Disorder Severity as Determined by Item Response Theory Models. *Drug and Alcohol Dependence*, *107*(1), 31–38. https://doi.org/10.1016/j.drugalcdep.2009.08.019

De Ayala, R. J., & Santiago, S. Y. (2017). An introduction to mixture item response theory models. *Journal of School Psychology, Current Topics in Methodology*, *60*, 25–40. https://doi.org/10.1016/j.jsp.2016.01.002

de Lauzon, B., Romon, M., Deschamps, V., Lafay, L., Borys, J.-M., Karlsson, J., Ducimetière, P., Charles, M. A., & Fleurbaix Laventie Ville Sante Study Group. (2004). The Three-

Factor Eating Questionnaire-R18 is able to distinguish among different eating patterns in a general population. *The Journal of Nutrition*, *134*(9), 2372–2380. https://doi.org/10.1093/jn/134.9.2372

De Ron, J., Fried, E. I., & Epskamp, S. (2021). Psychological networks in clinical populations: Investigating the consequences of Berkson's bias. *Psychological Medicine*, *51*(1), 168–176. https://doi.org/10.1017/S0033291719003209

de Ron, J., Robinaugh, D. J., Fried, E. I., Pedrelli, P., Jain, F. A., Mischoulon, D., & Epskamp, S. (2022). Quantifying and addressing the impact of measurement error in network models. *Behaviour Research and Therapy*, *157*, 104163. https://doi.org/10.1016/j.brat.2022.104163

DeMars, C. E., & Lau, A. (2011). Differential Item Functioning Detection With Latent Classes: How Accurately Can We Detect Who Is Responding Differentially? *Educational and Psychological Measurement*, *71*(4), 597–616. https://doi.org/10.1177/0013164411404221

DeVellis, R. F. (2012). *Scale Development: Theory and Applications*. SAGE Publications.

Devlieger, I., Mayer, A., & Rosseel, Y. (2016). Hypothesis Testing Using Factor Score Regression: A Comparison of Four Methods. *Educational and Psychological Measurement*, *76*(5), 741–770. https://doi.org/10.1177/0013164415607618

Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. *arXiv: Machine Learning*. https://www.semanticscholar.org/paper/Towards-A-Rigorous-Science-of-Interpretable-Machine-Doshi-Velez-Kim/5c39e37022661f81f79e481240ed9b175dec6513

Douissa, M. R., & Jabeur, K. (2020). A non-compensatory classification approach for multi-criteria ABC analysis. *Soft Computing*, *24*(13), 9525–9556. https://doi.org/10.1007/s00500-019-04462-w

Droubay, B. A., Lefmann, D. D., Zhang, S., Snow, J. L., & Hoy-Ellis, C. P. (2025). A new avenue to pathologize sexual minority populations? Internalized sexual stigma and endorsement of compulsive sexual behavior disorder. *Sexual and Relationship Therapy*, *40*(4), 773–793. https://doi.org/10.1080/14681994.2025.2507799

Edelen, M. O., & Reeve, B. B. (2007). Applying Item Response Theory (IRT) Modeling to Questionnaire Development, Evaluation, and Refinement. *Quality of Life Research*, *16*, 5. https://doi.org/10.1007/s11136-007-9198-0

Edwards, M. C. (2009). An Introduction to Item Response Theory Using the Need for Cognition Scale. *Social and Personality Psychology Compass*, *3*(4), 507–529. https://doi.org/10.1111/j.1751-9004.2009.00194.x

Edwards, M. C., & Edelen, M. O. (2009). Special topics in item response theory. In *The Sage handbook of quantitative methods in psychology* (pp. 178–198). Sage Publications Ltd. https://doi.org/10.4135/9780857020994.n8

Efron, B. (1983). Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation. *Journal of the American Statistical Association*, *78*(382), 316–331. https://doi.org/10.1080/01621459.1983.10477973

Efron, B., & Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. Chapman and Hall/CRC. https://doi.org/10.1201/9780429246593

Egleston, B. L., Miller, S. M., & Meropol, N. J. (2011). The impact of misclassification due to survey response fatigue on estimation and identifiability of treatment effects. *Statistics in Medicine*, *30*(30), 3560–3572. https://doi.org/10.1002/sim.4377

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment*, *8*(4), 341–349. https://doi.org/10.1037/1040-3590.8.4.341

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists* (pp. xi, 371). Lawrence Erlbaum Associates Publishers.

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2007). On the Consistency of Individual Classification using Short Scales. *Psychological Methods*, *12*(1), 105–120. https://doi.org/10.1037/1082-989X.12.1.105

Epistola, J. J. (2022). *Comparing the Validity & Fairness of Machine Learning to Regression in Personnel Selection* [Ph.D., University of Maryland, College Park]. https://www.proquest.com/pqdtglobal/docview/2688580521/abstract/3BFCBD6F717D42BDPQ/23?sourcetype=Dissertations%20&%20Theses

Fallon, L. M., Cathcart, S. C., & Johnson, A. H. (2021). Assessing Differential Item Functioning in a Teacher Self-Assessment of Cultural Responsiveness. *Journal of Psychoeducational Assessment*, *39*(7), 816–831. https://doi.org/10.1177/07342829211026464

Fernandez-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *Journal of Machine Learning Research*, *15*(1), 3133–3181.

Ferrando, P. J., & Lorenzo-Seva, U. (2021). The Appropriateness of Sum Scores as Estimates of Factor Scores in the Multiple Factor Analysis of Ordered-Categorical Responses.

Educational and Psychological Measurement, *81*(2), 205–228. https://doi.org/10.1177/0013164420938108

Fife, D. A., & D'Onofrio, J. (2023). Common, uncommon, and novel applications of random forest in psychological research. *Behavior Research Methods*, *55*(5), 2447–2466. https://doi.org/10.3758/s13428-022-01901-9

Fink, A., König, C., & Frey, A. (2025). Accounting for item calibration error in computerized adaptive testing. *Behavior Research Methods*, *57*(5), 126. https://doi.org/10.3758/s13428-025-02649-8

Flake, J. K. (2021). Strengthening the foundation of educational psychology by integrating construct validation into open science reform. *Educational Psychologist*, *56*(2), 132–141. https://doi.org/10.1080/00461520.2021.1898962

Flake, J. K., Davidson, I. J., Wong, O., & Pek, J. (2022). Construct validity and the validity of replication studies: A systematic review. *American Psychologist*, *77*(4), 576–588. https://doi.org/10.1037/amp0001006

Flake, J. K., & Fried, E. I. (2020). Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465. https://doi.org/10.1177/2515245920952393

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

Frick, H., Strobl, C., Leisch, F., & Zeileis, A. (2012). Flexible Rasch Mixture Models with Package psychomix. *Journal of Statistical Software*, *48*, 1–25. https://doi.org/10.18637/jss.v048.i07

Frick, H., Strobl, C., & Zeileis, A. (2015). Rasch Mixture Models for DIF Detection: A Comparison of Old and New Score Specifications. *Educational and Psychological Measurement*, *75*(2), 208–234. https://doi.org/10.1177/0013164414536183

Fried, E. I., & Nesse, R. M. (2015). Depression sum-scores don't add up: Why analyzing specific depression symptoms is essential. *BMC Medicine*, *13*(1), 72. https://doi.org/10.1186/s12916-015-0325-4

Furr, R. M. (2020). Psychometrics in Clinical Psychological Research. In A. G. C. Wright & M. N. Hallquist (Eds.), *The Cambridge Handbook of Research Methods in Clinical Psychology* (1st ed., pp. 54–65). Cambridge University Press. https://doi.org/10.1017/9781316995808.008

Gallop, M., & Weschle, S. (2019). Assessing the Impact of Non-Random Measurement Error on Inference: A Sensitivity Analysis Approach. *Political Science Research and Methods*, *7*(2), 367–384. https://doi.org/10.1017/psrm.2016.53

Giannouli, V., & Kampakis, S. (2024). Can Machine Learning Assist Us in the Classification of Older Patients Suffering from Dementia Based on Classic Neuropsychological Tests and a New Financial Capacity Test Performance? *Journal of Neuropsychology*. https://doi.org/10.1111/jnp.12409

Gibbons, R. D., Weiss, D. J., Frank, E., & Kupfer, D. (2016). Computerized Adaptive Diagnosis and Testing of Mental Health Disorders. *Annual Review of Clinical Psychology*, *12*, 83–104. https://doi.org/10.1146/annurev-clinpsy-021815-093634

Gilbert, J. B. (2025). How Measurement Affects Causal Inference: Attenuation Bias Is (Usually) More Important Than Outcome Scoring Weights. *Methodology*, *21*(2), 91–122. https://doi.org/10.5964/meth.15773

Gonzalez, O. (2021). Psychometric and Machine Learning Approaches for Diagnostic Assessment and Tests of Individual Classification. *Psychological Methods*, *26*(2), 236–254. https://doi.org/10.1037/met0000317

Gonzalez, O. (2025). Combining psychometric and machine learning approaches to select items and score responses. *Behaviormetrika*, *52*(2), 259–292. https://doi.org/10.1007/s41237-025-00257-6

Gonzalez, O., & Georgeson, A. R. (2020). Developing tools to assess fairness and equity in screening measures. *The Score*. https://www.apadivisions.org/division-5/publications/score/2020/01/screening-for-fairness

Gonzalez, O., Georgeson, A. R., & Pelham, W. E. (2024). Estimating Classification Consistency of Machine Learning Models for Screening Measures. *Psychological Assessment*, *36*(6–7), 395–406. https://doi.org/10.1037/pas0001313

Gonzalez, O., Georgeson, A. R., Pelham, W. E., & Fouladi, R. T. (2021). Estimating Classification Consistency of Screening Measures and Quantifying the Impact of Measurement Bias. *Psychological Assessment*, *33*(7), 596–609. https://doi.org/10.1037/pas0000938

Gonzalez, O., O'Rourke, H. P., Wurpts, I. C., & Grimm, K. J. (2018). Analyzing Monte Carlo Simulation Studies With Classification and Regression Trees. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(3), 403–413. https://doi.org/10.1080/10705511.2017.1369353

Gonzalez, O., & Pelham III, W. E. (2021). When Does Differential Item Functioning Matter for Screening? A Method for Empirical Evaluation. *Assessment*, *28*(2), 446–456. https://doi.org/10.1177/1073191120913618

Goretzko, D., & Bühner, M. (2022). Robustness of factor solutions in exploratory factor analysis. *Behaviormetrika*, *49*(1), 131–148. https://doi.org/10.1007/s41237-021-00152-w

Goretzko, D., Partsch, M. V., & Sterner, P. (2025). Embrace the heterogeneity in exploratory factor analysis but be transparent about what you do—A commentary on Manapat et al. (2023). *Psychological Methods*. https://doi.org/10.1037/met0000759

Graham, J. M. (2006). Congeneric and (Essentially) Tau-Equivalent Estimates of Score Reliability: What They Are and How to Use Them. *Educational and Psychological Measurement*, *66*(6), 930–944. https://doi.org/10.1177/0013164406288165

Grimm, K. J., Jacobucci, R., & McArdle, J. J. (2023). Decision trees and ensemble methods in the behavioral sciences. In H. Cooper, M. N. Coutanche, L. M. McMullen, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology: Data analysis and research publication (Vol. 3) (2nd ed.).* (pp. 429–448). American Psychological Association. https://doi.org/10.1037/0000320-019

Groen, M. M. van. (2014). *Adaptive Testing for Making Unidimensional and Multidimensional Classification Decisions* [Dissertation, University of Twente]. https://research.utwente.nl/en/publications/adaptive-testing-for-making-unidimensional-and-multidimensional-c

Hailu Gebrie, M. (2018). An Analysis of Beck Depression Inventory 2nd Edition (BDI-II). *Global Journal of Endocrinological Metabolism*, *2*(3). https://doi.org/10.31031/GJEM.2018.02.000540

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications, 73*, 220–239. https://doi.org/10.1016/j.eswa.2016.12.035

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (2011). *Fundamentals of item response theory* (Nachdr.). Sage.

Han, I., Chandler, J. S., & Liang, T.-P. (1996). The Impact of Measurement Scale and Correlation Structure on Classification Performance of Inductive Learning and Statistical Methods. *Expert Systems with Applications*, *10*(2), 209–221. https://doi.org/10.1016/0957-4174(95)00047-X

Han, K. T., & Paek, I. (2014). A Review of Commercial Software Packages for Multidimensional IRT Modeling. *Applied Psychological Measurement*, *38*(6), 486–498. https://doi.org/10.1177/0146621614536770

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29–36. https://doi.org/10.1148/radiology.143.1.7063747

Haroz, E. E., Bolton, P., Gross, A., Chan, K. S., Michalopoulos, L., & Bass, J. (2016). Depression symptoms across cultures: An IRT analysis of standard depression symptoms using data from eight countries. *Social Psychiatry and Psychiatric Epidemiology*, *51*(7), 981–991. https://doi.org/10.1007/s00127-016-1218-3

Hasan, R., & Chu, C. (2022). Noise in Datasets: What Are the Impacts on Classification Performance? [Noise in Datasets: What Are the Impacts on Classification Performance?]. *Proceedings of the 11th International Conference on Pattern Recognition Applications and Methods*. https://doi.org/10.5220/0010782200003122

Haslam, N., McGrath, M. J., Viechtbauer, W., & Kuppens, P. (2020). Dimensions over categories: A meta-analysis of taxometric research. *Psychological Medicine*, *50*(9), 1418–1432. https://doi.org/10.1017/S003329172000183X

Haslbeck, J. M. B., Ryan, O., & Dablander, F. (2022). The Sum of All Fears: Comparing Networks Based on Symptom Sum-Scores. *Psychological Methods, 27*(6), 1061–1068. https://doi.org/10.1037/met0000418

Hassanzad, M., & Hajian-Tilaki, K. (2024). Methods of determining optimal cut-point of diagnostic biomarkers with application of clinical data in ROC analysis: An update review. *BMC Medical Research Methodology, 24*(1), 84. https://doi.org/10.1186/s12874-024-02198-2

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer New York. https://doi.org/10.1007/978-0-387-84858-7

Henry, J. D., & Crawford, J. R. (2005). The short-form version of the Depression Anxiety Stress Scales (DASS-21): Construct validity and normative data in a large non-clinical sample. *The British Journal of Clinical Psychology*, *44*(Pt 2), 227–239. https://doi.org/10.1348/014466505X29657

Hill, C. D. (2004). *Precision of parameter estimates for the graded item response model [Unpublished master's thesis]* [. The University of North Carolina at Chapel Hill]. https:// catalog.lib.unc.edu/catalog/UNCb4556245

Höfler, M. (2005). The effect of misclassification on the estimation of association: A review. *International Journal of Methods in Psychiatric Research, 14*(2), 92–101. https://doi.org/10.1002/mpr.20

Holden, J. E., Finch, W. H., & Kelley, K. (2011). A Comparison of Two-Group Classification Methods. *Educational and Psychological Measurement, 71*(5), 870–901. https://doi.org/10.1177/0013164411398357

Holsclaw, T., Hallgren, K. A., Steyvers, M., Smyth, P., & Atkins, D. C. (2015). Measurement error and outcome distributions: Methodological issues in regression analyses of behavioral coding data. *Psychology of Addictive Behaviors : Journal of the Society of Psychologists in Addictive Behaviors*, *29*(4), 1031–1040. https://doi.org/10.1037/adb0000091

Holzinger, K. J., & Swineford, F. (1937). The Bi-factor method. *Psychometrika*, *2*(1), 41–54. https://doi.org/10.1007/BF02287965

Hoover, J. C. (2022). *Using Machine Learning to Identify Causes of Differential Item Functioning* [Ph.D., University of Kansas]. https://www.proquest.com/docview/2682816398/abstract/368EB15307B444F9PQ/1?sourcetype=Dissertations%20&%20Theses

Hoshino, T., & Bentler, P. (2013). Bias in factor score regression and a simple solution. In A. De Leon & K. Chough (Eds.), *Analysis of Mixed Data* (pp. 43–61). Chapman and Hall/CRC. https://doi.org/10.1201/b14571-5

Huang, H.-Y., Chen, P.-H., & Wang, W.-C. (2012). Computerized Adaptive Testing Using a Class of High-Order Item Response Theory Models. *Applied Psychological Measurement*, *36*(8), 689–706. https://doi.org/10.1177/0146621612459552

Hussey, I., & Hughes, S. (2020). Hidden Invalidity Among 15 Commonly Used Measures in Social and Personality Psychology. *Advances in Methods and Practices in Psychological Science*, *3*(2), 166–184. https://doi.org/10.1177/2515245919882903

Imai, K., & Yamamoto, T. (2010). Causal Inference with Differential Measurement Error: Nonparametric Identification and Sensitivity Analysis. *American Journal of Political Science*, *54*(2), 543–560. https://doi.org/10.1111/j.1540-5907.2010.00446.x

Jacobsen, G., Casanova, M. P., Dluzniewski, A., Reeves, A. J., & Baker, R. T. (2024). Examining the Factor Structure and Validity of the Depression Anxiety Stress Scale-21. *European Journal of Investigation in Health, Psychology and Education*, *14*(11), 2932–2943. https://doi.org/10.3390/ejihpe14110192

Jacobucci, R., & Grimm, K. J. (2020). Machine Learning and Psychological Research: The Unexplored Effect of Measurement. *Perspectives on Psychological Science*, *15*(3), 809–816. https://doi.org/10.1177/1745691620902467

Jacobucci, R., Grimm, K. J., & Zhang, Z. (2023). *Machine Learning for Social and Behavioral Research*. The Guilford Press.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: With Applications in R*. Springer US. https://doi.org/10.1007/978-1-0716-1418-1

Jank, W. (2005). Quasi-Monte Carlo sampling to improve the efficiency of Monte Carlo EM. *Computational Statistics & Data Analysis*, *48*(4), 685–701. https://doi.org/10.1016/j.csda.2004.03.019

Jennings, T. L., Gleason, N., & Kraus, S. W. (2022). Assessment of compulsive sexual behavior disorder among lesbian, gay, bisexual, transgender, and queer clients •. *Journal of Behavioral Addictions*, *11*(2), 216–221. https://doi.org/10.1556/2006.2022.00028

Jiang, S., Wang, C., & Weiss, D. J. (2016). Sample Size Requirements for Estimation of Item Parameters in the Multidimensional Graded Response Model. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00109

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating Type I Error and Power Rates Using an Effect Size Measure With the Logistic Regression Procedure for DIF Detection. *Applied*

Measurement in Education, *14*(4), 329–349. https://doi.org/10.1207/S15324818AME1404_2

Jones, C. R. G., Livingston, L. A., Fretwell, C., Uljarević, M., Carrington, S. J., Shah, P., & Leekam, S. R. (2024). Measuring self and informant perspectives of Restricted and Repetitive Behaviours (RRBs): Psychometric evaluation of the Repetitive Behaviours Questionnaire-3 (RBQ-3) in adult clinical practice and research settings. *Molecular Autism*, *15*(1), 24. https://doi.org/10.1186/s13229-024-00603-7

Jones, H. E., Gatsonsis, C. A., Trikalinos, T. A., Welton, N. J., & Ades, A. e. (2019). Quantifying how diagnostic test accuracy depends on threshold in a meta-analysis. *Statistics in Medicine*, *38*(24), 4789–4803. https://doi.org/10.1002/sim.8301

Junker, B. W., & Sijtsma, K. (2001). Cognitive Assessment Models with Few Assumptions, and Connections with Nonparametric Item Response Theory. *Applied Psychological Measurement*, *25*(3), 258–272. https://doi.org/10.1177/01466210122032064

Kalibatseva, Z., Leong, F. T. L., & Ham, E. H. (2014). A Symptom Profile of Depression among Asian Americans: Is There Evidence for Differential Item Functioning of Depressive Symptoms? *Psychological Medicine*, *44*, 2567–2578. https://doi.org/10.1017/S0033291714000130

Karadavut, T. (2021). Characterizing the Latent Classes in a Mixture IRT Model Using DIF. *Applied Measurement in Education*, *34*(4), 301–311. https://doi.org/10.1080/08957347.2021.1987900

Karukivi, M., Vahlberg, T., Horjamo, K., Nevalainen, M., & Korkeila, J. (2017). Clinical importance of personality difficulties: Diagnostically sub-threshold personality disorders. *BMC Psychiatry*, *17*. https://doi.org/10.1186/s12888-017-1200-y

Kawilapat, S., Maneeton, B., Maneeton, N., Prasitwattanaseree, S., Kongsuk, T., Arunpongpaisal, S., Leejongpermpoon, J., Sukhawaha, S., & Traisathit, P. (2022). Comparison of unweighted and item response theory-based weighted sum scoring for the Nine-Questions Depression-Rating Scale in the Northern Thai Dialect. *BMC Medical Research Methodology*, *22*, 268. https://doi.org/10.1186/s12874-022-01744-0

Ketchen, D. J., & Shook, C. L. (1996). The Application of Cluster Analysis in Strategic Management Research: An Analysis and Critique. *Strategic Management Journal*, *17*(6), 441–458. https://doi.org/10.1002/(SICI)1097-0266(199606)17:6%3C441::AID-SMJ819%3E3.0.CO;2-G

Kim, J., Kim, H., Kang, M., Oh, H., & Jo, M. (2024). Development and Psychometric Evaluation of the End-of-Life Nursing Competency Scale for Clinical Nurses. *Healthcare*, *12*(16), 1580. https://doi.org/10.3390/healthcare12161580

Kim, J., & Oshima, T. C. (2013). Effect of Multiple Testing Adjustment in Differential Item Functioning Detection. *Educational and Psychological Measurement*, *73*(3), 458–470. https://doi.org/10.1177/0013164412467033

Kohrt, B. A., Gurung, D., Singh, R., Rai, S., Neupane, M., Turner, E. L., Platt, A., Sun, S., Gautam, K., Luitel, N. P., & Jordans, M. J. D. (2025). Is there a mental health diagnostic crisis in primary care? Current research practices in global mental health cannot answer that question. *Epidemiology and Psychiatric Sciences*, *34*, e7. https://doi.org/10.1017/S2045796025000010

Krantz, C. (2022). *Modeling the outcomes of a longitudinal tie-breaker regression discontinuity design to assess an in-home training program for families at risk of child abuse and neglect*. https://hdl.handle.net/11244/335699

Kraus, E. B., Wild, J., & Hilbert, S. (2022). *Using Interpretable Machine Learning for Local Unfairness Detection in Psychometric Tests*. https://doi.org/10.31234/osf.io/mkq9r

Kraus, E. B., Wild, J., & Hilbert, S. (2024). Using Interpretable Machine Learning for Differential Item Functioning Detection in Psychometric Tests. *Applied Psychological Measurement*, *48*(4–5), 167–186. https://doi.org/10.1177/01466216241238744

Kroc, E., & Olvera Astivia, O. L. (2022). The Importance of Thinking Multivariately When Setting Subscale Cutoff Scores. *Educational and Psychological Measurement*, *82*(3), 517–538. https://doi.org/10.1177/00131644211023569

Krogsgaard, M. R., Brodersen, J., Christensen, K. B., Siersma, V., Jensen, J., Hansen, C. F., Engebretsen, L., Visnes, H., Forssblad, M., & Comins, J. D. (2021). How to translate and locally adapt a PROM. Assessment of cross-cultural differential item functioning. *Scandinavian Journal of Medicine & Science in Sports*, *31*(5), 999–1008. https://doi.org/10.1111/sms.13854

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, *28*(5), 1–26. https://doi.org/10.18637/jss.v028.i05

Kutscher, T., Eid, M., & Crayen, C. (2019). Sample Size Requirements for Applying Mixed Polytomous Item Response Models: Results of a Monte Carlo Simulation Study. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.02494

Kyriazos, T. A., & Stalikas, A. (2018). Applied Psychometrics: The Steps of Scale Development and Standardization Process. *Psychology*, *9*(11), 2531–2560. https://doi.org/10.4236/psych.2018.911145

Lahey, B. B., Applegate, B., Hakes, J. K., Zald, D. H., Hariri, A. R., & Rathouz, P. J. (2012). Is there a general factor of prevalent psychopathology during adulthood? *Journal of Abnormal Psychology*, *121*(4), 971–977. https://doi.org/10.1037/a0028355

LaHuis, D. M., Blackmore, C. E., & Ammons, G. M. (2025). Comparing Approaches to Estimating Person Parameters for the MUPP Model. *Applied Psychological Measurement*, *49*(4–5), 163–176. https://doi.org/10.1177/01466216251316278

Lai, M. H. C., Kwok, O., Yoon, M., & Hsiao, Y.-Y. (2017). Understanding the Impact of Partial Factorial Invariance on Selection Accuracy: An R Script. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(5), 783–799. https://doi.org/10.1080/10705511.2017.1318703

Lai, M. H. C., Richardson, G. B., & Mak, H. W. (2019). Quantifying the impact of partial measurement invariance in diagnostic research: An application to addiction research. *Addictive Behaviors, Improving the Implementation of Quantitative Methods in Addiction Research*, *94*, 50–56. https://doi.org/10.1016/j.addbeh.2018.11.029

Lai, M. H. C., & Tse, W. W.-Y. (2024). Are Factor Scores Measurement Invariant? *Psychological Methods*. https://doi.org/10.1037/met0000658

Lai, M. H. C., & Zhang, Y. (2022). Classification Accuracy of Multidimensional Tests: Quantifying the Impact of Noninvariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*(4), 620–629. https://doi.org/10.1080/10705511.2021.1977936

Lavigne, J. V., Feldman, M., & Meyers, K. M. (2016). Screening for Mental Health Problems: Addressing the Base Rate Fallacy for a Sustainable Screening Program in Integrated

Primary Care. *Journal of Pediatric Psychology*, *41*(10), 1081–1090. https://doi.org/10.1093/jpepsy/jsw048

LeDell, E., Petersen, M., & van der Laan, M. (2015). Computationally efficient confidence intervals for cross-validated area under the ROC curve estimates. *Electronic Journal of Statistics*, *9*(1), 1583–1607. https://doi.org/10.1214/15-EJS1035

Lee, J. K., Collins, B., Pepper, E., Alvarez, N. A., & Warholak, T. (2023). Improving a Leadership Scale: Applying Rasch Analysis to Student Pharmacists' Attitudes and Beliefs About Leadership. *American Journal of Pharmaceutical Education*, *87*(6), 100063. https://doi.org/10.1016/j.ajpe.2023.100063

Lee, S., Bulut, O., & Suh, Y. (2017). Multidimensional Extension of Multiple Indicators Multiple Causes Models to Detect DIF. *Educational and Psychological Measurement*, *77*(4), 545–569. https://doi.org/10.1177/0013164416651116

Liaw, A., & Wiener, M. (2022). *Classification and Regression by randomForest.* (Version R package version 4.7-1.2) [Computer software]. https://CRAN.R-project.org/doc/Rnews/

Linde, K., Olm, M., Teusen, C., Akturk, Z., Schrottenberg, V., Hapfelmeier, A., Dawson, S., Rücker, G., Löwe, B., & Schneider, A. (2022). The diagnostic accuracy of widely used self-report questionnaires for detecting anxiety disorders in adults. *The Cochrane Database of Systematic Reviews*, *2022*(9), CD015292. https://doi.org/10.1002/14651858.CD015292

Liu, R. (2018). Misspecification of Attribute Structure in Diagnostic Measurement. *Educational and Psychological Measurement*, *78*(4), 605–634. https://doi.org/10.1177/0013164417702458

Liu, X. (2012). Classification Accuracy and Cut Point Selection. *Statistics in Medicine*, *31*(23), 2676–2686. https://doi.org/10.1002/sim.4509

Liu, Y., Magnus, B., O'Connor, H., & Thissen, D. (2018). Multidimensional Item Response Theory. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The Wiley Handbook of Psychometric Testing* (1st ed., pp. 445–493). Wiley. https://doi.org/10.1002/9781118489772.ch16

Lo, B. C. Y., Zhao, Y., Kwok, A. W. Y., Chan, W., & Chan, C. K. Y. (2017). Evaluation of the Psychometric Properties of the Asian Adolescent Depression Scale and Construction of a Short Form: An Item Response Theory Analysis. *Assessment*, *24*(5), 660–676. https://doi.org/10.1177/1073191115614393

Lord, F. M. (2012). *Applications of Item Response Theory To Practical Testing Problems*. Taylor and Francis.

Lord, F. M., & Novick, M. R. (2008). *Statistical Theories of Mental Test Scores*. IAP.

Lundberg, S., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions* (arXiv:1705.07874). arXiv. https://doi.org/10.48550/arXiv.1705.07874

Lundström, S., Taylor, M., Larsson, H., Lichtenstein, P., Kuja-Halkola, R., & Gillberg, C. (2022). Perceived child impairment and the "autism epidemic." *Journal of Child Psychology and Psychiatry*, *63*(5), 591–598. https://doi.org/10.1111/jcpp.13497

Luong, R., & Flake, J. K. (2023). Measurement invariance testing using confirmatory factor analysis and alignment optimization: A tutorial for transparent analysis planning and reporting. *Psychological Methods*, *28*(4), 905–924. https://doi.org/10.1037/met0000441

Mahalanobis, P. C. (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India*, *2*(1), 49–55.

Manapat, P. D., Anderson, S. F., & Edwards, M. C. (2025). Evaluating avoidable heterogeneity in exploratory factor analysis results. *Psychological Methods*, *30*(3). https://doi.org/10.1037/met0000589

Manapat, P. D., & Edwards, M. C. (2022). Examining the Robustness of the Graded Response and 2-Parameter Logistic Models to Violations of Construct Normality. *Educational and Psychological Measurement*, *82*(5), 967–988. https://doi.org/10.1177/00131644211063453

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.

McNeish, D. (2022). Psychometric properties of sum scores and factor scores differ even when their correlation is 0.98: A response to Widaman and Revelle. *Behavior Research Methods*, *55*(8), 4269–4290. https://doi.org/10.3758/s13428-022-02016-x

McNeish, D. (2024). Practical Implications of Sum Scores Being Psychometrics' Greatest Accomplishment. *Psychometrika*, *89*(4), 1148–1169. https://doi.org/10.1007/s11336-024-09988-z

McNeish, D. (2026). How Do Psychologists Determine Whether a Measurement Scale Is Good? A Quarter-Century of Scale Validation with Hu & Bentler (1999). *Annual Review of Psychology*, *77*(Volume 77, 2026), 567–591. https://doi.org/10.1146/annurev-psych-121924-104021

McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods*, *52*(6), 2287–2305. https://doi.org/10.3758/s13428-020-01398-0

Messick, S. (1995). Validity of Psychological Assessment. *American Psychologist*, *50*(9), 741–749.

Meyers, J. L. (2018). Scoring models in competency-based educational assessment. *The Journal of Competency-Based Education*, *3*(3), e01173. https://doi.org/10.1002/cbe2.1173

Monaghan, T. F., Rahman, S. N., Agudelo, C. W., Wein, A. J., Lazar, J. M., Everaert, K., & Dmochowski, R. R. (2021). Foundational Statistical Principles in Medical Research: Sensitivity, Specificity, Positive Predictive Value, and Negative Predictive Value. *Medicina*, *57*(5), 503. https://doi.org/10.3390/medicina57050503

Mongelli, F., Georgakopoulos, P., & Pato, M. T. (2020). Challenges and Opportunities to Meet the Mental Health Needs of Underserved and Disenfranchised Populations in the United States. *Focus: Journal of Life Long Learning in Psychiatry*, *18*(1), 16–24. https://doi.org/10.1176/appi.focus.20190028

Morgado, E., Martino, L., & Millan-Castillo, R. S. (2023). Universal and Automatic Elbow Detection for Learning the Effective Number of Components in Model Selection Problems. *Digital Signal Processing*, *140*, 104103. https://doi.org/10.1016/j.dsp.2023.104103

Navarro-González, M. C., Padilla, J.-L., & Benítez, I. (2024). Analyzing Measurement Invariance for Studying the Gender Gap in Educational Testing: A Mixed Studies Systematic Review. *European Journal of Psychological Assessment*, *40*(5), 412–426. https://doi.org/10.1027/1015-5759/a000820

Nielsen, T., Elklit, A., Vang, M. L., Nielsen, S. B., Auning-Hansen, M., & Palic, S. (2023). Cross-cultural validity and psychometric properties of the International Trauma Questionnaire in a clinical refugee sample. *European Journal of Psychotraumatology*, *14*(1), 2172256. https://doi.org/10.1080/20008066.2023.2172256

Nikan, F., Jafarabadi, M. A., Mohammad-Alizadeh-Charandabi, S., & Mirghafourvand, M. (2018). Designation and psychometric properties of the short form postpartum quality of life questionnaire (SF-PQOL): An application of multidimensional item response theory and genetic algorithm. *Health Promotion Perspectives*, *8*(3), 215–224. https://doi.org/10.15171/hpp.2018.29

O'Connor, B. P. (2018). An illustration of the effects of fluctuations in test information on measurement error, the attenuation of effect sizes, and diagnostic reliability. *Psychological Assessment, 30*(8), 991–1003. https://doi.org/10.1037/pas0000471

Ohi, K., Tanaka, Y., Otowa, T., Shimada, M., Kaiya, H., Nishimura, F., Sasaki, T., Tanii, H., Shioiri, T., & Hara, T. (2024). Discrimination between healthy participants and people with panic disorder based on polygenic scores for psychiatric disorders and for intermediate phenotypes using machine learning. *Australian and New Zealand Journal of Psychiatry*, *58*(7), 603. https://doi.org/10.1177/00048674241242936

Onumanyi, A. J., Molokomme, D. N., Isaac, S. J., & Abu-Mahfouz, A. M. (2022). AutoElbow: An Automatic Elbow Detection Method for Estimating the Number of Clusters in a Dataset. *Applied Sciences*, *12*(15), 7515. https://doi.org/10.3390/app12157515

Orlando, M., & Marshall, G. N. (2002). Differential item functioning in a Spanish translation of the PTSD Checklist: Detection and evaluation of impact. *Psychological Assessment*, *14*(1), 50–59. https://doi.org/10.1037/1040-3590.14.1.50

Orrù, G., De Marchi, B., Sartori, G., Gemignani, A., Scarpazza, C., Monaro, M., Mazza, C., & Roma, P. (2022). Machine learning item selection for short scale construction: A proof-of-concept using the SIMS. *The Clinical Neuropsychologist*, *0*(0), 1–18. https://doi.org/10.1080/13854046.2022.2114548

Orrù, G., Gemignani, A., Ciacchini, R., Bazzichi, L., & Conversano, C. (2020). Machine Learning Increases Diagnosticity in Psychometric Evaluation of Alexithymia in Fibromyalgia. *Frontiers in Medicine, 6*. https://doi.org/10.3389/fmed.2019.00319

Ozcan, M., & Lai, M. H. C. (2025). Exploring the Impact of Deleting (or Retaining) a Biased Item: A Procedure Based on Classification Accuracy. *Assessment*, *32*(8), 1211–1225. https://doi.org/10.1177/10731911241298081

Öztürk Özkan, G., Karaman, A., Zafer, B., & Evci, S. A. (2025). A cross-sectional study to evaluate the relationship between emotional eating, social media use and body perception in young adults. *European Review of Applied Psychology*, *75*(3), 101080. https://doi.org/10.1016/j.erap.2025.101080

Pankowska, P., Oberski, D., Garnier-Villarreal, M., & Pavlopoulos, D. (2025). The effect of measurement error on clustering. *Quality & Quantity*, *59*(5), 4825–4860. https://doi.org/10.1007/s11135-025-02177-9

Paulsen, J., Svetina, D., Feng, Y., & Valdivia, M. (2020). Examining the Impact of Differential Item Functioning on Classification Accuracy in Cognitive Diagnostic Models. *Applied Psychological Measurement*, *44*(4), 267–281. https://doi.org/10.1177/0146621619858675

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling*, *8*(2), 287–312. https://doi.org/10.1207/S15328007SEM0802_7

Penfield, R. D., & Camilli, G. (2006). Differential Item Functioning and Item Bias. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics* (Vol. 26, pp. 125–167). Elsevier. https://doi.org/10.1016/S0169-7161(06)26005-X

Quan, Y., & Wang, C. (2025). *Using Multi-label Classification Neural Network to Detect Intersectional DIF with Small Sample Sizes*.

Quan, Y., & Wang, C. (2026). Using multilabel classification neural network to detect intersectional DIF with small sample sizes. *British Journal of Mathematical and Statistical Psychology*, *00*, 1–38. https://doi.org/https://doi.org/10.1111/bmsp.70041

R Core Team. (2024). *R: A Language and Environment for Statistical Computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Ramsay, J., Li, J., & Wiberg, M. (2020). Better Rating Scale Scores with Information–Based Psychometrics. *Psych*, *2*(4), 347–369. https://doi.org/10.3390/psych2040026

Ravand, H., & Baghaei, P. (2019). Diagnostic Classification Models: Recent Developments, Practical Issues, and Prospects. *International Journal of Testing*. https://doi.org/10.1080/15305058.2019.1588278

Raykov, T. (1997a). Estimation of Composite Reliability for Congeneric Measures. *Applied Psychological Measurement*, *21*(2), 173–184. https://doi.org/10.1177/01466216970212006

Raykov, T. (1997b). Scale Reliability, Cronbach's Coefficient Alpha, and Violations of Essential Tau-Equivalence with Fixed Congeneric Components. *Multivariate Behavioral Research*, *32*(4), 329–353. https://doi.org/10.1207/s15327906mbr3204_2

Raykov, T., & Zhang, B. (2025). Sum Scores of Behavioral Measurement Scales: When Not to Recommend Them and How Structural Equation Modeling Can Help. *Structural Equation Modeling: A Multidisciplinary Journal*, *0*(0), 1–9. https://doi.org/10.1080/10705511.2025.2571741

Raymond, K. E., & Butler, B. E. (2025). Measuring Misophonia: Assessing the Psychometric Properties of the MisoQuest and Its Ability to Predict Cognitive Impacts of Triggering Sounds. *Journal of Clinical Psychology*, 1–15. https://doi.org/10.1002/jclp.70033

Reckase, M. D. (2009). Historical Background for Multidimensional Item Response Theory (MIRT). In M. D. Reckase (Ed.), *Multidimensional Item Response Theory* (pp. 57–77). Springer. https://doi.org/10.1007/978-0-387-89976-3_3

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, *16 Suppl 1*, 19–31. https://doi.org/10.1007/s11136-007-9183-7

Rhemtulla, M., & Savalei, V. (2025). Estimated Factor Scores Are Not True Factor Scores. *Multivariate Behavioral Research*, *60*(3), 598–619. https://doi.org/10.1080/00273171.2024.2444943

Rhemtulla, M., van Bork, R., & Borsboom, D. (2020). Worse than measurement error: Consequences of inappropriate latent variable measurement models. *Psychological Methods*, *25*(1), 30–45. https://doi.org/10.1037/met0000220

Rijmen, F. (2010). Formal Relations and an Empirical Comparison among the Bi-Factor, the Testlet, and a Second-Order Multidimensional IRT Model. *Journal of Educational Measurement*, *47*(3), 361–372. https://doi.org/10.1111/j.1745-3984.2010.00118.x

Robbins, H., & Monro, S. (1951). A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, *22*(3), 400–407.

Rome, L., & Zhang, B. (2018). Investigating the Effects of Differential Item Functioning on Proficiency Classification. *Applied Psychological Measurement*, *42*(4), 259–274. https://doi.org/10.1177/0146621617726789

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, *104*(3), 192–233. https://doi.org/10.1037/0096-3445.104.3.192

Rosenthal, M. Z., Anand, D., Cassiello-Robbins, C., Williams, Z. J., Guetta, R. E., Trumbull, J., & Kelley, L. D. (2021). Development and Initial Validation of the Duke Misophonia Questionnaire. *Frontiers in Psychology*, *12*, 709928. https://doi.org/10.3389/fpsyg.2021.709928

Ross, S. M., Haegele, J. A., Anderson, K., & Healy, S. (2023). Evidence of item bias in a national flourishing measure for autistic youth. *Autism Research*, *16*(4), 841–854. https://doi.org/10.1002/aur.2900

Rost, J. (1990). Rasch Models in Latent Classes: An Integration of Two Approaches to Item Analysis. *Applied Psychological Measurement*, *14*(3), 271–282. https://doi.org/10.1177/014662169001400305

Rudin, C. (2019). Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications* (pp. xx, 348). The Guilford Press.

Russell, M., & Kaplan, L. (2021). An Intersectional Approach to Differential Item Functioning: Reflecting Configurations of Inequality. *Practical Assessment, Research, and Evaluation*, *26*(1). https://doi.org/10.7275/20614854

Ryan, E. P., & Oquendo, M. A. (2020). Suicide Risk Assessment and Prevention: Challenges and Opportunities. *Focus: Journal of Life Long Learning in Psychiatry*, *18*(2), 88–99. https://doi.org/10.1176/appi.focus.20200011

Sajobi, T. T., Lix, L. M., Russell, L., Schulz, D., Liu, J., Zumbo, B. D., & Sawatzky, R. (2022). Accuracy of mixture item response theory models for identifying sample heterogeneity in patient-reported outcomes: A simulation study. *Quality of Life Research*, *31*(12), 3423–3432. https://doi.org/10.1007/s11136-022-03169-0

Salloum, S. A., Basiouni, A., Alfaisal, R., Salloum, A., & Shaalan, K. (2024). Predicting Student Retention in Higher Education Using Machine Learning. In A. Basiouni & C. Frasson (Eds.), *Breaking Barriers with Generative Intelligence. Using GI to Improve Human Education and Well-Being* (pp. 197–206). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-65996-6_17

Samejima, F. (1969). Estimation of Latent Ability Using a Response Pattern of Graded Scores. *Psychometrika*, *34*(S1), 1–97. https://doi.org/10.1007/BF03372160

Samejima, F. (1997). Graded Response Model. In W. J. Van Der Linden & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 85–100). Springer New York. https://doi.org/10.1007/978-1-4757-2691-6_5

Satopaa, V., Albrecht, J., Irwin, D., & Raghavan, B. (2011). Finding a “Kneedle” in a Haystack: Detecting Knee Points in System Behavior. *2011 31st International Conference on Distributed Computing Systems Workshops*, 166–171. https://doi.org/10.1109/ICDCSW.2011.20

Schilling, S., & Bock, R. D. (2005). High-Dimensional Maximum Marginal Likelihood Item Factor Analysis by Adaptive Quadrature. *Psychometrika*, *70*(3), 533–555. https://doi.org/10.1007/s11336-003-1141-x

Schiltz, H. K., & Magnus, B. E. (2020). Gender-based differential item functioning on the child behavior checklist in youth on the autism spectrum: A brief report. *Research in Autism Spectrum Disorders*, *79*, 101669. https://doi.org/10.1016/j.rasd.2020.101669

Schiltz, H. K., & Magnus, B. E. (2021). Differential Item Functioning Based on Autism Features, IQ, and Age on the Screen for Child Anxiety Related Disorders (SCARED) Among Youth on the Autism Spectrum. *Autism Research*, *14*(6), 1220–1236. https://doi.org/10.1002/aur.2481

Schröder, A., Vulink, N., & Denys, D. (2013). Misophonia: Diagnostic Criteria for a New Psychiatric Disorder. *PLoS ONE*, *8*(1), e54706. https://doi.org/10.1371/journal.pone.0054706

Schroeders, U., & Gnambs, T. (2025). Sample-Size Planning in Item-Response Theory: A Tutorial. *Advances in Methods and Practices in Psychological Science*, *8*(1), 25152459251314798. https://doi.org/10.1177/25152459251314798

Schwartz, R. C., & Blankenship, D. M. (2014). Racial disparities in psychotic disorder diagnosis: A review of empirical literature. *World Journal of Psychiatry*, *4*(4), 133–140. https://doi.org/10.5498/wjp.v4.i4.133

Schwartzman, J. M., Williams, Z. J., & Corbett, B. A. (2022). Diagnostic- and sex-based differences in depression symptoms in autistic and neurotypical early adolescents. *Autism*, *26*(1), 256–269. https://doi.org/10.1177/13623613211025895

Sen, S., Cohen, A. S., & Kim, S.-H. (2016). The Impact of Non-Normality on Extraction of Spurious Latent Classes in Mixture IRT Models. *Applied Psychological Measurement*, *40*(2), 98–113. https://doi.org/10.1177/0146621615605080

Shaw, M., Cloos, L. J. R., Luong, R., Elbaz, S., & Flake, J. K. (2020). Measurement practices in large-scale replications: Insights from Many Labs 2. *Canadian Psychology / Psychologie Canadienne*, *61*(4), 289–298. https://doi.org/10.1037/cap0000220

Shealy, R., & Stout, W. (1993). A Model-Based Standardization Approach that Separates True Bias/DIF from Group Ability Differences and Detects Test Bias/DTF as well as Item Bias/DIF. *Psychometrika*, *58*(2), 159–194. https://doi.org/10.1007/BF02294572

Sheldrick, R. C., Benneyan, J. C., Kiss, I. G., Briggs-Gowan, M. J., Copeland, W., & Carter, A. S. (2015). Thresholds and accuracy in screening tools for early detection of psychopathology. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *56*(9), 936–948. https://doi.org/10.1111/jcpp.12442

Sijtsma, K., Ellis, J. L., & Borsboom, D. (2024). Recognize the Value of the Sum Score, Psychometrics' Greatest Accomplishment. *Psychometrika*, *89*(1), 84–117. https://doi.org/10.1007/s11336-024-09964-7

Silva, L. de A., Noll, M., Siqueira, G. C., & Barbosa, A. K. N. (2024). Assessing Misophonia in Young Adults: The Prevalence and Psychometric Validation of the MisoQuest Questionnaire. *Healthcare*, *12*(18), 1888. https://doi.org/10.3390/healthcare12181888

Smith, G. T., McCarthy, D. M., & Zapolski, T. C. B. (2009). On the Value of Homogeneous Constructs for Construct Validation, Theory Testing, and the Description of Psychopathology. *Psychological Assessment*, *21*(3), 272–284. https://doi.org/10.1037/a0016699

Smith, J. D. (2014). Prototypes, exemplars, and the natural history of categorization. *Psychonomic Bulletin & Review, 21*(2), 312–331. https://doi.org/10.3758/s13423-013-0506-0

Soland, J., Kuhfeld, M., & Edwards, K. (2024). How survey scoring decisions can influence your study's results: A trip through the IRT looking glass. *Psychological Methods*, *29*(5), 1003–1024. https://doi.org/10.1037/met0000506

Song, Q. C., Tang, C., & Wee, S. (2021). Making Sense of Model Generalizability: A Tutorial on Cross-Validation in R and Shiny. *Advances in Methods and Practices in Psychological Science*, *4*(1), 2515245920947067. https://doi.org/10.1177/2515245920947067

Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology*, *15*(2), 201–293.

Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A Comparison of Random Forest Variable Selection Methods for Classification Prediction Modeling. *Expert Systems with Applications*, *134*, 93–101. https://doi.org/10.1016/j.eswa.2019.05.028

Spence, I. (1983). Monte Carlo Simulation Studies. *Applied Psychological Measurement*, *7*(4), 405–425. https://doi.org/10.1177/014662168300700403

Staccio, B. (2025). ADHD: Is it over-pathologized and over-diagnosed? Understanding potential reasons for the over-diagnosis of ADHD and how different theoretical lenses may explain ADHD and it's rising prevalence rates. *SUNY New Paltz Journal of Psychology*, *1*. https://ojs.newpaltz.edu/index.php/SNPJP/article/view/22

Stals, Y., du Plessis, E., Pretorius, P. J., Nel, M., & Boateng, A. (2024). Depression, anxiety and coping mechanisms among mental healthcare practitioners during COVID-19. *The South*

African Journal of Psychiatry : SAJP : The Journal of the Society of Psychiatrists of South Africa, *30*, 2307. https://doi.org/10.4102/sajpsychiatry.v30i0.2307

Stefana, A., Damiani, S., Granziol, U., Provenzani, U., Solmi, M., Youngstrom, E. A., & Fusar-Poli, P. (2025). Psychological, psychiatric, and behavioral sciences measurement scales: Best practice guidelines for their development and validation. *Frontiers in Psychology*, *15*. https://doi.org/10.3389/fpsyg.2024.1494261

Stickle, T. R., & Weems, C. F. (2006). Improving Prediction From Clinical Assessment: The Roles of Measurement, Psychometric Theory, and Decision Theory. In *Strengthening research methodology: Psychological measurement and evaluation* (pp. 213–230). American Psychological Association. https://doi.org/10.1037/11384-011

Stout, W., Li, H.-H., Nandakumar, R., & Bolt, D. (1997). MULTISIB: A Procedure to Investigate DIF When a Test is Intentionally Two-Dimensional. *Applied Psychological Measurement*, *21*(3), 195–213. https://doi.org/10.1177/01466216970213001

Strobl, C. (2013). Data Mining. In *The Oxford handbook of quantitative methods: Statistical analysis* (Vol. 2, pp. 678–700). Oxford University Press.

Strobl, C., Boulesteix, A.-L., & Augustin, T. (2007). Unbiased split selection for classification trees based on the Gini Index. *Computational Statistics & Data Analysis*, *52*(1), 483–501. https://doi.org/10.1016/j.csda.2006.12.030

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, *8*(1), 25. https://doi.org/10.1186/1471-2105-8-25

Stucky, B. D., & Edelen, M. O. (2014). Using Hierarchical IRT Models to Create Unidimensional Measures From Multidimensional Data. In *Handbook of Item Response Theory Modeling* (pp. 183–251). Routledge.

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

Sun, X., Liu, Y., Xin, T., & Song, N. (2020). The Impact of Item Calibration Error on Variable-Length Cognitive Diagnostic Computerized Adaptive Testing. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.575141

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., & Feuston, B. P. (2003). Random forest: A classification and regression tool for compound classification and QSAR modeling. *Journal of Chemical Information and Computer Sciences*, *43*(6), 1947–1958. https://doi.org/10.1021/ci034160g

Sweet, L., Müller, C., Anand, M., & Zscheischler, J. (2023). Cross-Validation Strategy Impacts the Performance and Interpretation of Machine Learning Models. *Artificial Intelligence for the Earth Systems*, *2*(4). https://doi.org/10.1175/AIES-D-23-0026.1

Tamano, H., Hino, H., & Mochihashi, D. (2025). *Misspecifying non-compensatory as compensatory IRT: Analysis of estimated skills and variance* (arXiv:2507.15222). arXiv. https://doi.org/10.48550/arXiv.2507.15222

Tan, Z., Torre, J., ma, W., Huh, D., Larimer, M., & Mun, E.-Y. (2022). A Tutorial on Cognitive Diagnosis Modeling for Characterizing Mental Health Symptom Profiles Using Existing Item Responses. *Prevention Science*, *24*. https://doi.org/10.1007/s11121-022-01346-8

Tang, X., Schalet, B. D., Peipert, J. D., & Cella, D. (2023). Does Scoring Method Impact Estimation of Significant Individual Changes Assessed by Patient-Reported Outcome Measures? Comparing Classical Test Theory Versus Item Response Theory. *Value in Health: The Journal of the International Society for Pharmacoeconomics and Outcomes Research*, *26*(10), 1518–1524. https://doi.org/10.1016/j.jval.2023.06.002

Tay, L., Woo, S. E., Hickman, L., Booth, B. M., & D'Mello, S. (2022). A Conceptual Framework for Investigating and Mitigating Machine-Learning Measurement Bias (MLMB) in Psychological Assessment. *Advances in Methods and Practices in Psychological Science*, *5*(1), 25152459211061337. https://doi.org/10.1177/25152459211061337

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*(3), 287–305. https://doi.org/10.1037/1082-989X.11.3.287

Templin, J. L., Henson, R. A., Templin, S. E., & Roussos, L. (2008). Robustness of Hierarchical Modeling of Skill Association in Cognitive Diagnosis Models. *Applied Psychological Measurement*, *32*(7), 559–574. https://doi.org/10.1177/0146621607300286

Teplin, L. A., Abram, K. M., & Luna, M. J. (2023). Racial and Ethnic Biases and Psychiatric Misdiagnoses: Toward More Equitable Diagnosis and Treatment. *American Journal of Psychiatry*. (Washington, DC). https://doi.org/10.1176/appi.ajp.20230294

Terluin, B., Unalan, P. C., Turfaner Sipahioğlu, N., Arslan Özkul, S., & van Marwijk, H. W. J. (2016). Cross-cultural validation of the Turkish Four-Dimensional Symptom Questionnaire (4DSQ) using differential item and test functioning (DIF and DTF) analysis. *BMC Family Practice*, *17*(1), 53. https://doi.org/10.1186/s12875-016-0449-4

Therneau, T., & Atkinson, B. (2025). *rpart: Recursive Partitioning and Regression Trees* (Version R package version 4.1.24, p. 4.1.24) [Computer software]. https://doi.org/10.32614/CRAN.package.rpart

Thiele, C., & Hirschfeld, G. (2023). Confidence intervals and sample size planning for optimal cutpoints. *PLOS ONE*, *18*(1), e0279693. https://doi.org/10.1371/journal.pone.0279693

Thissen, D., & Thissen-Roe, A. (2022). Latent Variable Estimation in Factor Analysis and Item Response Theory. *Chinese/English Journal of Educational Measurement and Evaluation*, *3*(3). https://doi.org/10.59863/OPTZ4045

Thissen, D., & Wainer, H. (2001). *Test scoring* (pp. xii, 422). Lawrence Erlbaum Associates Publishers.

Thomson, G. H. (1935). The definition and measurement of "g" (general intelligence). *Journal of Educational Psychology*, *26*(4), 241–262. https://doi.org/10.1037/h0059873

Thomson, G. H. (1936). Some points of mathematical technique in the factorial analysis of ability. *Journal of Educational Psychology*, *27*(1), 37–54. https://doi.org/10.1037/h0062007

Thomson, G. H. (1938). Methods of Estimating Mental Factors. *Nature*, *141*(3562), 246–246. https://doi.org/10.1038/141246a0

Thorndike, R. L. (1953). Who Belongs in the Family? *Psychometrika*, *18*(4), 267–276. https://doi.org/10.1007/BF02289263

Thurstone, L. L. (1935). *The vectors of mind: Multiple-factor analysis for the isolation of primary traits.* University of Chicago Press. https://doi.org/10.1037/10018-000

Thurstone, L. L. (1947). *Multiple-factor analysis; a development and expansion of The Vectors of Mind* (pp. xix, 535). University of Chicago Press.

Tiego, J., Martin, E. A., DeYoung, C. G., Hagan, K., Cooper, S. E., Pasion, R., Satchell, L., Shackman, A. J., Bellgrove, M. A., & Fornito, A. (2023). Precision behavioral phenotyping as a strategy for uncovering the biological correlates of psychopathology. *Nature Mental Health*, *1*(5), 304–315. https://doi.org/10.1038/s44220-023-00057-5

Uebelacker, L. A., Strong, D., Weinstock, L. M., & Miller, I. W. (2009). Use of item response theory to understand differential functioning of DSM-IV major depression symptoms by race, ethnicity and gender. *Psychological Medicine*, *39*(4), 591–601. https://doi.org/10.1017/S0033291708003875

Valeri, L., & Vanderweele, T. J. (2014). The estimation of direct and indirect causal effects in the presence of misclassified binary mediator. *Biostatistics (Oxford, England)*, *15*(3), 498–512. https://doi.org/10.1093/biostatistics/kxu007

Van Groen, M. M., Eggen, T. J. H. M., & Veldkamp, B. P. (2019). Multidimensional Computerized Adaptive Testing for Classifying Examinees. In *Theoretical and Practical Advances in Computer-based Educational Measurement*.

Vitoratou, S., Uglik-Marucha, N., Hayes, C., & Gregory, J. (2021). Listening to People with Misophonia: Exploring the Multiple Dimensions of Sound Intolerance Using a New Psychometric Tool, the S-Five, in a Large Sample of Individuals Identifying with the Condition. *Psych*, *3*(4), 639–662. https://doi.org/10.3390/psych3040041

Vowels, M. J. (2023). Misspecification and unreliable interpretations in psychology and social science. *Psychological Methods*, *28*(3), 507–526. https://doi.org/10.1037/met0000429

Wakefield, J. C. (2007). The concept of mental disorder: Diagnostic implications of the harmful dysfunction analysis. *World Psychiatry*, *6*(3), 149–156.

Wakefield, J. C. (2010). False positives in psychiatric diagnosis: Implications for human freedom. *Theoretical Medicine and Bioethics*, *31*(1), 5–17. https://doi.org/10.1007/s11017-010-9132-2

Wakefield, J. C. (2015). Psychological Justice: Dsm-5, False Positive Diagnosis, and Fair Equality of Opportunity. *Public Affairs Quarterly*, *29*(1), 32–75.

Wakefield, J. C. (2016). Diagnostic Issues and Controversies in DSM-5: Return of the False Positives Problem. *Annual Review of Clinical Psychology*, *12*(Volume 12, 2016), 105–132. https://doi.org/10.1146/annurev-clinpsy-032814-112800

Wald, A. (1947). *Sequential analysis*.

Walker, C. M., & Beretvas, S. N. (2003). Comparing Multidimensional and Unidimensional Proficiency Classifications: Multidimensional IRT as a Diagnostic Aid. *Journal of Educational Measurement*, *40*(3), 255–275. https://doi.org/10.1111/j.1745-3984.2003.tb01107.x

Wallin, G., & Huang, Q. (2026). *A Hybrid Latent-Class Item Response Model for Detecting Measurement Non-Invariance in Ordinal Scales* (arXiv:2601.17612). arXiv. https://doi.org/10.48550/arXiv.2601.17612

Wang, C. (2015). On Latent Trait Estimation in Multidimensional Compensatory Item Response Models. *Psychometrika*, *80*(2), 428–449. https://doi.org/10.1007/s11336-013-9399-0

Wang, C., Xu, G., & Zhang, X. (2019). Correction for Item Response Theory Latent Trait Measurement Error in Linear Mixed Effects Models. *Psychometrika*, *84*(3), 673–700. https://doi.org/10.1007/s11336-019-09672-7

Wang, D., Gao, X., Cai, Y., & Tu, D. (2019). Development of a New Instrument for Depression With Cognitive Diagnosis Models. *Frontiers in Psychology, 10*. https://doi.org/10.3389/fpsyg.2019.01306

Wang, J., & Tian, L. (2024). Optimal Cut-Point Selection Methods Under Binary Classification When Subclasses Are Involved. *Pharmaceutical Statistics*, *23*(6), 984–1030. https://doi.org/10.1002/pst.2413

Wang, W.-C., & Su, Y.-H. (2004). Effects of Average Signed Area Between Two Item Characteristic Curves and Test Purification Procedures on the DIF Detection via the Mantel-Haenszel Method. *Applied Measurement in Education, 17*(2), 113–144. https://doi.org/10.1207/s15324818ame1702_2

Watts, A. L., Greene, A. L., Bonifay, W., & Fried, E. I. (2024). A critical evaluation of the p-factor literature. *Nature Reviews Psychology*, *3*(2), 108–122. https://doi.org/10.1038/s44159-023-00260-2

Weiss, B. A., & Dardick, W. (2021). Making the Cut: Comparing Methods for Selecting Cut-Point Location in Logistic Regression. *The Journal of Experimental Education*, *89*(4), 721–741. https://doi.org/10.1080/00220973.2019.1689375

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Widaman, K. F., & Revelle, W. (2023). Thinking thrice about sum scores, and then some more about measurement and analysis. *Behavior Research Methods*, *55*(2), 788–806. https://doi.org/10.3758/s13428-022-01849-w

Widaman, K. F., & Revelle, W. (2024). Thinking About Sum Scores Yet Again, Maybe the Last Time, We Don't Know, Oh No . . .1: A Comment on McNeish (2023). *Educational and*

Psychological Measurement, *84*(4), 637–659. https://doi.org/10.1177/00131644231205310

Widiger, T. A., & Samuel, D. B. (2005). Diagnostic categories or dimensions? A question for the Diagnostic And Statistical Manual Of Mental Disorders--fifth edition. *Journal of Abnormal Psychology*, *114*(4), 494–504. https://doi.org/10.1037/0021-843X.114.4.494

William, G. J. (2011). *Data Mining with {Rattle} and {R}: The art of excavating data for knowledge discovery* [R]. Springer. https://rd.springer.com/book/10.1007/978-1-4419-9890-3

Williams, Z. J., Cascio, C. J., & Woynaroski, T. G. (2022). Psychometric validation of a brief self-report measure of misophonia symptoms and functional impairment: The duke-vanderbilt misophonia screening questionnaire. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.897901

Wilson, M., & Hoskens, M. (2005). Multidimensional Item Responses: Multimethod-Multitrait Perspectives. In *Applied Rasch measurement: A book of exemplars: Papers in honour of John P. Keeves* (pp. 287–307). Dordrecht: Springer.

Winsberg, S., & De Soete, G. (1993). A latent class approach to fitting the weighted Euclidean model, clascal. *Psychometrika*, *58*(2), 315–330. https://doi.org/10.1007/BF02294578

Woods, C. M. (2007). Empirical Histograms in Item Response Theory With Ordinal Data. *Educational and Psychological Measurement*, *67*(1), 73–87. https://doi.org/10.1177/0013164406288163

Woods, C. M. (2009). Evaluation of MIMIC-Model Methods for DIF Testing With Comparison to Two-Group Analysis. *Multivariate Behavioral Research*, *44*(1), 1–27. https://doi.org/10.1080/00273170802620121

Wortmann, J. H., Jordan, A. H., Weathers, F. W., Resick, P. A., Dondanville, K. A., Hall-Clark, B. N., Foa, E. B., Young-McCaughan, S., Yarvis, J. S., Hembree, E. A., Mintz, J., Peterson, A. L., Litz, B. T., & Consortium, S. S. (2016). Psychometric analysis of the PTSD Checklist-5 (PCL-5) among treatment-seeking military service members. *Psychological Assessment*, *28*(11), 1392–1403. https://doi.org/10.1037/pas0000260

Wu, M. S., Lewin, A. B., Murphy, T. K., & Storch, E. A. (2014). Misophonia: Incidence, Phenomenology, and Clinical Correlates in an Undergraduate Student Sample: Misophonia. *Journal of Clinical Psychology*, *70*(10), 994–1007. https://doi.org/10.1002/jclp.22098

Xu, S., & Liu, Y. (2022). Characterizing Sampling Variability for Item Response Theory Scale Scores in a Fixed-Parameter Calibrated Projection Design. *Applied Psychological Measurement*, *46*(6), 509–528. https://doi.org/10.1177/01466216221108136

Yang, J. S., Hansen, M., & Cai, L. (2012). Characterizing Sources of Uncertainty in IRT Scale Scores. *Educational and Psychological Measurement*, *72*(2), 264–290. https://doi.org/10.1177/0013164411410056

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Yavuz Temel, G. (2023). A Simulation and Empirical Study of Differential Test Functioning (DTF). *Psych*, *5*, 478–496. https://doi.org/10.3390/psych5020032

Yland, J. J., Wesselink, A. K., Lash, T. L., & Fox, M. P. (2022). Misconceptions About the Direction of Bias From Nondifferential Misclassification. *American Journal of Epidemiology*, *191*(8), 1485–1495. https://doi.org/10.1093/aje/kwac035

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, *3*(1), 32–35. https://doi.org/10.1002/1097-0142(1950)3:1%3C32::AID-CNCR2820030106%3E3.0.CO;2-3

Youngstrom, E. A. (2013). A Primer on Receiver Operating Characteristic Analysis and Diagnostic Efficiency Statistics for Pediatric Psychology: We Are Ready to ROC. *Journal of Pediatric Psychology*, *39*(2), 204. https://doi.org/10.1093/jpepsy/jst062

Zanon, C., Brenner, R. E., Baptista, M. N., Vogel, D. L., Rubin, M., Al-Darmaki, F. R., Gonçalves, M., Heath, P. J., Liao, H.-Y., Mackenzie, C. S., Topkaya, N., Wade, N. G., & Zlati, A. (2021). Examining the Dimensionality, Reliability, and Invariance of the Depression, Anxiety, and Stress Scale–21 (DASS-21) Across Eight Countries. *Assessment*, *28*(6), 1531–1544. https://doi.org/10.1177/1073191119887449

Zheng, G., & Tian, X. (2005). The impact of diagnostic error on testing genetic association in case–control studies. *Statistics in Medicine*, *24*(6), 869–882. https://doi.org/10.1002/sim.1976

Zimmerman, M. (2022). Positive Predictive Value: A Clinician's Guide to Avoid Misinterpreting the Results of Screening Tests. *The Journal of Clinical Psychiatry*, *83*(5), 22com14513. https://doi.org/10.4088/JCP.22com14513

Zimmerman, M., & Holst, C. G. (2018). Screening for psychiatric disorders with self-administered questionnaires. *Psychiatry Research*, *270*, 1068–1073. https://doi.org/10.1016/j.psychres.2018.05.022

Zou, D. (2025). *Investigating the impact of hidden subpopulations on individual test scores using IRT-based methods* [University of British Columbia]. https://doi.org/10.14288/1.0450045

Zumbo, B. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-Type (Ordinal) Item Scores*. Directorate of Human Resources Research and Evaluation.

Zumbo, B., & Gelin, M. N. (2005). A Matter of Test Bias in Educational Policy Research: Bringing the Context into Picture by Investigating Sociological/Community Moderated (or Mediated) Test and Item Bias. *Journal of Educational Research & Policy Studies*, *5*(1), 1–23.

Zweig, M. H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, *39*(4), 561–577. https://doi.org/10.1093/clinchem/39.4.561

Tables

Table 1. Proposed Diagnostic Criteria for Misophonia (Schröder et al., 2013).

A. The presence or anticipation of a specific sound, produced by a human being (e.g. eating sounds, breathing sounds), provokes an impulsive aversive physical reaction which starts with irritation or disgust that instantaneously becomes anger.
B. This anger initiates a profound sense of loss of self-control with rare but potentially aggressive outbursts.
C. The person recognizes that the anger or disgust is excessive, unreasonable, or out of proportion to the circumstances or the provoking stressor.
D. The individual tends to avoid the misophonic situation, or if he/she does not avoid it, endures encounters with the misophonic sound situation with intense discomfort, anger or disgust.
E. The individual's anger, disgust or avoidance causes significant distress (i.e. it bothers the person that he or she has the anger or disgust) or significant interference in the person's day-to-day life. For example, the anger or disgust may make it difficult for the person to perform important tasks at work, meet new friends, attend classes, or interact with others.
F. The person's anger, disgust, and avoidance are not better explained by another disorder, such as obsessive-compulsive disorder (e.g. disgust in someone with an obsession about contamination) or post-traumatic stress disorder (e.g. avoidance of stimuli associated with a trauma related to threatened death, serious injury or threat to the physical integrity of self or others).

Table 2. Confusion Table for Classification in the Context of Psychological Diagnosis

		Predicted		
		Diagnosed	Not Diagnosed	Prediction Error
Gold Standard	Diagnosed	TP	FN	FN/(TP+FN)
	Not Diagnosed	FP	TN	FP/(FP+TN)
	Use Error	FP/(TP+FP)	FN/(FN+TN)	Total Error = $\frac{FP+FN}{TP+FN+FP+TN}$

Note. TP represents true positives (diagnosed individuals correctly predicted by the assessment). TN represents true negatives (non-diagnosed individuals correctly predicted by the assessment). FP represents false positives (non-diagnosed individuals incorrectly predicted as diagnosed). FN represents false negatives (diagnosed individuals incorrectly predicted as non-diagnosed). The main diagonal contains the number of cases correctly predicted by the assessment, and the off-diagonal contains the number of cases incorrectly predicted by the assessment.

Table 3. Manipulated Simulation Variables

Variable	Levels
Varying	
Dimensionality	One, Three
Correlation between $\theta_{Diagnosis}$ and θ_{Scale} (measurement error)	1, 0.5
Number of Items	
% of Total Items with DIF	0, 0.33, 0.66[9]
Factor with DIF	Single Factor, Across all factors
b DIF Strength	0, 1, 3
Impact	0, -1, 1
Fixed	
DIF Groups Structure	Interaction of Sex and Race
DIF Scope	Threshold parameter only
DIF Balance	Unbalanced
Prevalence of Diagnosed Class	30%
Sample Size	1,000

[9] Note that a DIF percentage of 66% cannot occur in a three-factor model when DIF occurs in only a single factor, as only 33% of items will load onto a single factor. As such, all 66% conditions occur in either (a) a one factor model, or (b) with DIF across multiple factors.

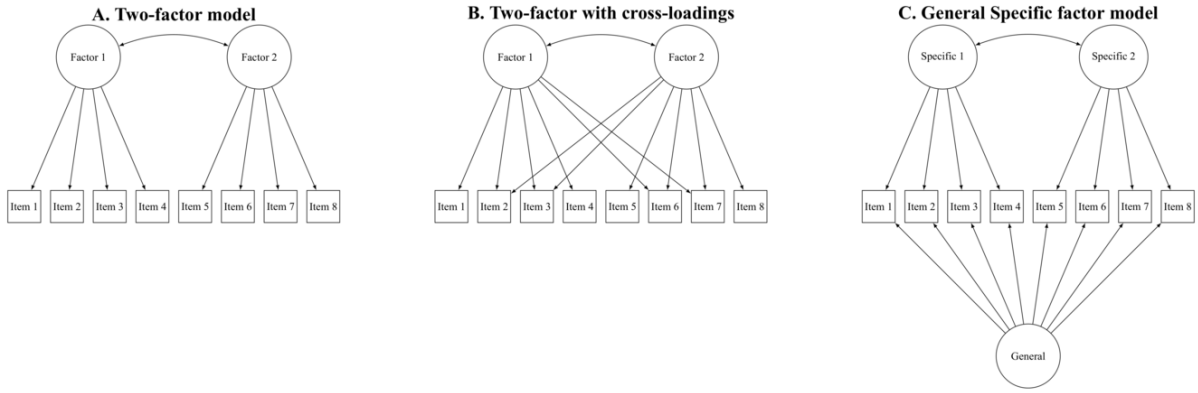
Table 4. Comparison of Fit Metrics for All CART Models with Second-Order IRT-Based and RF Methods Only

				Train		Test	
Metric (s^2)	Tuning Strategy	CP	Splits	MSE	R^2	MSE	R^2
Accuracy (0.033)							
	Full	0	303	0.0058	0.8210	0.0058	0.8203
	1SE	0.0001	70	0.0059	0.8185	0.0059	0.8177
	Elbow	0.0072	10	0.0077	0.7607	0.0078	0.7597
	Δ1SE - Elbow		60	-0.0019	0.0578	-0.0019	0.0579
Sensitivity (0.039)							
	Full	0	303	0.0064	0.8358	0.0065	0.8346
	1SE	0.0001	110	0.0065	0.8338	0.0066	0.8328
	Elbow	0.0041	18	0.0091	0.7666	0.0092	0.7653
	Δ1SE - Elbow		92	-0.0026	0.0671	-0.0027	0.0675
Specificity (0.036)							
	Full	0	303	0.0078	0.7841	0.0078	0.7834
	1SE	0.0001	75	0.0079	0.7816	0.0079	0.7810
	Elbow	0.0113	11	0.0109	0.6976	0.0110	0.6957
	Δ1SE - Elbow		64	-0.0030	0.0840	-0.0031	0.0853
Precision (0.039)							
	Full	0	303	0.0057	0.8565	0.0057	0.8555
	1SE	0.0001	84	0.0057	0.8545	0.0058	0.8536
	Elbow	0.0065	10	0.0082	0.7930	0.0083	0.7916
	Δ1SE - Elbow		74	-0.0024	0.0615	-0.0025	0.0620
NPV (0.020)							
	Full	0	303	0.0039	0.8060	0.0039	0.8057
	1SE	0.0001	87	0.0039	0.8032	0.0040	0.8030
	Elbow	0.0052	10	0.0051	0.7440	0.0052	0.7432
	Δ1SE - Elbow		77	-0.0012	0.0592	-0.0012	0.0598
F1 (0.038)							
	Full	0	303	0.0052	0.8643	0.0053	0.8631
	1SE	0.0001	109	0.0053	0.8623	0.0054	0.8611
	Elbow	0.0060	10	0.0081	0.7903	0.0082	0.7882
	Δ1SE - Elbow		99	-0.0028	0.0721	-0.0028	0.0729

Figures

Figure 1

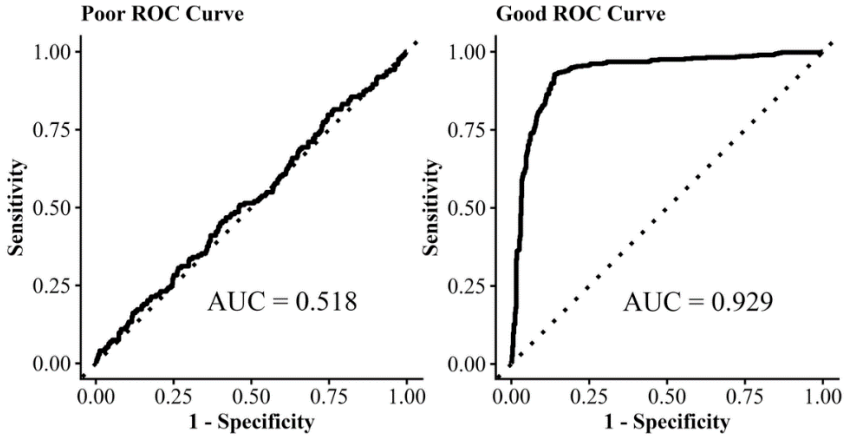
Conceptual Models of Multidimensionality



Note: Panel A shows a two-factor model representing between-item multidimensionality, where each item loads exclusively on one of two distinct latent factors, reflecting separate constructs. Panel B depicts a two-factor model with cross-loadings, illustrating within-item multidimensionality through items that load on multiple latent factors simultaneously. Panel C presents a general-specific factor model, a second form of within-item multidimensionality, in which each item loads on both a general factor (capturing shared variance across all items) and one or more specific factors (capturing unique variance attributable to subgroup or trait-specific dimensions). This structure can arise due to phenomena like differential item functioning (DIF), where item responses vary systematically across groups (e.g., by race or culture) despite equivalent levels of the underlying trait.

Figure 2

ROC curves illustrating the performance of two binary classifiers.



Note: The left panel illustrates a classifier with poor discrimination ability, yielding an area under the curve (AUC) of 0.54, which is close to random guessing. The right panel shows a classifier with good discrimination ability, achieving an AUC of 0.93, indicating strong predictive performance. The dashed diagonal line represents the performance of a random classifier (AUC = 0.5).

Figure 3

The Steps Required to Form Classifications Via IRT.

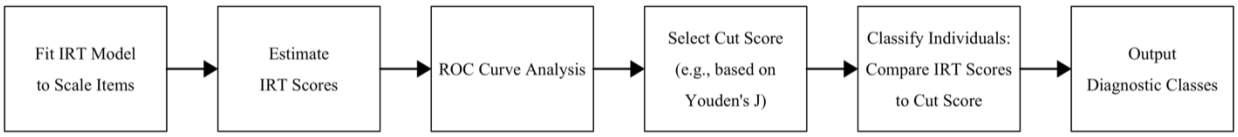
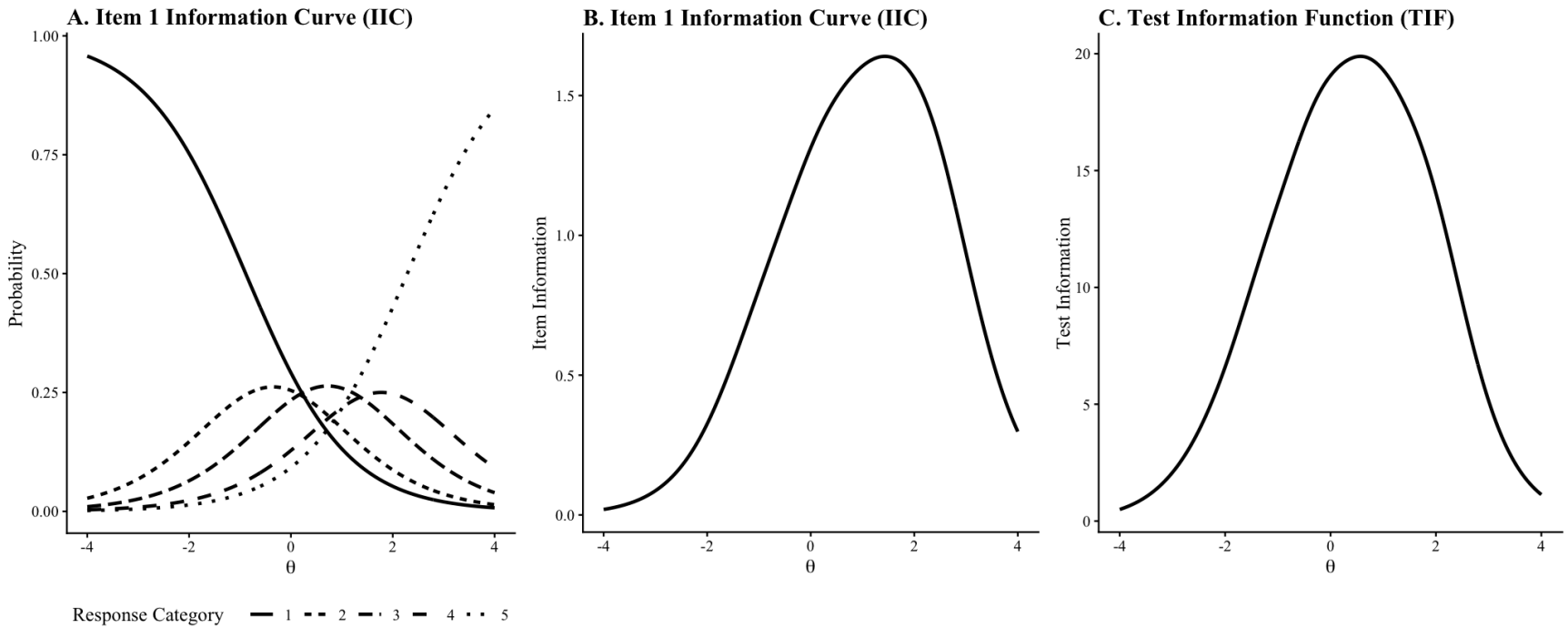


Figure 4

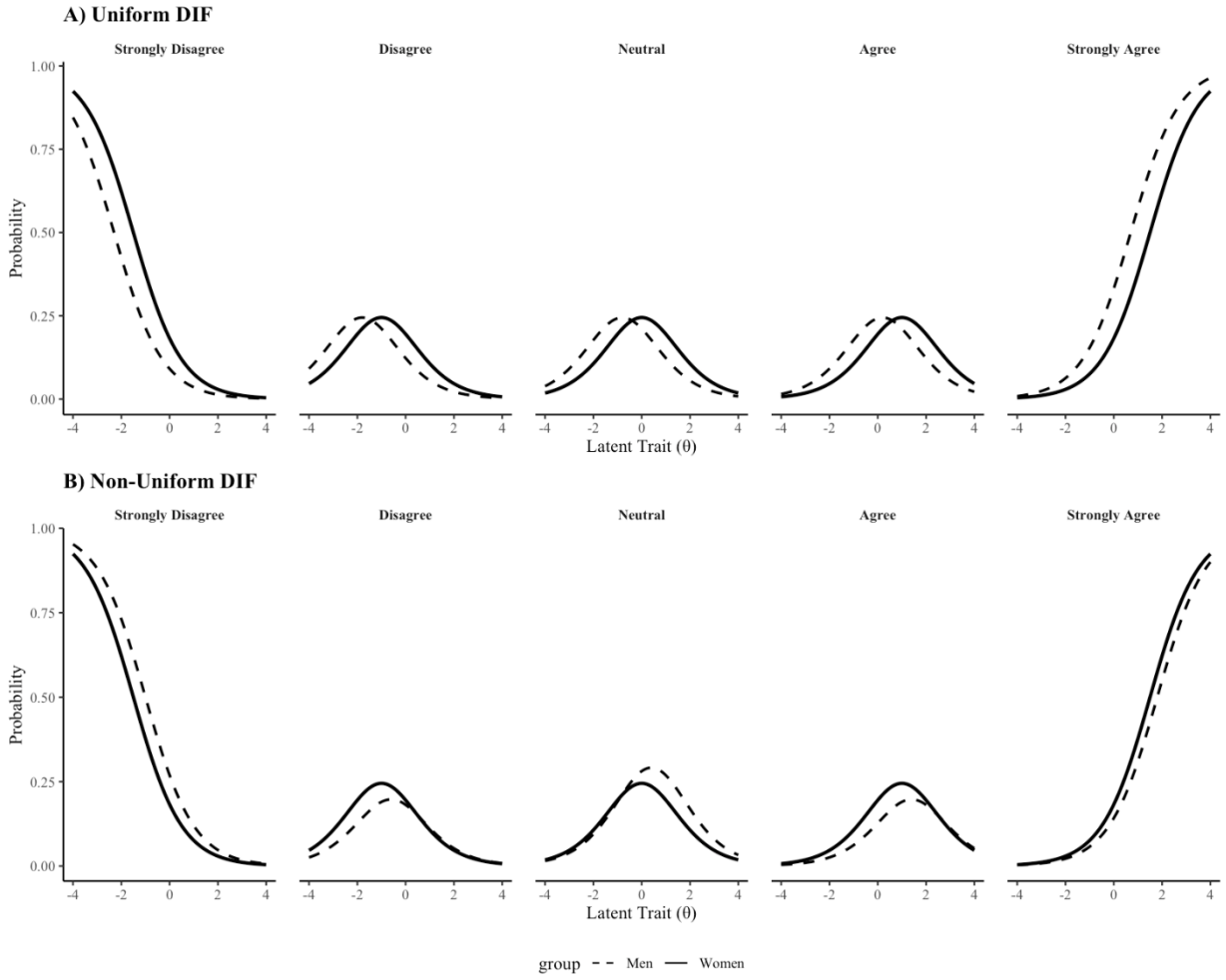
Trace Lines, Item Information Curves, and Test Information Curves.



Note: Panel A shows the category response curves for Item 1, plotting the probability of endorsing each response category (1 through 5) across levels of the latent trait (θ). Each curve represents the probability of selecting a specific response option as a function of θ, indicating how item responses vary with the underlying trait. Panel B depicts the Item Information Curve (IIC) for Item 1, showing the amount of information this item provides at different values of θ. The peak of the curve indicates where the item is most informative. Panel C displays the Test Information Function (TIF), which aggregates item information across all items in the test, illustrating the precision of the test at various levels of θ. Higher peaks in the TIF indicate greater measurement precision at those θ values.

Figure 5

Category Characteristic Curves for a 5-Point Likert Scale Item Exhibiting Differential Item Functioning (DIF).



Note: Panel A shows uniform DIF, where the probability curves for men (dashed lines) are consistently shifted relative to women (solid lines) across all response categories (1=Strongly Disagree to 5=Strongly Agree) such that men are less likely to endorse the item than women. Panel B shows non-uniform DIF, where the differences between men and women vary across the latent trait continuum, with curves crossing in some categories. Both panels display the probability of each response option as a function of the latent trait (θ), demonstrating how DIF manifests differently in these two forms. Trace lines are typically plotted on a single x-axis but are separated here for readability.

Figure 6

Example Decision Tree Illustrating a Simplified Diagnostic Pathway for Attention-Deficit/Hyperactivity Disorder (ADHD) Based on Representative Symptom Screening Questions.

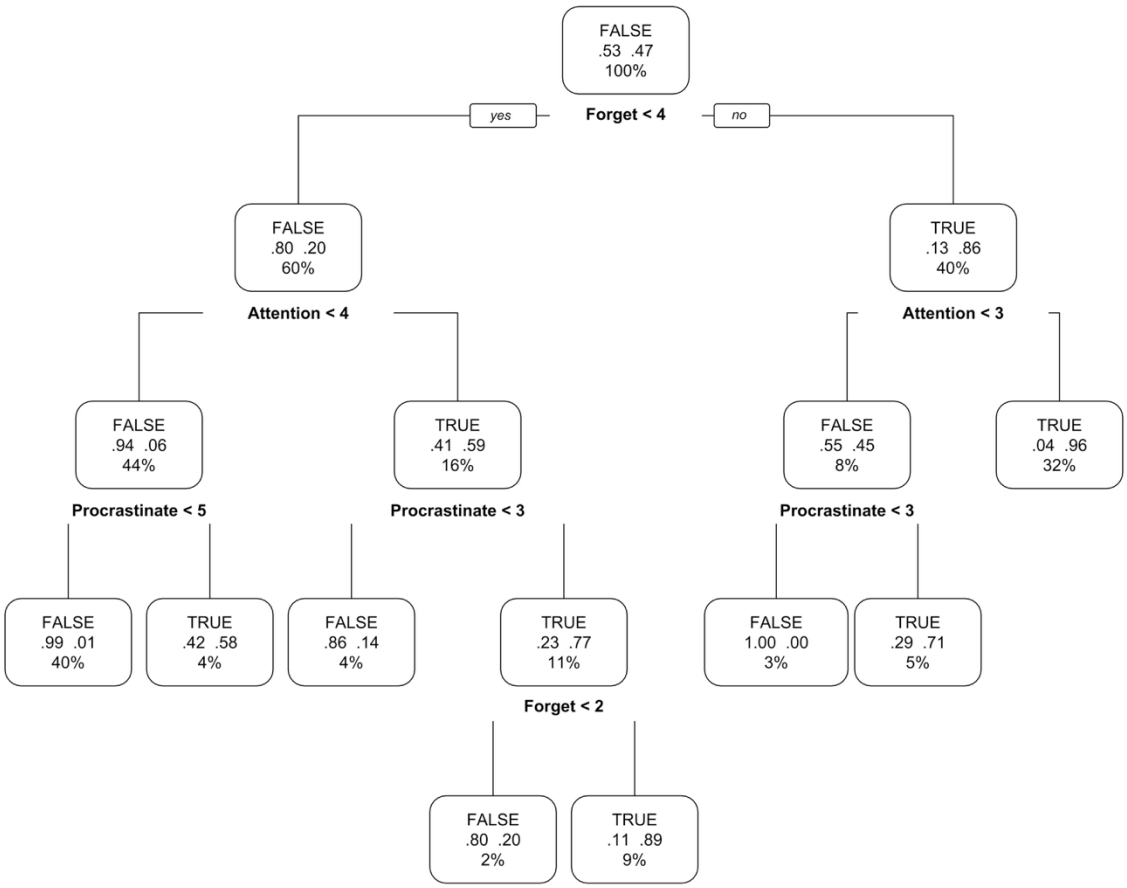


Figure 7

Diagram Illustrating the Two Possible Processes for Classification via RF.

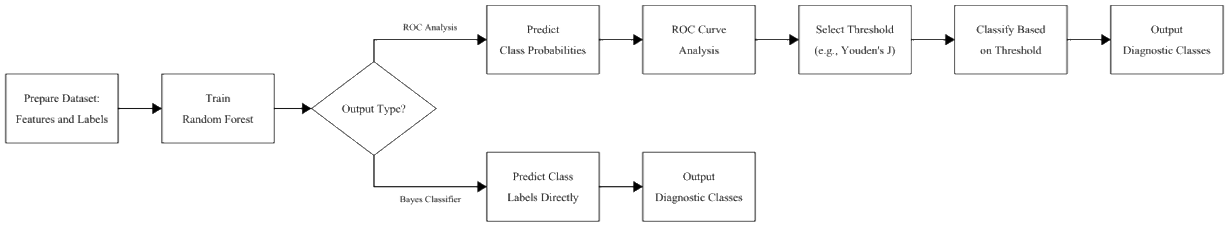
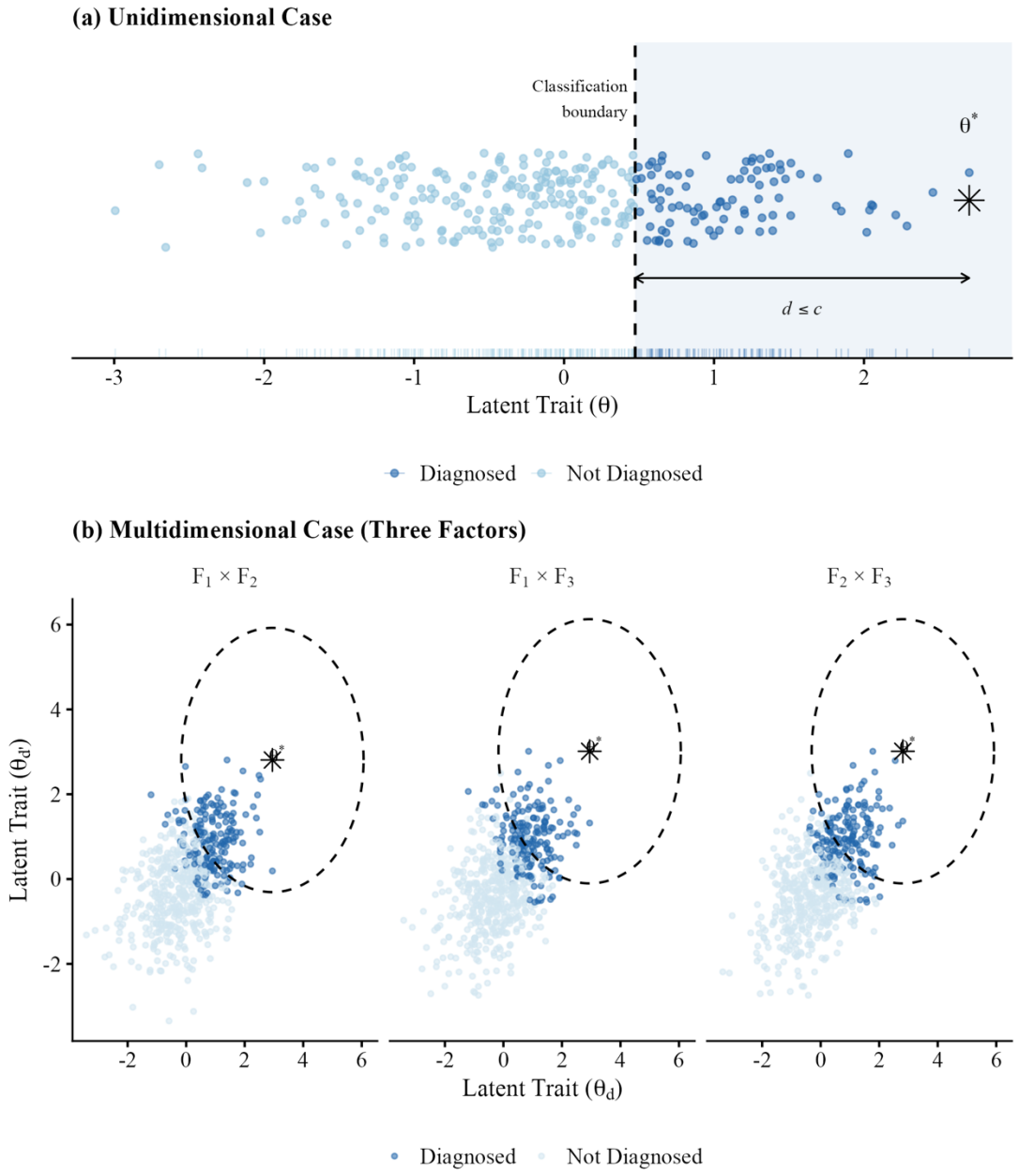


Figure 8

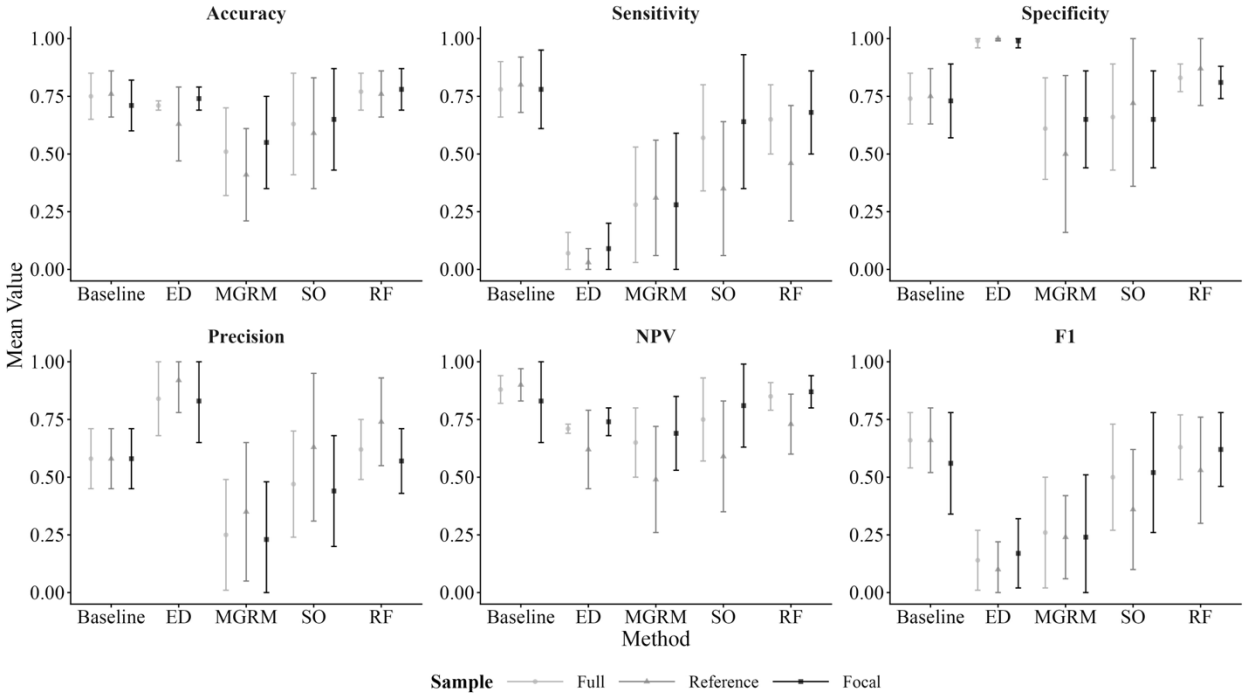
Illustration of the Euclidean Distance Approach for Generating Classes



Note. Panel (a) illustrates the unidimensional case. Individuals to the right of the boundary (distance $\leq c$ from θ^*) are classified as diagnosed. Panel (b) illustrates the three-dimensional case via pairwise 2D projections. This dashed circle approximates the projected decisions. Points closer to θ^* across all three dimensions simultaneously are classified as diagnosed.

Figure 9

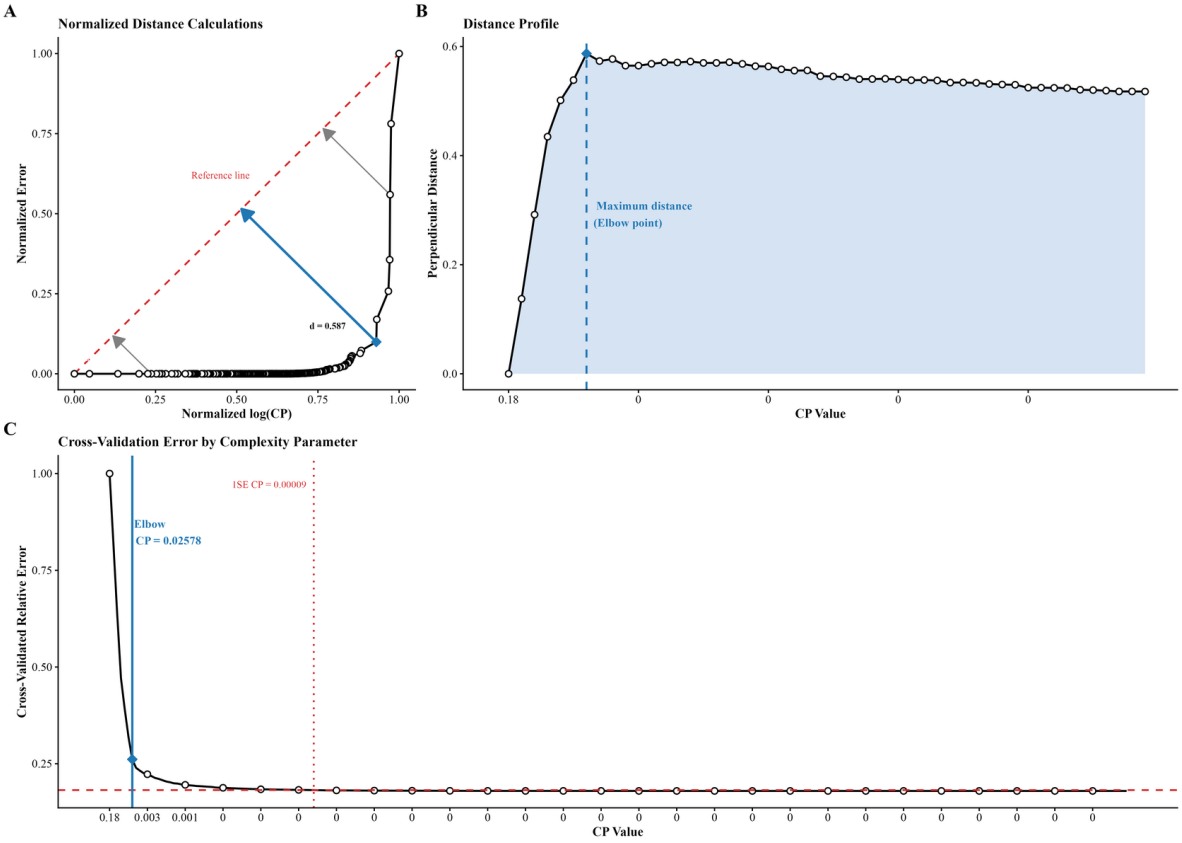
Marginal Means and Standard Deviations of Classification Performance Metrics by Method and Sample Group



Note: Mean performance across all outcome metrics and the four classification methods across all simulated conditions. Error bars represent ±1 standard deviation. Classification methods include: Baseline = baseline item response approach; ED = item response theory (IRT) using Euclidean distance threshold; Multi = IRT using dimension-specific thresholds for each dimension; S-O = second-order IRT based classifications; RF = random forest. Note that in conditions where the scale only contained one dimension, Multi and second-order IRT are the same model. Performance metrics are displayed separately for the full sample, reference group (female × White), and focal group (all other demographic combinations: female × Black, female × Latinx, and male × Black). All metrics are bounded between 0 and 1, with higher values indicating better classification performance. The full sample represents overall classification performance, while reference and focal group comparisons reveal potential differential performance across demographic subgroups in conditions with differential item functioning (DIF). Results are based on 1,000 Monte Carlo replications per condition across 168 simulation conditions varying in dimensionality, measurement error, scale length, DIF prevalence, DIF location, DIF strength, and impact.

Figure 10

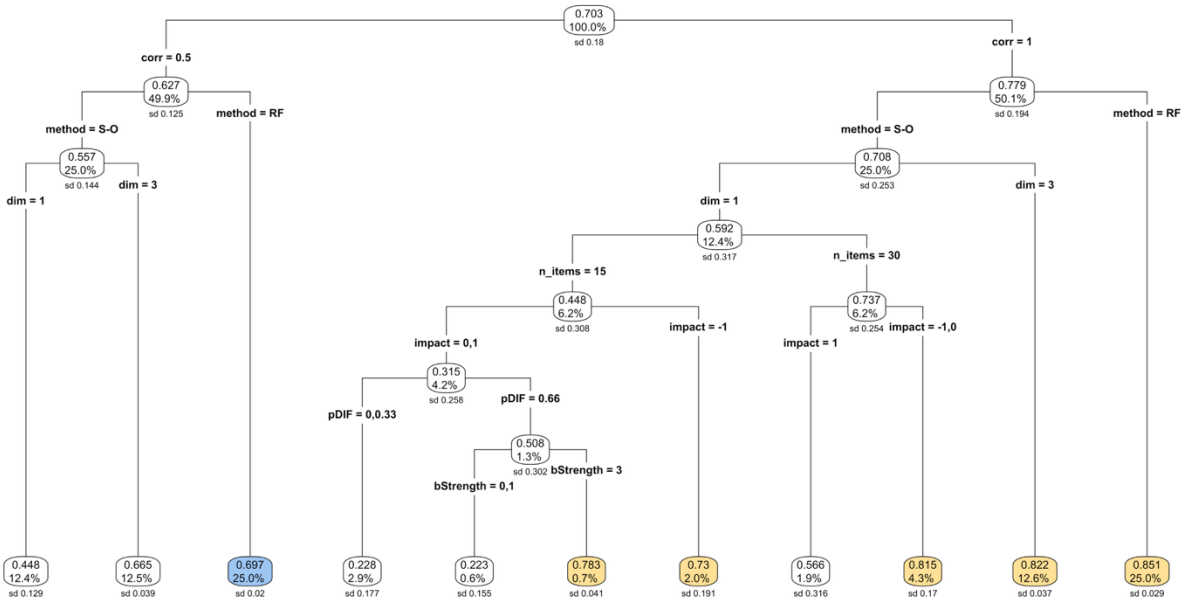
Distance-Based Elbow Detection for CART Complexity Parameter Selection



Note: This figure demonstrates the distance-based elbow detection method used to select complexity parameters (CP) for Classification and Regression Tree (CART) models analyzing accuracy outcomes from Monte Carlo simulations. Panel A illustrates the geometric basis of the elbow detection algorithm in normalized space, where both log(CP) and cross-validation error are scaled to the unit interval [0,1]. The dashed red line connects the first and last points of the normalized curve. Blue (maximum) and gray (additional examples) arrows demonstrate the perpendicular distance calculations from points on the curve to this reference line, with the point of maximum perpendicular distance (shown as a blue diamond) identifying the elbow. This geometric approach detects the point of maximum curvature where the trade-off between model complexity and prediction error is optimized. Panel B presents the distance profile across all CP values, showing how perpendicular distance varies along the sequence. The peak of this profile (marked with a blue diamond and vertical dashed line) corresponds to the selected elbow point. The distance-based method identified a CP value that balanced interpretability and predictive performance, yielding a more parsimonious tree than the 1SE rule while maintaining acceptable cross-validation error. All panels share the same CP sequence on the x-axis, with every 10th value displayed for clarity. Panel C displays the cross-validation error curve across the CP sequence. The vertical dashed red line indicates the CP value selected by the one-standard-error (1SE) rule. The vertical solid blue line marks the CP value selected by the distance-based elbow detection method, representing a more parsimonious model. The x-axis in Panel C is limited to plot only every 10^{th} value, as displaying all tested CP values decreased clarity.

Figure 11

Regression Tree Predicting Classification Accuracy of Second-Order IRT and RF Across Simulation Conditions.

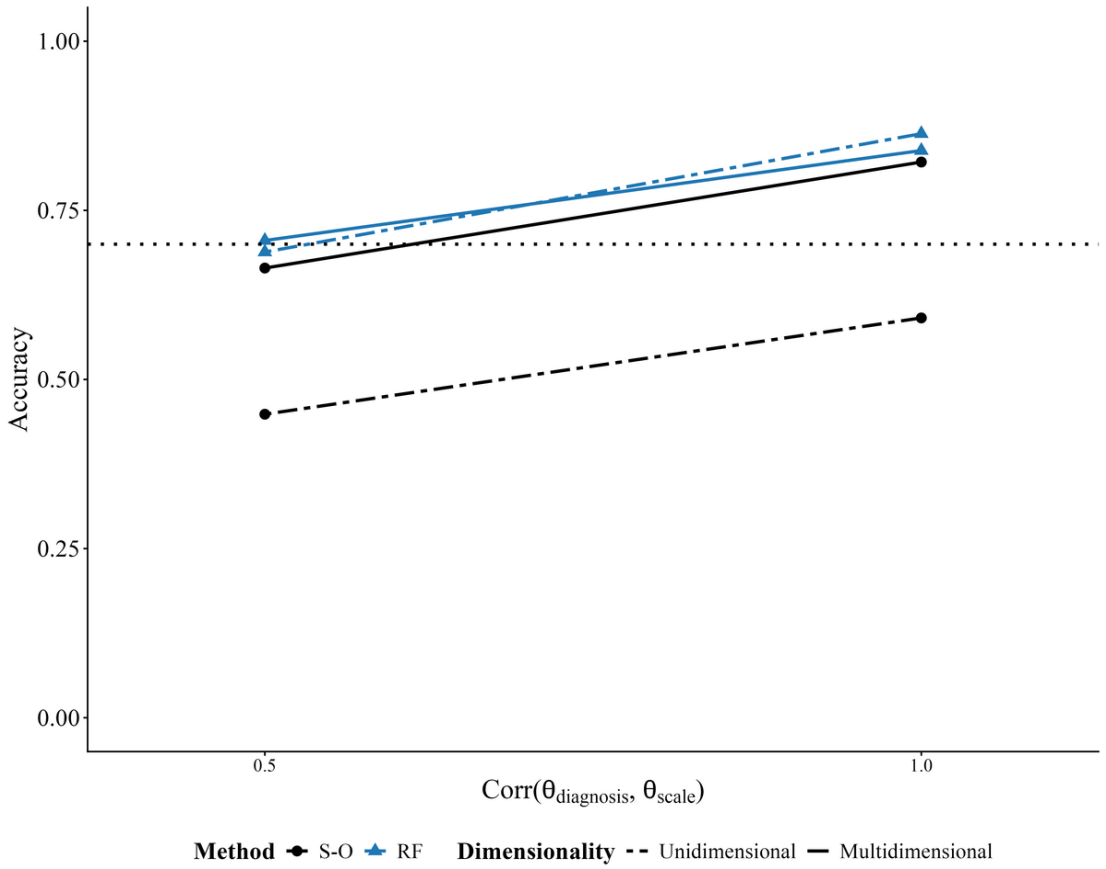


Note. S-O = second-order IRT approach; RF = random forest; dim = number of dimensions in the scale; pDIF = percent of items that contained DIF; bStrength = strength of DIF in the *b* parameter; impact value indicates the μ of the θ distribution of the focal group (μ for the reference group distribution was always zero). Yellow terminal nodes indicate conditions where second-order IRT-based and RF-based classifications were more accurate, on average, than a base rate of 0.7. Blue terminal nodes indicate where RF average classification accuracy was worse than the base rate, but better than second-order IRT-based classification accuracy.

Figure 12

Line Graph of Classification Accuracy for Second-Order IRT (S-O) and Random Forest (RF)

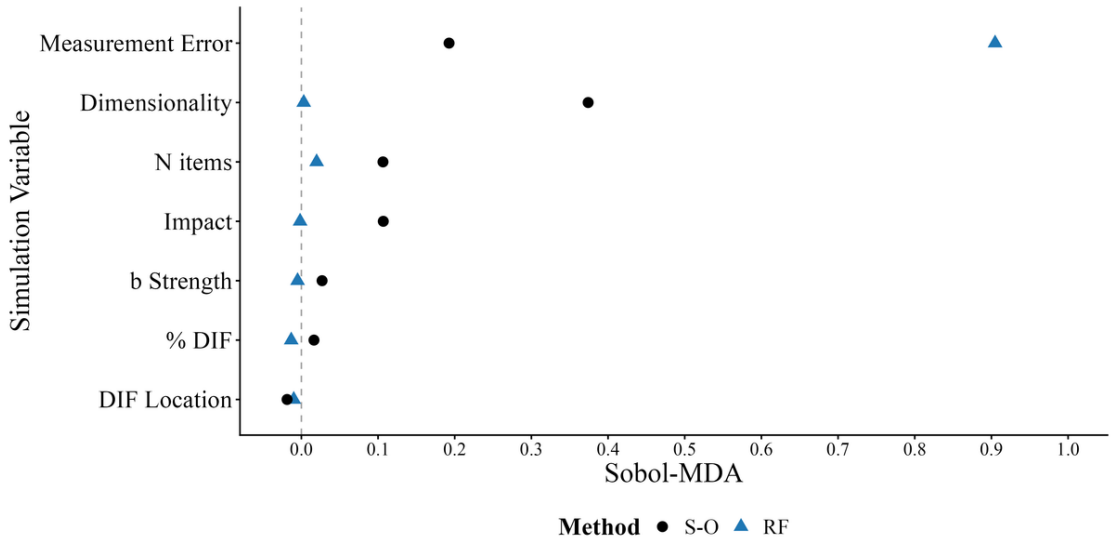
Methods Across Simulation Conditions Where S-O and RF Methods Do Not Meet Base Rates



Note: Line graphs of mean classification accuracy for second-order IRT (S-O) and RF classifications (color and shape) as a function of measurement error (x-axis) and scale dimensions (line type), and number of items and impact (columns). In all panels, the dotted horizontal line demarcates a base rate accuracy of 0.70. These data conditions led to terminal nodes in CART models where average accuracy of second-order IRT-based and RF classifications fell below 0.70 but RF classifications were more accurate than second-order IRT-based classifications.

Figure 13

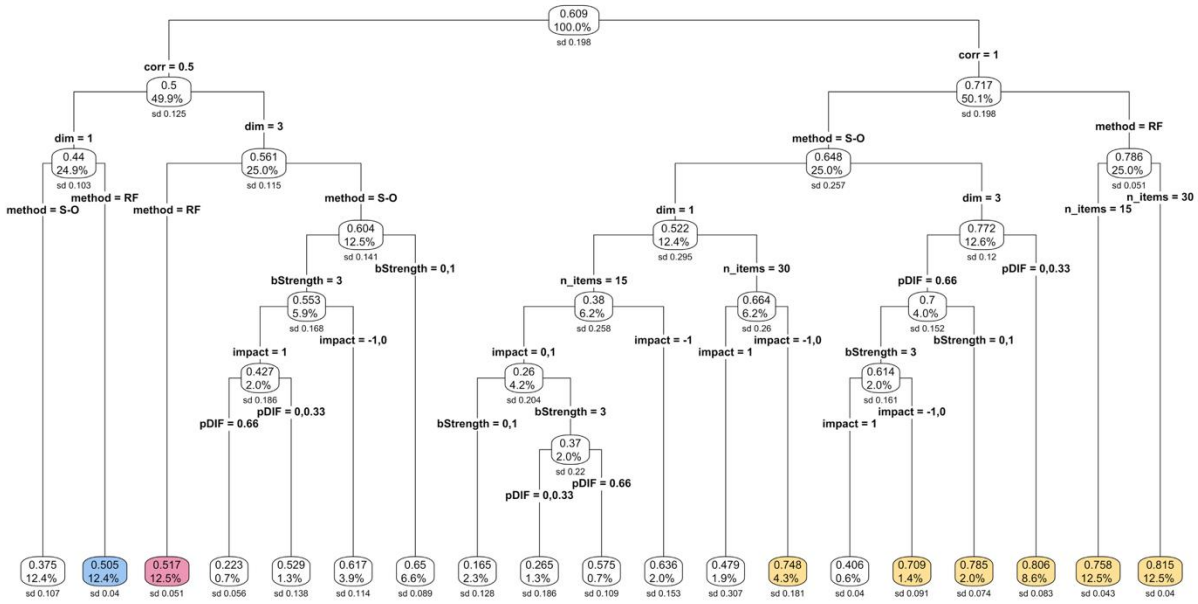
Sobol MDA Variable Importance Measures for Predicting Classification Accuracy Using Either the RF or Second-Order IRT-Based Approach.



Note: RF = random forest; S-O = second-order IRT; DIF location = where DIF occurred in the scale (i.e., in a single factor versus across multiple factors). Dark red circles and blue triangles correspond to second-order IRT and RF approaches, respectively. A vertical line is included to demarcate the zero point. Negative Sobol MDA values indicate that removing the predictor from the model improves predictive accuracy.

Figure 14

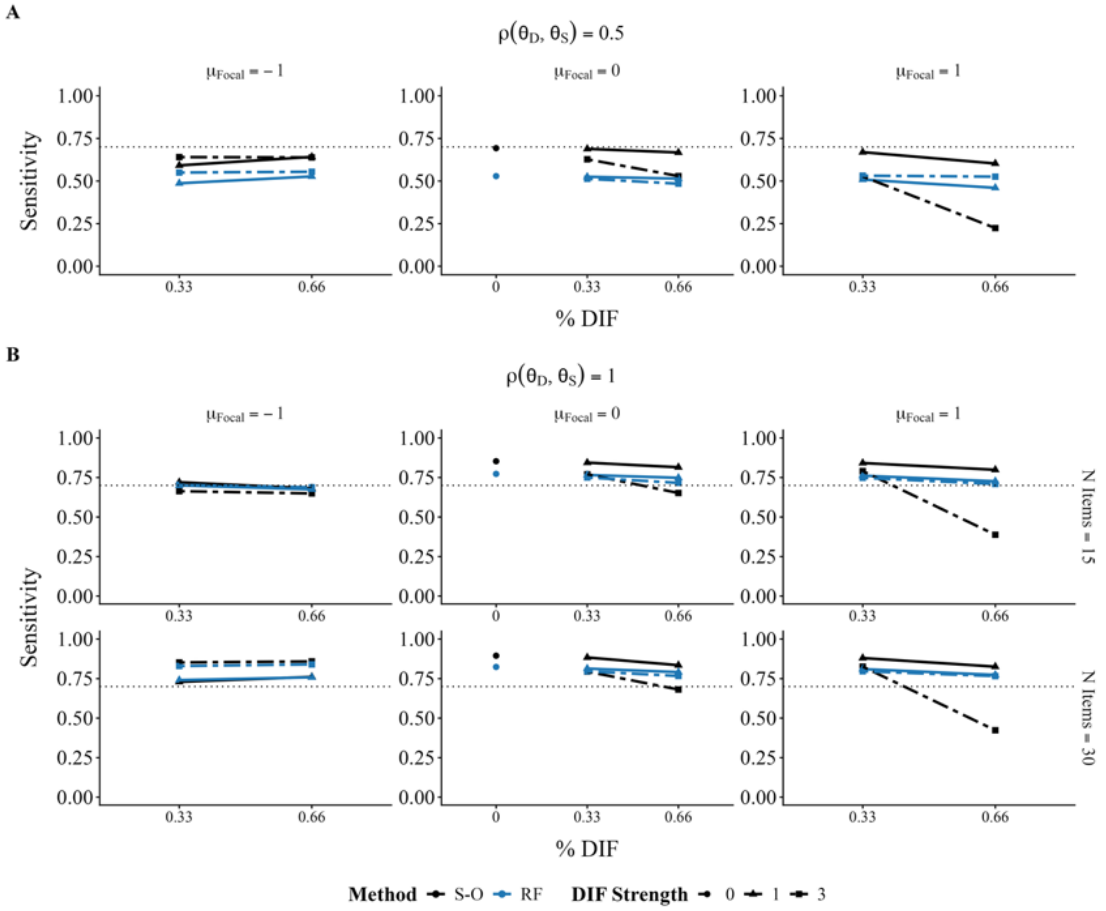
Regression Tree Predicting Classification Specificity of Second-Order IRT and RF Across Simulation Conditions.



Note: S-O = second-order IRT approach; RF = random forest; dim = number of dimensions in the scale; pDIF = percent of items that contained DIF; bStrength = strength of DIF in the *b* parameter; impact value indicates the μ of the θ distribution of the focal group (μ for the reference group distribution was always zero). Yellow terminal nodes indicate conditions where second-order IRT and RF classifications were more specific, on average, than a base rate of 0.7. Blue terminal nodes indicate where RF average classification specificity was worse than the base rate, but better than second-order IRT classification specificity. Pink terminal nodes indicate where RF average classification specificity was higher than some, but not all, second-order IRT nodes within the same branch.

Figure 15.

Line Graphs of Classification Sensitivity of the Second-Order IRT (S-O) and Random Forest (RF) Methods Under Multidimensional Scale Conditions as a Function of DIF Percentage, DIF Strength, and Impact.



Note: Panel A displays results when measurement error is present in the scale, collapsed across number of items. Panel B displays results when no measurement error is present, presented separately for 15-item and 30-item tests. In both panels, columns correspond to impact (μ_{Focal} = -1, 0, 1; μ_{Ref} was always zero). The proportion of DIF items (pDIF) is displayed on the x-axis (0.33 and 0.66), and line type indicates bStrength (0, 1, 3). The horizontal dotted line represents an acceptable sensitivity threshold of 0.70. S-O = second-order IRT approach; RF = random forest; dim = number of dimensions in the scale; pDIF = percent of items that contained DIF; bStrength = strength of DIF in the *b* parameter; impact value indicates the μ of the θ distribution of the focal group (μ for the reference group distribution was always zero).

Figure 16

Sobol MDA Variable Importance Measures for Predicting Sensitivity of Classifications Using Either the RF or Second-Order IRT Approach.

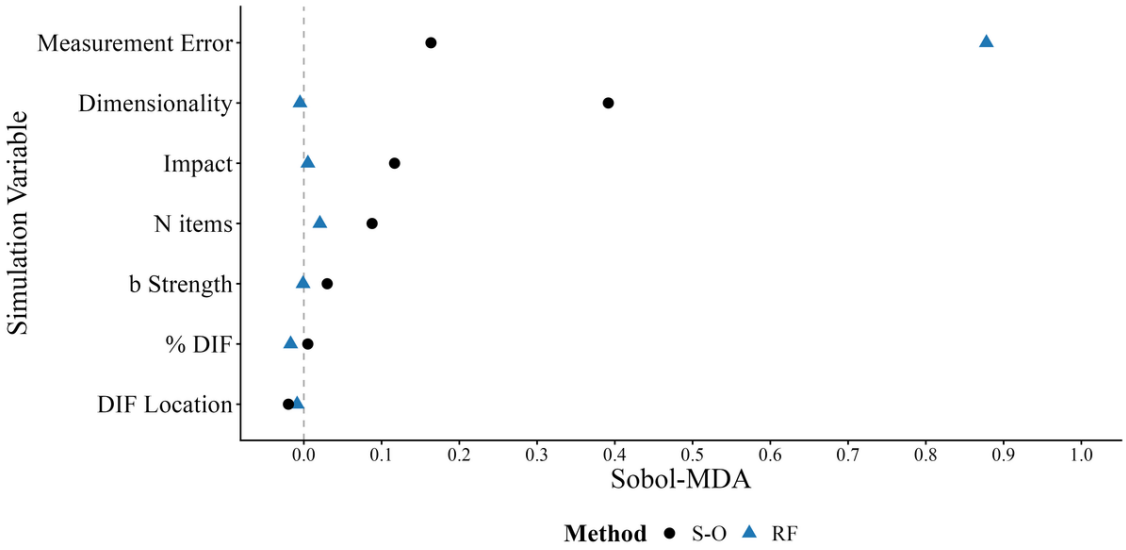
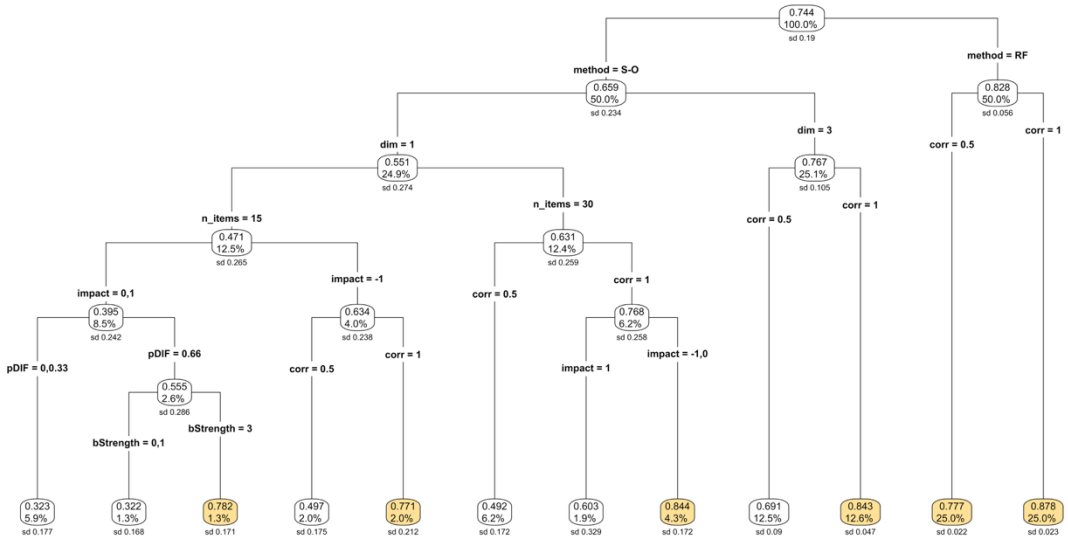


Figure 17

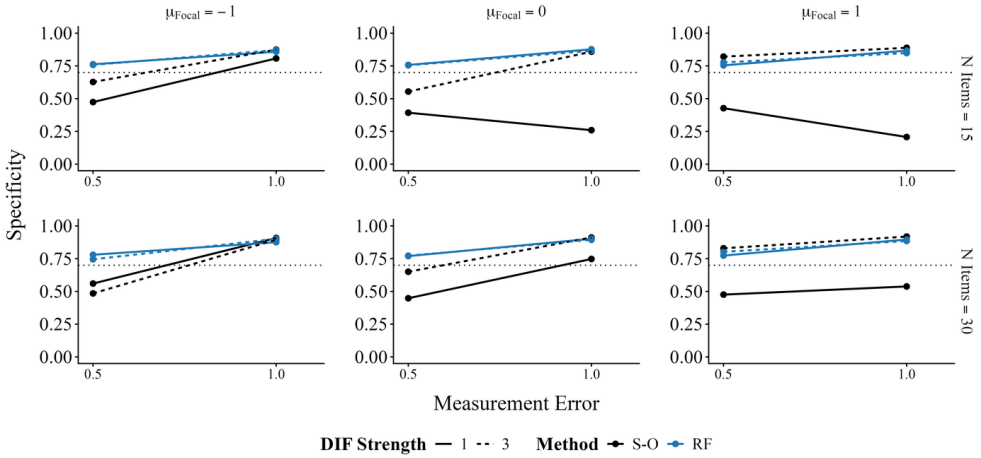
Regression Tree Predicting Classification Specificity of Second-Order IRT and RF Across Simulation Conditions.



Note. S-O = second-order IRT approach; RF = random forest; dim = number of dimensions in the scale; pDIF = percent of items that contained DIF; bStrength = strength of DIF in the *b* parameter; impact value indicates the μ of the θ distribution of the focal group (μ for the reference group distribution was always zero). Yellow terminal nodes indicate conditions where second-order IRT and RF classifications were more specific, on average, than an acceptable base rate of 0.7.

Figure 18

Line Graphs of Classification Specificity for Second-Order IRT (S-O) and Random Forest (RF) Approaches in Unidimensional Conditions where Average Second-Order IRT-Based Classification Performance was Acceptable



Note: Line graphs of mean classification specificity for second-order IRT and RF classifications as a function of the measurement error, number of items (line type), and impact. In all panels, the dotted horizontal line demarcates an acceptable specificity of 0.70, and black and blue colors correspond to second-order IRT and RF methods, respectively.

Figure 19

Sobol MDA Variable Importance Measures for Predicting Specificity of Classifications Using Either the RF or Second-Order IRT Approach.

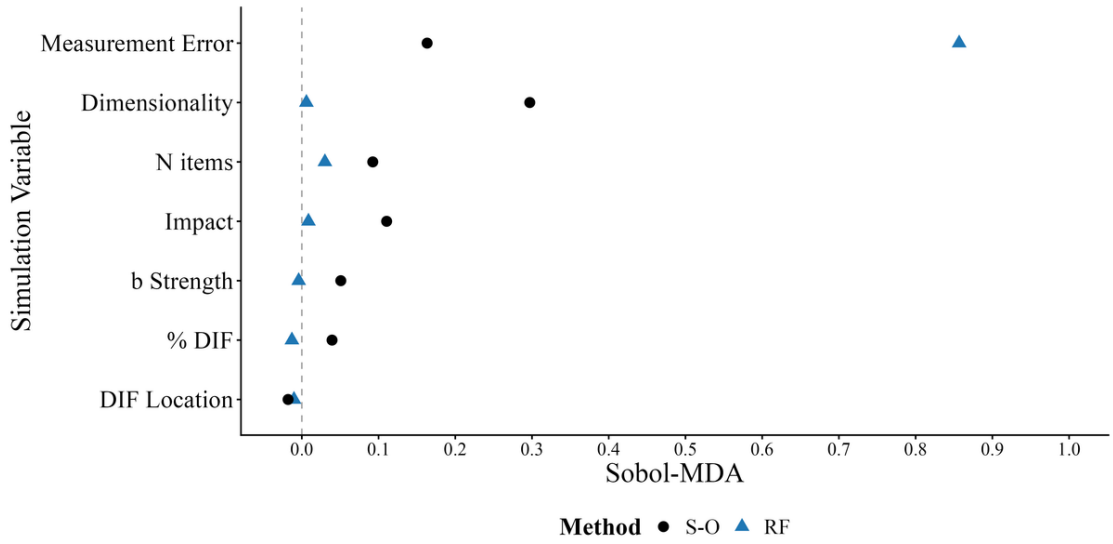
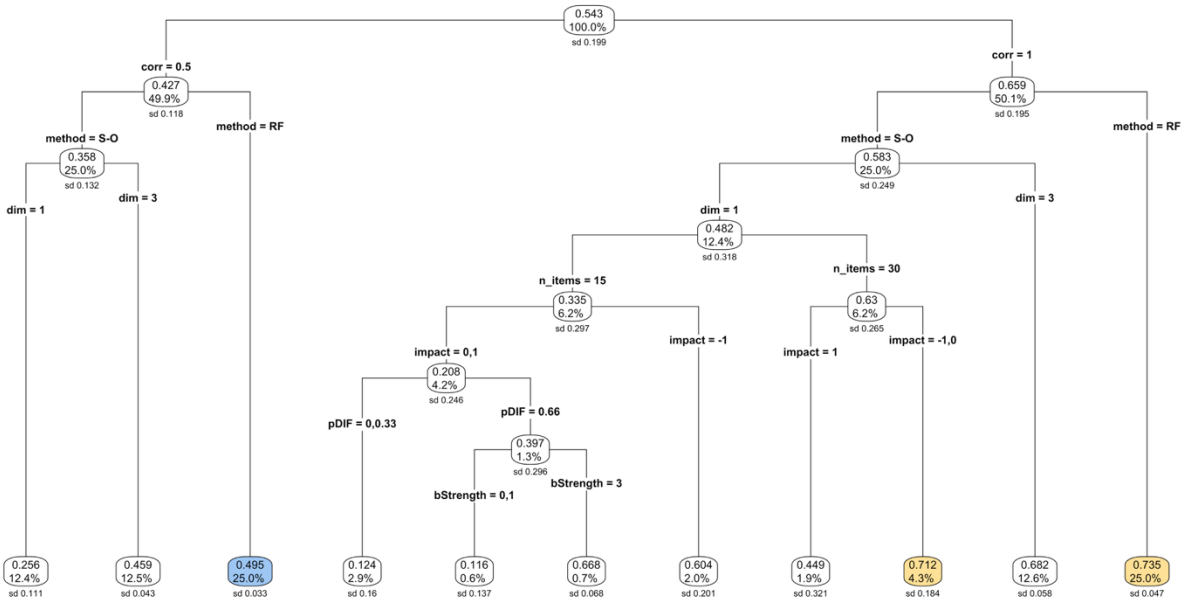


Figure 20

Regression Tree Predicting Classification Precision of Second-Order IRT and RF Across Simulation Conditions



Note: S-O = second-order IRT approach; RF = random forest; dim = number of dimensions in the scale; pDIF = percent of items that contained DIF; bStrength = strength of DIF in the *b* parameter; impact value indicates the μ of the θ distribution of the focal group (μ for the reference group distribution was always zero). Yellow terminal nodes indicate conditions where second-order IRT and RF classifications had an acceptable precision (M > 0.70). Blue terminal nodes indicate conditions where second-order IRT and RF classifications did not have an acceptable precision, but RF outperformed second-order IRT.

Figure 21

Sobol MDA Variable Importance Measures for Predicting Precision of Classifications Using Either the RF or Second-Order IRT Approach.

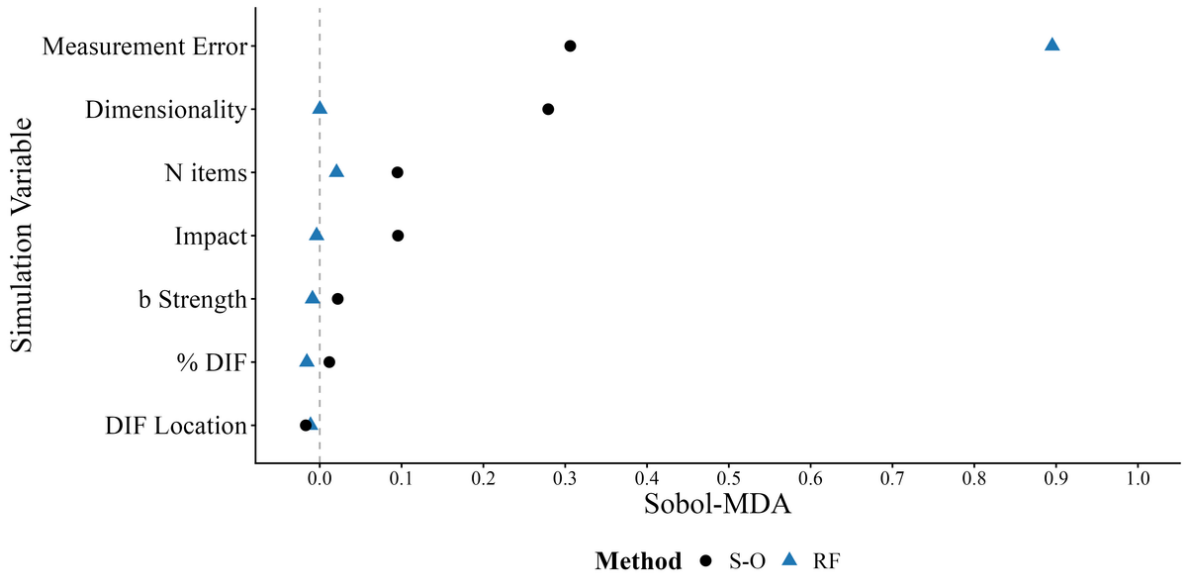
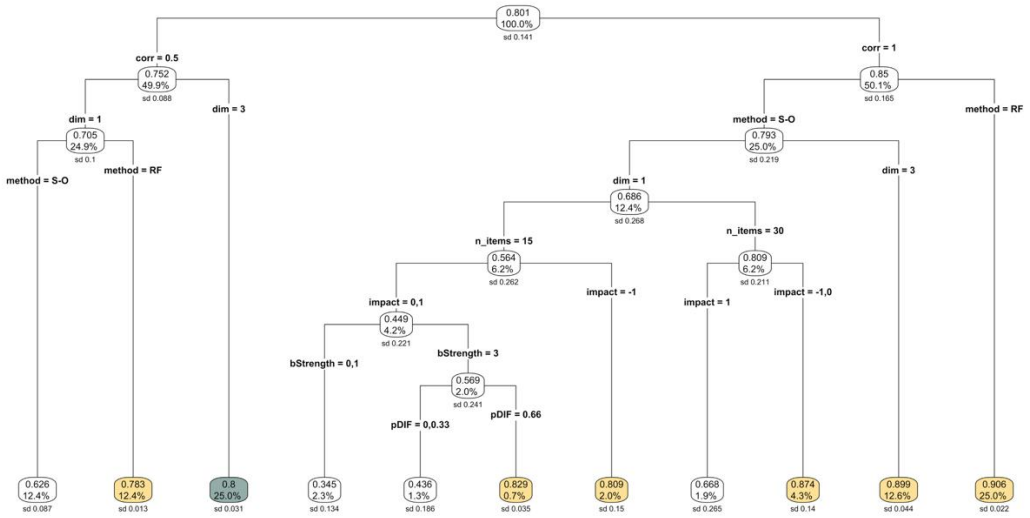


Figure 22.

Regression Tree Predicting the Negative Predictive Value of Second-Order IRT and RF Across Simulation Conditions



Note: S-O = second-order IRT approach; RF = random forest; dim = number of dimensions in the scale; pDIF = percent of items that contained DIF; bStrength = strength of DIF in the *b* parameter; impact value indicates the μ of the θ distribution of the focal group (μ for the reference group distribution was always zero). Yellow terminal nodes indicate conditions where second-order IRT and RF classifications had an acceptable NPV (M > 0.70). Green terminal nodes did not contain method as a split in their branch.

Figure 23

Sobol MDA Variable Importance Measures for Predicting NPV of Classifications Using Either the RF or Second-Order IRT Approach.

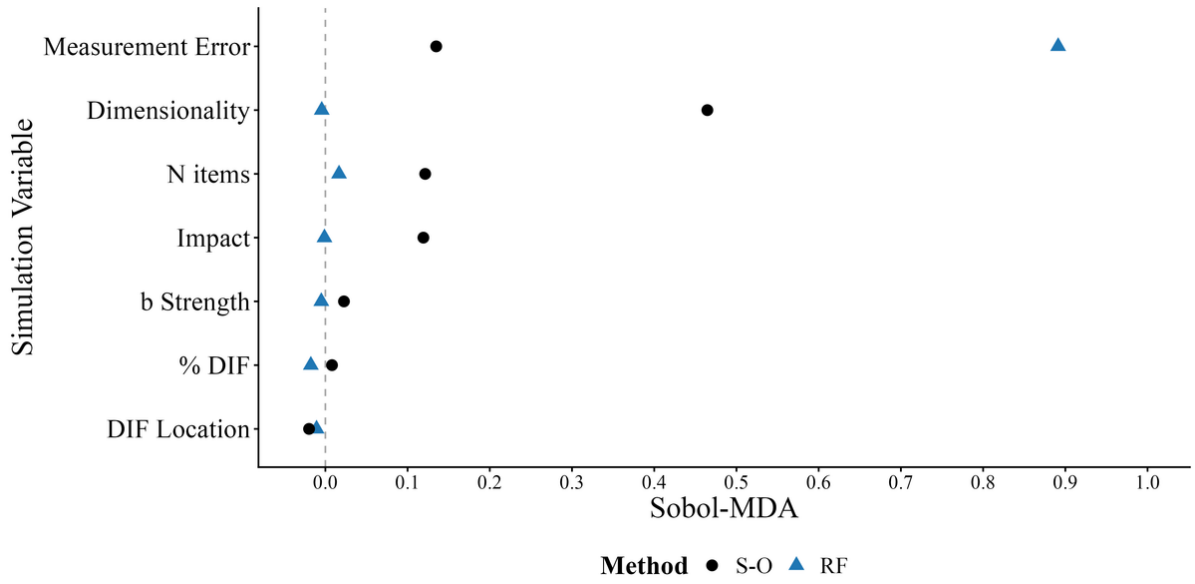
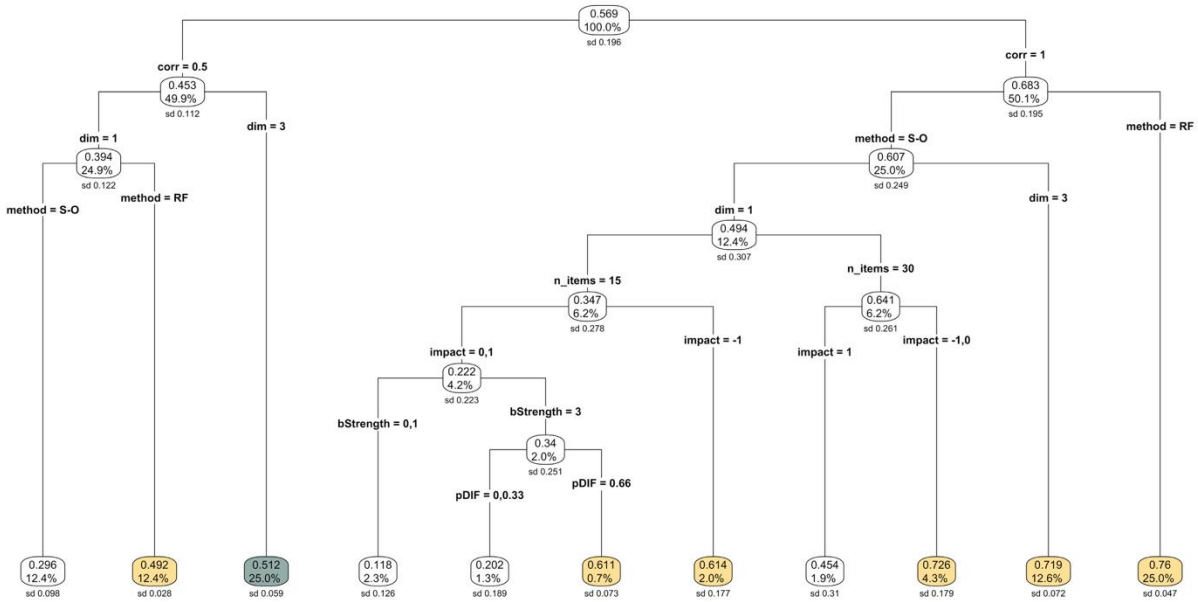


Figure 24

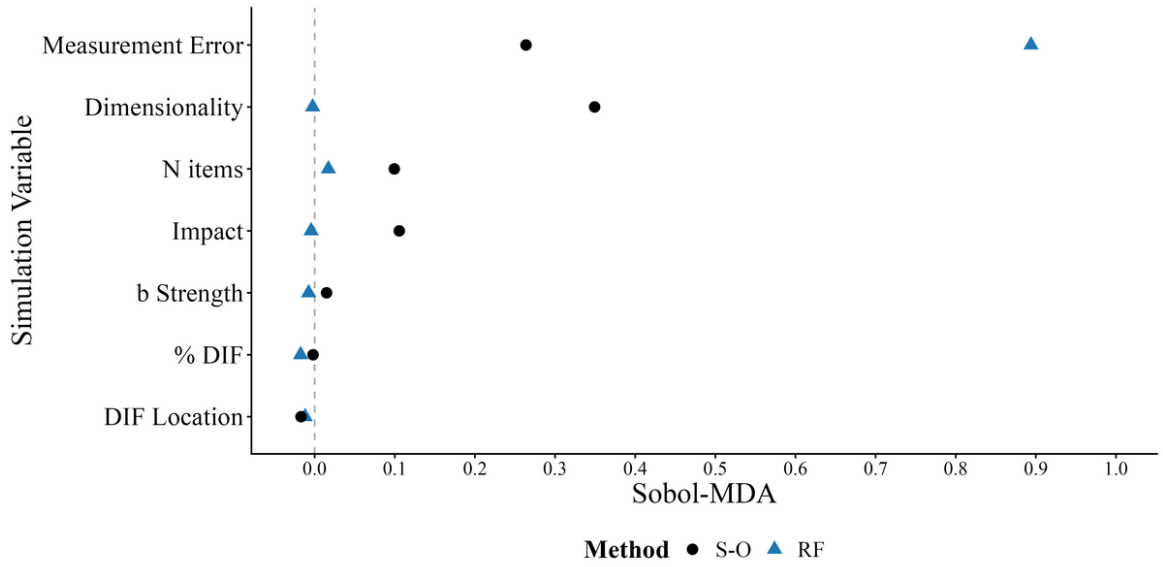
Regression Tree Predicting F1 Scores of Second-Order and RF Approaches Across Simulation Conditions



Note: S-O = second-order IRT approach; RF = random forest; dim = number of dimensions in the scale; pDIF = percent of items that contained DIF; bStrength = strength of DIF in the *b* parameter; impact value indicates the μ of the θ distribution of the focal group (μ for the reference group distribution was always zero). Yellow terminal nodes indicate conditions where second-order IRT and RF classifications had F1s that exceed the base rate (M > 0.46; calculated via the assumption that all people were placed into class 1). Green terminal nodes did not have a parent node that split on method.

Figure 25

Sobol MDA Variable Importance Measures for Predicting F1 Scores of Classifications Using Either the RF or Second-Order IRT-Based Approach.



Appendix A

Table S1. Marginal Means and Standard Deviation of Performance of Each Classification Method and DIF Group Across All Conditions

	Marginal Mean (SD)					
Method	Accuracy	Sensitivity	Specificity	Precision	NPV	F1
Baseline IRT						
Full	0.75 (0.10)	0.78 (0.12)	0.74 (0.11)	0.58 (0.13)	0.88 (0.06)	0.66 (0.12)
Reference	0.76 (0.10)	0.80 (0.12)	0.75 (0.12)	0.58 (0.13)	0.90 (0.07)	0.66 (0.14)
Focal	0.71 (0.11)	0.78 (0.17)	0.73 (0.16)	0.58 (0.13)	0.83 (0.18)	0.56 (0.22)
IRT_{ED}						
Full	0.71 (0.02)	0.07 (0.09)	0.99 (0.03)	0.84 (0.16)	0.71 (0.02)	0.14 (0.13)
Reference	0.63 (0.16)	0.03 (0.06)	1.00 (0.01)	0.92 (0.14)	0.62 (0.17)	0.10 (0.12)
Focal	0.74 (0.05)	0.09 (0.11)	0.99 (0.03)	0.83 (0.18)	0.74 (0.06)	0.17 (0.15)
SO IRT						
Full	0.63 (0.22)	0.57 (0.23)	0.66 (0.23)	0.47 (0.23)	0.75 (0.18)	0.50 (0.23)
Reference	0.59 (0.24)	0.35 (0.29)	0.72 (0.36)	0.63 (0.32)	0.59 (0.24)	0.36 (0.26)
Focal	0.65 (0.22)	0.64 (0.29)	0.65 (0.21)	0.44 (0.24)	0.81 (0.18)	0.52 (0.26)
IRT_{multi}						
Full	0.51 (0.19)	0.28 (0.25)	0.61 (0.22)	0.25 (0.24)	0.65 (0.15)	0.26 (0.24)
Reference	0.41 (0.20)	0.31 (0.25)	0.50 (0.34)	0.35 (0.30)	0.49 (0.23)	0.24 (0.18)
Focal	0.55 (0.20)	0.28 (0.31)	0.65 (0.21)	0.23 (0.25)	0.69 (0.16)	0.24 (0.27)
RF						
Full	0.77 (0.08)	0.65 (0.15)	0.83 (0.06)	0.62 (0.13)	0.85 (0.06)	0.63 (0.14)
Reference	0.76 (0.10)	0.46 (0.25)	0.87 (0.16)	0.74 (0.19)	0.73 (0.13)	0.53 (0.23)
Focal	0.78 (0.09)	0.68 (0.18)	0.81 (0.07)	0.57 (0.14)	0.87 (0.07)	0.62 (0.16)

Note: IRT_{ED} = item response theory using a Euclidean distance threshold; SO = second-order item response theory using a general factor threshold; IRT_{multi} = item response theory using a dimension specific threshold for each dimension; RF = random forest; Marginal means and standard deviations collapsed across conditions within classification method and sub-group.

Table S2. Marginal Means and Standard Deviations of Performance of Each Classification Method and DIF Group in Multidimensional Conditions (i.e., D = 3).

	Marginal Mean (SD)					
Method	Accuracy	Sensitivity	Specificity	Precision	NPV	F1
Baseline IRT						
Full	0.75 (0.09)	0.77 (0.11)	0.73 (0.10)	0.57 (0.11)	0.88 (0.05)	0.65 (0.10)
Reference	0.76 (0.09)	0.79 (0.11)	0.75 (0.11)	0.57 (0.11)	0.89 (0.07)	0.65 (0.12)
Focal	0.71 (0.12)	0.78 (0.16)	0.72 (0.16)	0.57 (0.11)	0.83 (0.18)	0.55 (0.24)
IRT_{ED}						
Full	0.72 (0.02)	0.10 (0.10)	0.99 (0.03)	0.86 (0.15)	0.72 (0.02)	0.17 (0.14)
Reference	0.63 (0.18)	0.04 (0.07)	1.00 (0.01)	0.94 (0.12)	0.62 (0.19)	0.12 (0.13)
Focal	0.75 (0.06)	0.12 (0.13)	0.99 (0.03)	0.85 (0.17)	0.75 (0.06)	0.19 (0.16)
SO IRT						
Full	0.74 (0.09)	0.69 (0.16)	0.77 (0.10)	0.57 (0.12)	0.85 (0.06)	0.62 (0.13)
Reference	0.73 (0.14)	0.33 (0.30)	0.93 (0.09)	0.83 (0.18)	0.70 (0.16)	0.43 (0.30)
Focal	0.75 (0.10)	0.80 (0.12)	0.72 (0.12)	0.53 (0.13)	0.91 (0.06)	0.63 (0.12)
IRT_{multi}						
Full	0.50 (0.08)	0.12 (0.11)	0.67 (0.12)	0.14 (0.14)	0.64 (0.04)	0.12 (0.10)
Reference	0.36 (0.13)	0.25 (0.22)	0.50 (0.25)	0.23 (0.21)	0.49 (0.19)	0.19 (0.15)
Focal	0.54 (0.09)	0.09 (0.12)	0.72 (0.12)	0.11 (0.15)	0.68 (0.07)	0.08 (0.11)
RF						
Full	0.77 (0.07)	0.64 (0.13)	0.83 (0.05)	0.61 (0.11)	0.84 (0.06)	0.63 (0.12)
Reference	0.76 (0.10)	0.43 (0.25)	0.87 (0.16)	0.74 (0.18)	0.72 (0.14)	0.51 (0.24)
Focal	0.78 (0.08)	0.67 (0.17)	0.81 (0.06)	0.56 (0.13)	0.87 (0.06)	0.61 (0.14)

Note: IRT_{ED} = item response theory using a Euclidean distance threshold; SO = item response theory using a second-order general factor threshold; IRT_{multi} = item response theory using a dimension specific threshold for each dimension; RF = random forest; Marginal means and standard deviations collapsed across conditions within classification method and sub-group.

Table S3. Random Forest Cross-Validated Fit Measures and Sobol MDA Variable Importance Measures Across Second-Order IRT-, MGRM-, and RF-Based Classification Performance Metrics.

		Fit Measures				Variable Importance						
		Train		Test		Train Sobol MDA						
Outcome	Method	MSE	R2	MSE	R2	Dim	Corr	Items	% DIF	Local	*b*	Impact
Accuracy												
	SO IRT	0.016	0.677	0.016	0.676	0.369	0.189	0.103	0.012	–	0.025	0.102
	MGRM	0.017	0.511	0.017	0.512	0.250	0.188	0.122	0.002	–	0.027	0.128
	RF	< 0.001	0.958	< 0.001	0.958	0.001	0.902	0.018	–	–	–	–
Sensitivity												
	SO IRT	0.018	0.673	0.018	0.672	0.392	0.164	0.088	0.005	–	0.03	0.117
	MGRM	0.017	0.718	0.017	0.716	0.545	0.103	0.076	–	–	0.020	0.083
	RF	0.001	0.944	0.001	0.944	–	0.878	0.021	–	–	–	0.005
Specificity												
	SO IRT	0.020	0.638	0.020	0.637	0.297	0.163	0.092	0.039	–	0.051	0.111
	MGRM	0.023	0.528	0.023	0.529	0.303	0.151	0.089	0.031	–	0.051	0.123
	RF	< 0.001	0.921	< 0.001	0.920	0.006	0.857	0.030	–	–	–	0.009
Precision												
	SO IRT	0.015	0.713	0.015	0.711	0.279	0.306	0.095	0.012	–	0.022	0.096
	MGRM	0.019	0.664	0.020	0.662	0.375	0.164	0.074	0.021	–	0.053	0.100
	RF	0.001	0.955	0.001	0.955	–	0.895	0.021	–	–	–	–
NPV												
	SO IRT	0.010	0.678	0.010	0.678	0.465	0.135	0.121	0.008	–	0.023	0.119
	MGRM	0.011	0.505	0.011	0.506	0.229	0.164	0.173	0.006	–	0.024	0.161
	RF	< 0.001	0.952	< 0.001	0.952	–	0.891	0.017	–	–	–	–
F1 Score												
	SO IRT	0.014	0.719	0.014	0.717	0.349	0.264	0.099	–	–	0.015	0.106
	MGRM	0.017	0.698	0.017	0.697	0.478	0.151	0.086	–	–	0.016	0.084
	RF	0.001	0.952	0.001	0.954	–	0.894	0.017	–	–	–	–

Note: SO IRT = Second-order IRT; MGRM = multidimensional graded response model IRT (IRT_{multi}); RF = random forest; Dim = number of dimensions in the scale; Corr = correlation between gold standard and scale; Local = location of the DIF; Impact = the mean of the θ distribution for the focal group {-1, 0, 1}. The mean of the reference group θ distribution is always zero. Only non-zero, positive Sobol MDA values were retained in the table for simplicity.