Journal of Behavioral Data Science, 2026, *6* (2), 1–39.
DOI: https://doi.org/10.35566/jbds/bainmmbg

When DIF Goes Unmodeled: Assessing the Viability of Random Forest for Diagnostic Classification

Catherine M. Bain[1 [0000−0002−2767−6882]], Patrick D. Manapat[1 [0000−0003−1027−6652]], Danielle M. Manapat[1 [0000−0002−3466−0758]], Keelyn B. Brennan[1 [0009−0008−3927−5614]], and Kevin J. Grimm[2 [0000−0002−8576−4469]]

[1] Department of Psychology, University of Oklahoma, Norman, OK 73072 USA
cbain1@ou.edu, pmanapat@ou.edu
[2] Department of Psychology, Arizona State University, Tempe, AZ 85287 USA

Abstract. Psychological assessments are often used to make classification decisions (i.e., identify whether an individual meets a set of criteria for a psychological diagnosis). A psychometric approach, like item response theory (IRT), first estimates a latent trait score from item responses and then compares that score to a cut-point to determine classes. In contrast, a machine learning approach, such as random forest (RF) predicts class membership directly from item responses. While both methods show promise for diagnostic classification, their relative robustness to differential item functioning (DIF) is unclear. This study compared the classification performance of an IRT- and a RF-based approach to classification under conditions varying DIF presence and severity, along with other sample and scale characteristics via Monte Carlo simulation. Single-group IRT served as a baseline representing standard psychometric practice, as it is widely used for diagnostic classification but assumes item invariance across groups. Results indicated that when DIF was absent or minimal, both approaches yielded comparable classification metrics. However, as DIF severity increased, IRT-based classification performance declined, whereas RF maintained robust performance across conditions. These findings suggest that RF may maintain more stable classification performance than IRT-based classification when DIF is present but not explicitly accounted for in the model, making RF a viable alternative for diagnostic classification when DIF is suspected but its source or structure is unknown, unmeasured, or complex. Strengths and limitations of each approach are discussed, with particular attention to the trade-off between interpretability and classification robustness in applied assessment contexts.

Keywords: Classification · Differential item functioning · Item response theory · Random forest · Diagnostic assessment · Bias · Fairness · Machine learning · Psychometrics

1 Introduction

One goal of psychometric scales is to classify individuals, though the purpose of these classifications varies across both field and scale. In clinical psychology, diagnostic scales (e.g., the ASP; Brown, Tollefson, Dunn, Cromwell, & Filion 2001; the DMQ; Rosenthal et al. 2021) can be used to classify individuals as having, or not having, a clinically relevant level of a specific disorder. The disorder is represented as a latent trait that relates to a gold standard defined by given diagnostic criteria (e.g., DSM-5TR; American Psychiatric Association, 2022). Direct assessment of the gold standard usually requires testing and documentation of symptoms by a licensed clinician, which can be costly, time-intensive, and potentially invasive. For these reasons, the use of a gold standard may be infeasible in some contexts, and researchers may instead administer a scale to approximate the gold standard through a set of questions or items (Gibbons et al., 2013; Gonzalez, 2021).

After obtaining a participant's item responses, researchers interested in making a binary diagnostic classification (e.g., diagnosed vs. not diagnosed) often take one of two approaches to create these classifications. The first requires the researcher to use a parametric approach to determine a single continuous score estimating the individual's level of the latent trait and then determine a cut-point such that only individuals with scores falling above the cut-point are placed in the diagnosed class. The second approach utilizes a non-parametric algorithm to predict the probability of diagnosis directly from the item responses and then classifies respondents based on whether the probability of diagnosis is above a given cut-point. This paper focuses on the use of item response theory (IRT) as the parametric approach and the use of random forest (RF; Breiman, 2001) with a Bayes classifier as the nonparametric approach.

Although both IRT and RF have been shown to perform well in diagnostic classification contexts, notably less research exists examining the use of RF (e.g., Giannouli & Kampakis, 2024; Ohiri, Momoh, Christopher, Ikeanumba, & Benedict, 2024) than IRT (e.g., Carvalho, Costa, Otoni, & Junqueira, 2019; R. Liu, Huggins-Manley, & Bulut, 2018; Rosenthal et al., 2021), with even less research comparing IRT-based classification and RF-based classification directly. One previous study found that IRT-based scoring outperformed nonparametric machine learning (ML) approaches (Gonzalez, 2021). To our knowledge, no studies have examined how differential item functioning (DIF) affects this comparison.

One key challenge for classification performance is DIF, which occurs when participants with the same level of a latent trait respond differently to a given item. Many published diagnostic scales contain DIF (e.g., Golay, Abrahamyan Empson, Mebdouhi, Conus, & Alameda, 2023; Spann, Cicero, Straub, Pellegrini, & Kerns, 2024; Wardell, Cunningham, Quilty, Carter, & Hendershot, 2020); therefore, the assumption that scales are free of DIF is unlikely to hold in practice. For example, previous research suggests that individuals with autism respond differently to misophonia questionnaires than those without autism (Williams, Cascio, & Woynaroski, 2022). Misophonia is a condition characterized by a decreased tolerance for specific auditory stimuli (i.e., triggers), followed by strong, negative

emotional, physiological, and behavioral responses, which often co-occurs with other neuropsychiatric conditions like autism (Schröder, Vulink, & Denys, 2013). One study illustrated that those with autism more readily indicate that their sound sensitivities interfere with their social life than those without (Williams et al., 2022). Score differences on these items, therefore, represent co-morbidity-based differences in the pattern of responding rather than true differences in misophonia levels. This over-endorsement leads to individuals with autism having greater observed scores for misophonia than those without, even when their true levels of misophonia are identical. In this case, these items exhibit DIF across co-morbidity groups (Williams et al., 2022). The presence of DIF can decrease classification performance, making it critical to evaluate how robust different classification methods are when DIF exists but cannot be detected.

If DIF is left unaddressed, IRT scores can be biased (Paulsen, Svetina, Feng, & Valdivia, 2020), making IRT-based classifications less accurate. Traditional DIF detection methods typically involve researchers predefining the groups across which DIF is thought to occur (Battauz, 2019; Woods, Cai, & Wang, 2013). In cases where DIF manifests in complex subpopulations (e.g., LatinX women whose highest degree obtained is a high school diploma) or unmeasured factors, like a participant's primary language, this approach is problematic.

When DIF is suspected but groups are unknown, mixture modeling approaches like mixture IRT and factor mixture modeling offer psychometric alternatives that explicitly model latent group structure. However, these methods require researchers to specify the number of classes and the structure of classes a priori. Additionally, mixture IRT can face identification challenges when groups have identical trait distributions but differ only in item parameters (i.e., DIF without impact), particularly when DIF affects relatively few items or is moderate in magnitude (Sen & Cohen, 2024). In our study, many conditions involve no impact (identical generating trait distributions across groups), which may make mixture IRT implementations particularly challenging. However, RF can flexibly model high-dimensional interactions between items, potentially making classifications more robust to unmodeled DIF. As such, we hypothesize that RF will maintain classification performance in the presence of unmodeled DIF by capturing complex non-linear patterns in the data, making it a promising alternative for classification with unmodeled DIF (Breiman, 2001). This study focuses on the comparison between single-group IRT-based classification (without group-specific parameters) and RF-based classification, representing a common applied scenario where DIF may be left unmodeled (e.g., the grouping variable is either unmeasured, unknown, or too complex to model). We use single-group IRT as a baseline representing psychometric practice when researchers suspect DIF exists but lack information about potential grouping variables and therefore likely leave DIF unmodeled. For simplicity, we use IRT throughout the remainder of the manuscript to refer to single-group IRT.

The goals of this study were to (1) evaluate how classification accuracy changes when DIF is present but unmodeled in IRT-based classification approaches, (2) evaluate whether RF-based classification is more robust to unmodeled DIF than

IRT, and (3) analyze the impact of various data characteristics on the performance of IRT and RF methods. The remainder of this paper is structured as follows. First, we outline the goals of diagnostic assessment. Then, we describe IRT- and RF-based classification methods. Next, the details of the Monte Carlo simulation study are provided, and results are discussed. This is followed by an empirical example illustrating the use of both IRT and RF on data related to misophonia. Finally, we conclude with a discussion of our findings and practical recommendations regarding the strengths and limitations of IRT and RF approaches to classification.

1.1 Diagnostic Assessment

To better understand the process of diagnostic assessment, consider diagnosing misophonia using the Duke Misophonia Questionnaire (DMQ; Rosenthal et al., 2021). The DMQ generates a score ($\theta_{Assessment}$) reflecting the severity of misophonia. A threshold (T) is then applied to classify individuals as having misophonia ($D_{Assessment} = 1$) or not ($D_{Assessment} = 0$).

$$\begin{aligned} D_{Assessment} = 1 &\iff \theta_{Assessment} \geq T \\ D_{Assessment} = 0 &\iff \theta_{Assessment} < T \end{aligned} \tag{1}$$

Researchers may then determine the diagnostic ability of this scale by comparing these classifications to those obtained from the gold standard (e.g., the misophonia questionnaire severity item; Wu, Lewin, Murphy, & Storch, 2014). Ideally, scores from the DMQ ($\theta_{Assessment}$) will correlate strongly with scores from the gold standard; however, prior research indicates that the strength of this correlation may vary (Bain, Norris, Conley, Manapat, & Ethridge, 2025). Moreover, there is often a direct relationship between classification performance and the strength of the correlation between $\theta_{Assessment}$ and the gold standard (Gonzalez, 2021). As such, misclassification is inevitable (Figure 1). Researchers aim to maximize both sensitivity (true positive rate) and specificity (true negative rate), but sensitivity and specificity have an indirect relationship, such that as one increases, the other decreases (Gonzalez, 2021; Youngstrom, 2013). Decreasing T would increase sensitivity but decrease specificity. Increasing T would have the opposite effect (low sensitivity and high specificity). Decision theory techniques, which allow researchers to manually weight the importance of sensitivity and specificity, are recommended to determine the optimal cut-point (Smits, Smit, Cuijpers, & De Graaf, 2007). The following sections detail IRT and RF approaches to diagnostic assessment.

Item Response Theory IRT is a collection of latent variable models that seek to uncover the underlying process that influences responses to observed variables (Edwards, 2009). They model the probability of endorsing a particular response on an item as a function of the latent trait (θ) and item properties (Embretson & Reise, 2000; Lord, 2012). These properties include the difficulty of the item

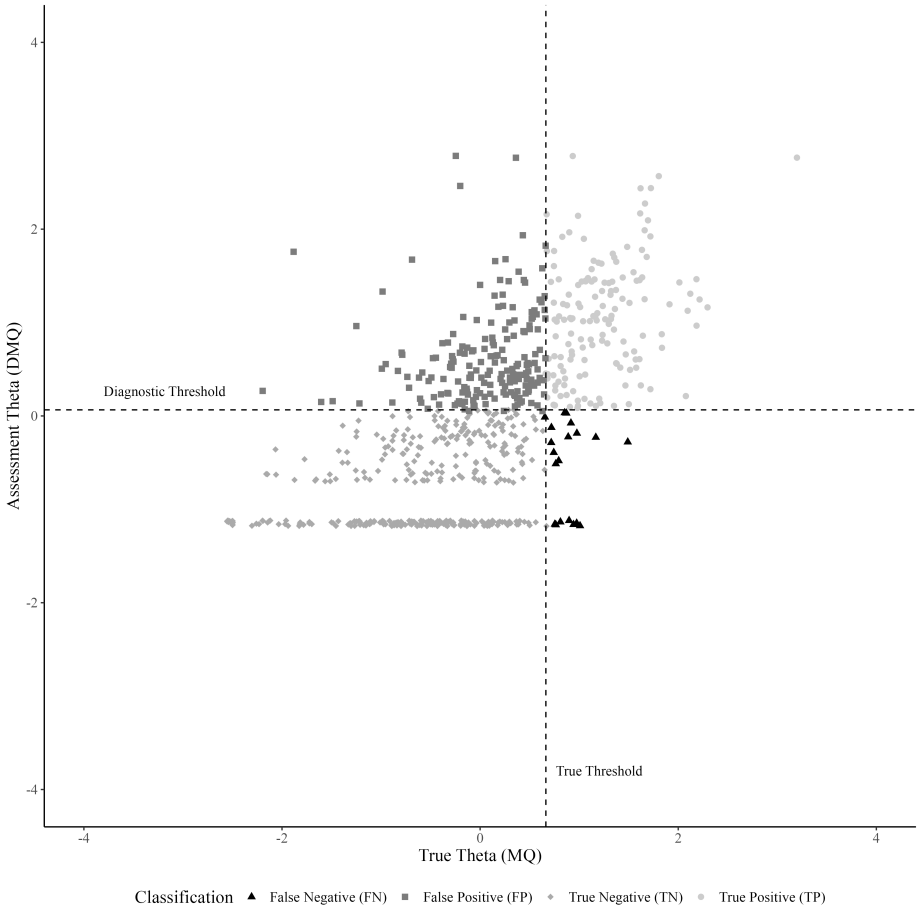

Figure 1. Scatterplot illustrating thresholds to classify observed values. *Note:* The x-axis represents the gold standard θ values (labeled true theta), while the y-axis represents the estimated θ values from the assessment. Two reference lines divide the plot (the diagnostic threshold and the true threshold). Each observation is classified into one of four categories: false negative, false positive, true negative, and true positive.

(i.e., threshold; b) and the item's ability to discriminate between individuals (i.e., slope; a). Individual scores (θ) are estimated using an individual's responses and these item parameters (Edwards, 2009). IRT calibration is advantageous as it provides standardized, interpretable scores with greater variability than summed scores, conditional standard errors, and straightforward score equating (De Ayala, 2009; Embretson & Reise, 2000).

Samejima's (1997) graded response model (GRM) is the most widely used IRT model for scales with items that consist of two or more ordinal response

options (e.g., Likert scales). The GRM is expressed as:

$$P(x_j = c \mid \theta) = \frac{1}{1 + \exp[-a_j(\theta - b_{cj})]} - \frac{1}{1 + \exp[a_j(\theta - b_{(c+1)j})]}, \quad (2)$$

which represents the probability of endorsing response option c given the latent trait (θ), resulting in the observed response (x_j) to item j. For more on the GRM and IRT in general, see work by Edwards (2009) or Embretson and Reise (2000).

The discrimination parameter (a) represents the slope of item j and indicates how strongly the item relates to the latent construct (i.e., a steeper slope reflects a stronger relationship). The threshold parameters (b) define the level of the latent trait required to endorse response c for item j, with higher thresholds indicating higher trait levels (Manapat, Edwards, MacKinnon, Poldrack, & Marsch, 2021; Samejima, 1997). The GRM, like other parametric IRT models, assumes local independence, monotonicity, and unidimensionality. These assumptions enable interpretable item and person parameters but impose more structural constraints than nonparametric methods.

Differential Item Functioning's Effects on Classification While DIF is traditionally studied as a measurement bias problem, its presence can also affect classification accuracy. An item is said to exhibit DIF if two respondents from distinct subgroups (e.g., gender, primary language, etc.) have equal levels of θ but different probabilities of endorsing a given response category for that item (Edwards, 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 2A and 2B, respectively). Uniform DIF occurs when the item uniformly requires a larger or smaller amount of the trait for one group, reflected by a difference in difficulty parameters but not in discrimination parameters ($a_{Focal} = a_{Reference}$ and $b_{Focal} \neq b_{Reference}$). 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., $a_{Focal} \neq a_{Reference}$). When classifying individuals using scores from scales with DIF, these group differences in item functioning can lead to systematic misclassification, particularly for members of the focal group (Gonzalez, Georgeson, Pelham, & Fouladi, 2021; Lai, Richardson, & Wa Mak, 2019; Paulsen et al., 2020).

DIF can affect the overall scale score, resulting in differential test functioning (DTF; Chalmers, Counsell, & Flora, 2016; Yavuz Temel, 2023) when biased items disproportionately influence measurement for certain groups. For example, if items favor native English speakers, the scale score could lead to systematic misclassification for non-native speakers (Hambleton, Swaminathan, & Rogers, 2011; Zumbo, 1999). However, DIF does not always result in DTF (e.g., if only a few items have DIF or if the DIF items have low discrimination parameters). This underscores the importance of evaluating a scale for both DIF and DTF (Camilli & Shepard, 1994).

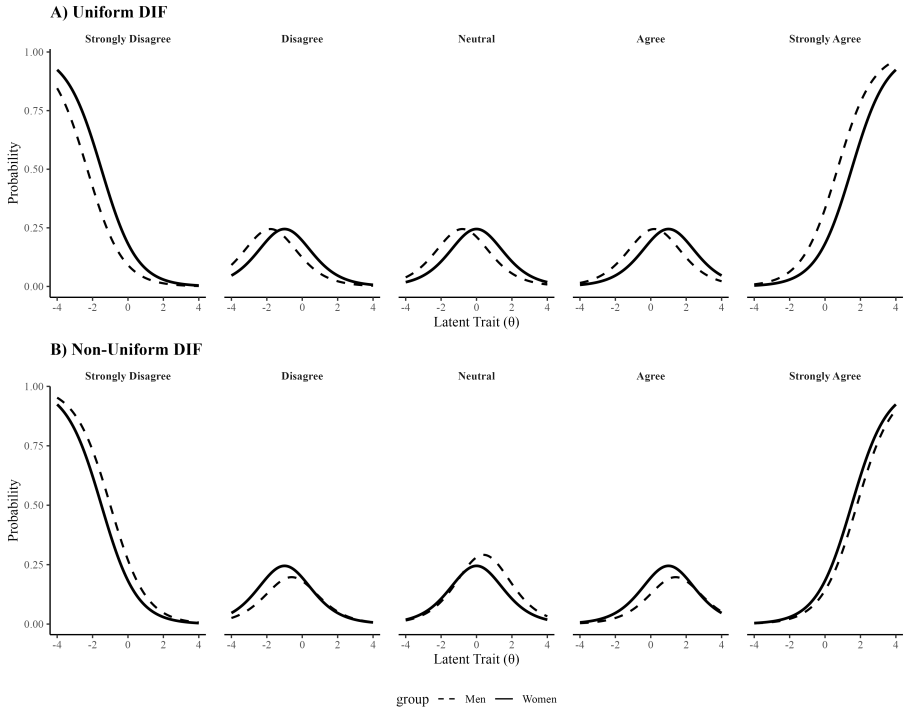

Figure 2. Category characteristic curves for a 5-point Likert scale item exhibiting differential item functioning (DIF). *Note:* Panel A shows uniform DIF, where the probability curves for men (dashed lines) are consistently shifted relative to women (solid lines) across all response categories. Panel B shows non-uniform DIF, where the differences between men and women vary across the latent trait continuum. Trace lines are typically plotted on a single x-axis but are separated here for readability.

Random Forest Random forests (RF) are an ensemble learning approach that improve prediction accuracy by aggregating the predictions of many decision trees built on bootstrapped samples of the data (Breiman, 2001). The RF additionally introduces randomness by selecting a subset of potential predictors at each split (Breiman, 2001; Jacobucci, Grimm, & Zhang, 2023). Specifically, for each of the decision trees, a bootstrap sample of size N is drawn from the data, which introduces sample-level randomness. A decision tree is then grown on each sample by recursively splitting the data using a randomly selected subset of m predictor variables at each node, which introduces variable-level randomness. This randomness helps reduce variance and improve generalizability to new samples. After all trees are trained, predictions are aggregated by averaging (for regression) or majority vote (for classification) across all trees.

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). Other RF tuning parameters include maximum tree depth, minimum node size, and impurity criterion (e.g., Gini or entropy). While these typically have less influence than the number of trees and predictors per split, they can affect model complexity and computational efficiency and may be adjusted depending on the application. For a full list of parameters, see 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. In this approach, larger weights are assigned to observations from minority classes during tree construction, which alters both the bootstrap sampling and the splitting criterion. By penalizing misclassification of minority-classes more heavily, weighting directly increases their influence on decision boundaries and reduces the tendency of RF to favor the majority class. This makes weighting particularly useful in diagnostic applications, where the minority class often represents the group of individuals in need of clinical intervention.

This study proposes using RF when DIF is present but unmodeled. RF has been shown to handle complex, high-dimensional data effectively, making RF suitable for psychometric applications where item responses may exhibit nonlinear relationships or interactions (Strobl, Hothorn, & Zeileis, 2009). Single-group IRT models assume item-invariance across subgroups (i.e., no DIF); however, the IRT framework can be extended to model DIF through multiple-group approaches when the cause of DIF is known. RF makes no such assumptions about item functioning. As a nonparametric method, RF can flexibly adapt to group-specific response patterns through its ensemble of trees. This flexibility may allow RF to maintain classification accuracy even when items function differently across subgroups, making it a potentially robust classification approach when DIF is unmodeled (e.g., DIF is suspected, but the source or structure of DIF is unknown).

1.2 Present Study

The current study aims to compare the performance of RF and IRT for classification in conditions when DIF is absent and present. We hypothesize that RF will demonstrate greater classification robustness in the presence of DIF because it can capture complex response patterns without assuming item-invariance. A secondary aim was to assess the impact of DIF on IRT-based classification performance. For example, we examine whether single-group IRT can maintain acceptable classification accuracy when DIF is minimal. We compare IRT, RF with predicted probabilities, and RF with predicted classifications in terms of accuracy, sensitivity, specificity, precision, negative predictive value (NPV), and F1 score (i.e., the harmonic mean of precision and sensitivity). We were also interested in the extent to which the classification performance of these models

depended on features of the data. Importantly, all approaches (IRT-based, RF probability-based, and RF classes) did not have "access" to any DIF grouping variable, reflecting what applied researchers would likely do when DIF sources are unknown, unmeasured, or too complex to model using multi-group IRT approaches. In this respect, our simulation was designed to reflect common applied settings where researchers are likely to fit a misspecified single-group model. The simulated two-group DIF structure (discussed below) represents the simplest case of a broader set of scenarios that fail to account for DIF.

2 Methods

A Monte Carlo simulation study was conducted to examine the performance of IRT and RF in classifying (i.e., diagnosing) individuals when DIF in items is not directly identified and specified. Conditions with no DIF were included to serve as a baseline comparison. All data generation and analyses were conducted in R (R Core Team, 2024), and R scripts used for the simulation have been made publicly available on GitHub (`https://github.com/cbain1/difml_open`). We manipulated the following features of the generating model: sample size of the reference group, ratio of focal group to reference group, prevalence rate of the diagnosis according to the gold standard, number of items, DIF patterns, and magnitude of DIF.

2.1 Data Generation

The reference group size was set to 400 (small) or 2000 (large). The small sample size was meant to simulate datasets that are common in psychological research (Shen et al., 2011), while the larger sample size was based on previous simulations examining DIF (e.g., Classe & Kern, 2024). The length of the scale was set to either 40 or 80 based on the lengths of published diagnostic scales (Raskin & Terry, 1988; Rosenthal et al., 2021). The prevalence rate of the disorder was set to either 12 or 30, representing prevalence rates of disorders seen in the literature (Rosenthal et al., 2021; Sims, Michaleff, Glasziou, & Thomas, 2021; Terry-McElrath & Patrick, 2018).

The ratio of focal group to reference group was set to either 1:1 (equal), 1:2, or 1:4; these values were chosen based on previous research (e.g., Classe & Kern, 2024). We chose a two-group DIF implementation that represents the simplest case for the research questions posed here. We recognize this is an "ideal" case and that the present study does not include conditions that reflect more complex DIF that is likely to be encountered in real data. However, the current simulation does reflect the approach applied researchers would likely take in situations where the grouping variable causing DIF is unmeasured or unknown.

In each simulation repetition, data were generated by a single continuous variable ($\theta_{Diagnosis}$) from a normal distribution with $\mu = 0$ and $\sigma = 1$ representing the latent variable determining the respondent's true diagnosis (Equation 1; Gonzalez, 2021). To determine a threshold ($T_{Diagnosis}$) that would ensure our

desired prevalence rate, we determined the $1 - p$ quantile of the $\theta_{Diagnosis}$ distribution, where p represents the desired prevalence rate. Using this $T_{Diagnosis}$ and Equation 1, the true diagnosis labels were generated. To examine the effects of impact (i.e., true group differences in θ), a set of additional simulations were performed manipulating the data-generating distribution for the focal group. In these conditions, $\theta_{Diagnosis}$ for the reference group simulees was generated from a normal distribution with a $\mu = 0$ and $\sigma = 1$, whereas $\theta_{Diagnosis}$ for the focal group simulees was generated from a normal distribution where $\mu = \{-1, 1\}$ and $\sigma = 1$.

The $\theta_{Diagnosis}$ was then used with the GRM (Equation 2) to generate item responses. Based on previous research, a-parameters were drawn from a normal distribution with $\mu = 1.7$ and $\sigma = 0.3$ (Hill, 2004; Manapat & Edwards, 2022). Also based on previous research, b-parameters were generated by 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). This process was repeated until a total of four b-parameters were created (e.g., $b_2 = b_1 +$ difference parameter 1).

To simulate DIF, we manipulated DIF proportion, type of DIF, strength of DIF, and whether DIF was balanced by changing the focal group parameters. DIF was set to be present in 0, 5, 40, and 70 percent of the items based on previous published DIF research (Classe & Kern, 2024; Patel, Robison, & Cougle, 2024; Paulsen et al., 2020). In conditions with DIF in at least five percent of items, DIF was simulated in either the a-parameter, the b-parameter, or both parameters. DIF in the a-parameter was simulated by multiplying the a-parameter of the reference group by a constant, $h = 0.2$, 0.5, or 0.8. DIF in the b-parameter was simulated by changing the b-parameter of the reference group by a constant $s = 0.5$, 1.0, or 2.0. In cases where DIF was unbalanced, all parameters were increased multiplicatively for the a-parameter or additively for the b-parameter. In conditions where DIF was balanced, half the items with DIF were increased, while half were decreased. To determine these values for h and s, a small simulation was performed to ensure that some levels of h and s would result in visible changes in the TIFs (i.e., DTF) while others would not. The results from this simulation can be seen in Supplementary Figure 1.

Example category characteristic curves from the five simulation conditions with the most extreme differences between the focal and reference groups can be found in Figure 3. These examples illustrate the true parameter differences between groups (dotted and dashed lines) as well as the biased parameters estimated using the misspecified single-group GRM (solid lines). The solid lines demonstrate how model misspecification leads to biased parameter estimates that inadequately represent either group's true response patterns and, in turn, biased θ estimates of an individual's level of the latent trait. They also highlight the practical consequences of failing to account for group differences in item functioning. Example distributions of both the generating θ and estimated θ from the same five simulation conditions can be found in Figure 4. Estimated θ values

were obtained via expected a posteriori (EAP) scoring from the misspecified single-group GRM fit to pooled data. These illustrate the bias in θ estimates resulting from the misspecified single-group GRM fit to pooled data when DIF is ignored (dashed lines). Together, Figures 3 and 4 illustrate both the item-level source and the person-level consequence of model misspecification under unmodeled DIF.

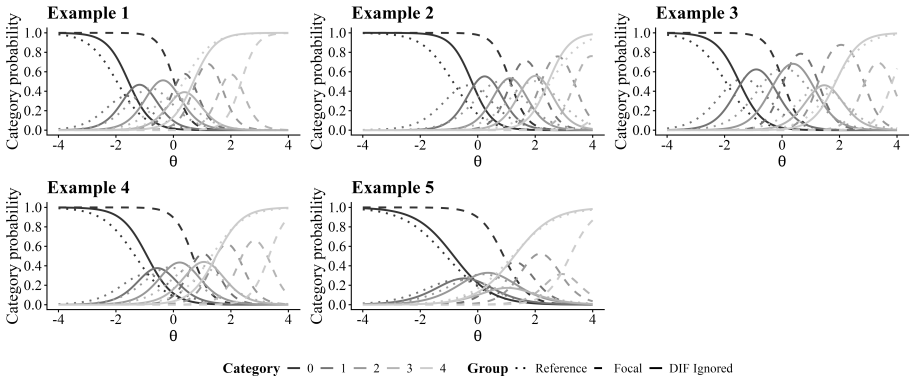

Figure 3. Example category characteristic curves for one item from the five most extreme conditions. *Note:* Each panel displays category characteristic curves (trace lines) for a single 5-point Likert-type item with differential item functioning (DIF) under one of the five most extreme simulation conditions. Curves represent the model-implied probability of endorsing each response category (0–4) as a function of the latent trait (θ). Dotted lines represent the reference group's true parameters, dashed lines represent the focal group's true parameters, and solid lines represent parameters estimated when DIF is ignored (i.e., fitting a single-group GRM to pooled data from both groups). Grayscale shading differentiates response categories. For items with DIF, discrepancies between reference and focal group curves reflect condition-specific shifts in item thresholds and/or discrimination parameters, resulting in different category response probabilities at the same level of θ. The solid lines illustrate the biased parameter estimates that result when group differences are not accounted for in model estimation. All curves are plotted over a common θ range to facilitate visual comparison of DIF magnitude and direction across conditions.

All non-DIF-related manipulated variables were fully crossed to create a total of 24 simulated conditions with no DIF. A total of 432 conditions with only a-parameter DIF were generated. An additional 432 conditions with only the b-parameter DIF were generated. When DIF was present in both the a- and b-parameters, all conditions were fully crossed to generate 1,296 conditions. Item responses were generated in R (R Core Team, 2024), with 1,000 data sets generated for each condition. All program specifications were set to the defaults. Of the 2,184 conditions, 24 did not contain DIF, 456 contained DIF with DTF, and 1,704 contained DIF without DTF. The additional conditions with impact

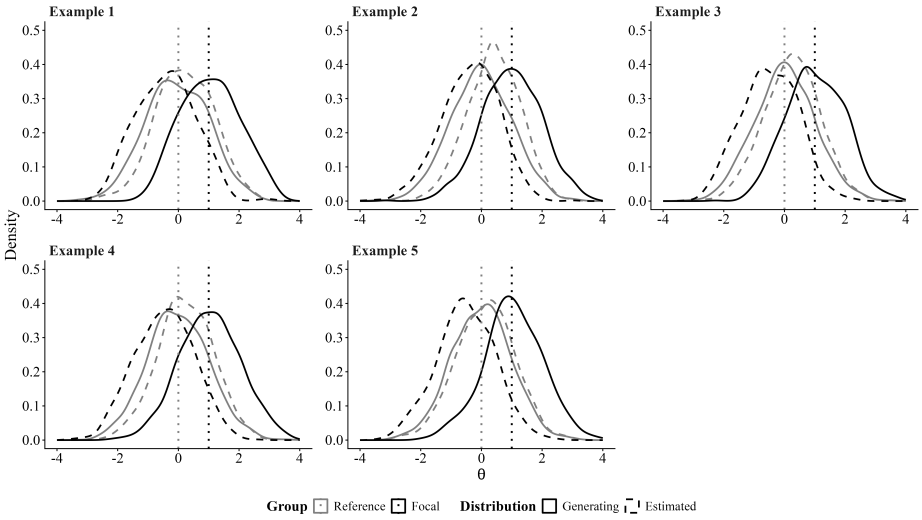

Figure 4. Example distributions of the generating and estimated θs for both the focal and reference groups from the five most extreme conditions. *Note:* Each panel displays θ distributions for one of the five most extreme simulation conditions. Vertical dotted lines represent the mean of the generating distribution. Line type distinguishes the generating (solid line) θ distribution from the estimated (dashed line) θ distribution. Color distinguishes the reference (gray) and focal (black) groups. All curves are plotted over a common θ range to facilitate visual comparison of bias magnitude and direction across conditions. The mean standard deviation of bias for Example 1 [Reference = -0.253 (0.188), Focal = 1.48 (0.131)], Example 2 [Reference = -0.399 (0.186), Focal = 1.37 (-.116)], Example 3 [Reference = -0.234 (0.149), Focal = 1.51 (0.102)], Example 4 [Reference = -0.245 (-.156), Focal = 1.50 (0.113)], and Example 5 [Reference = -0.133 (0.132), Focal = 1.560 (0.097)] are included here to summarize the level of bias in these example conditions.

were fully crossed with the 456 DIF conditions resulting in DTF. This resulted in an additional 912 conditions. As such, a total of 3,096 conditions were examined.

2.2 Classification Models

The simulated datasets were analyzed with one psychometric approach, IRT, and one nonparametric approach, RF. For the IRT model, the estimated $\theta_{Diagnosis}$ were obtained using IRT expected a posteriori (EAP) scores. Thresholds for IRT were determined through ROC curve analysis (Brown et al., 2001; Florkowski, 2008), discussed in the following section. RF does not require a threshold.

Determining Cut-Points and Probability Thresholds In ML, a default probability cut-point of .50 is often set, such that cases with a predicted probability of diagnosis of 0.5 or higher would be classified as diagnosed (Gonzalez, 2021).

This approach is referred to as using a Bayes classifier. However, this approach does not allow researchers to consider the relative importance of sensitivity and specificity within their research contexts. For example, when evaluating a screening tool for a disorder, researchers may wish to prioritize sensitivity over specificity, as they would rather send individuals on for (potentially unnecessary) further testing than miss individuals who would benefit from treatment (e.g., see Graham et al., 2019, for a discussion of this approach with an eating disorder measure) . However, other psychological studies have prioritized specificity over sensitivity (Parker-Guilbert, Leifker, Sippel, & Marshall, 2014; Wardell et al., 2020). See the following for discussions of the ethical issues of false positives in clinical psychology (Bradford, Meyer, Khan, Giardina, & Singh, 2024; Wakefield, 2010, 2015).

Determining a threshold with an ROC curve allows a researcher to prioritize sensitivity, specificity, or balance their importance. ROC curves illustrate the changes in sensitivity and specificity that occur as a function of the threshold. Sensitivity and specificity are calculated for each cut-point, and these values are then plotted, forming the ROC curve. For example, if researchers wanted a classifier that balances sensitivity and specificity, researchers could choose the threshold that maximizes the Youden index (X. Liu, 2012). The Youden index corresponds to the point on the ROC curve that is farthest 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. Once the cut-point has been set, the cut-point is then imposed on the θ estimates to get the predicted diagnosis. Given that we have no a priori reason to prioritize one over the other in this study, we have chosen the thresholds for the IRT models based on maximization of the Youden index, as was done in previous research (Gonzalez, 2021).[3]

One can also assess how well the diagnostic scale discriminates via the shape and height of the ROC curve. The area under the ROC curve (AUC) can also help compare the performance of classifiers. The AUC measures the likelihood of making a correct classification when one observation from each group (e.g., one individual who is diagnosed and one who is not) is chosen at random. 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 AUCs closer to 1.0.

[3] We initially included a second RF approach using probability-based predictions with ROC-optimized thresholds (matching the IRT approach). However, RF probability estimates exhibited substantially worse performance across all metrics compared to using RF's predicted classes with a Bayes classifier. This is consistent with known limitations of RF probability calibration, particularly with unbalanced class sizes. To streamline the presentation of results, we do not include the results for RF probability. However, we report these results in Supplementary Table 1.

Role of Training and Test Data Split-half cross-validation was implemented so that 50% of the sample was used to train the model (training set), and the remaining 50% was used for evaluation (test set). Sparseness checks were performed on the training data to ensure that each item response category was endorsed at least five times across all items to avoid zero or near-zero cases, as was done in previous IRT simulations (Gonzalez, 2021; Manapat & Edwards, 2022). If this was found not to be the case, this data set was discarded and resampled until the sparseness conditions were met. For the IRT models, the training set was used for item calibration and to determine the cut-point using ROC curves. The RF models were trained on the training set and evaluated on the test set to predict diagnostic classes. For the RF models, all tuning parameters were set to defaults, consistent with recommendations from previous research (Breiman, 2001; Gonzalez, 2021).

2.3 Assessing Classification

We evaluated the performance of the two classification approaches using a combination of metrics that assess different dimensions of classification quality: (a) accuracy which measures the proportion of correctly classified instances (both true positives and true negatives) out of the total instances, (b) sensitivity which quantifies the models' ability to correctly identify positive instances, (c) specificity which measures the models' ability to correctly identify negative instances, (d) precision which assesses the proportion of true positive predictions among all positive predictions, (e) NPV which quantifies the proportion of true negative predictions among all negative predictions, and (f) F1 score, the harmonic mean of precision and sensitivity, providing a balanced measure of the models' performance. The goal was to assess how well the models could correctly classify simulees into their generating categories.

Effect Sizes for Classification Metrics Results were first examined visually. Then, ANOVAs were conducted to assess the impact of manipulated data features (sample size of the reference group, ratio of focal to reference group, prevalence rate of the diagnosis according to the gold standard, number of items, DIF patterns, and magnitude of DIF) and the classification techniques (IRT and RF) on all classification outcomes. Main effects and interactions up to three-way interactions were included in the analysis. To ensure that the most influential effects were of focus, we probed all effect sizes (partial $\eta^2 s$) greater than or equal to .06 (medium effect; Cohen, 2009). All other effects were omitted.

3 Results

Given the 3,096 unique data conditions, we present a subset of results that focuses on key trends and significant interactions. The selection of conditions was guided by the magnitude and consistency of observed effects, with an emphasis on medium or large effect sizes ($\eta^2 \geq .06$; Cohen, 2009) and interactions that

exhibited divergent method performance. The results are organized according to the presence of both DIF and DTF. We first discuss the 24 conditions where DIF was not present in the data, followed by conditions with DIF with DTF, those with DIF without DTF, and finally, the additional conditions where impact was introduced.

3.1 No DIF Conditions

ANOVA models were used to analyze simulation results. The type of approach was influential on all classification measures (Table 1). IRT was found to produce models with higher specificity and precision, while RF produced models with higher accuracy, sensitivity, NPV, and F1 scores. It is worth noting, however, that many of these differences were quite small, occurring in the third or fourth decimal place (Table 1). Thus, there were no practical differences in performance across conditions. These results align with previous findings, supporting that IRT and RF are equally good at classifying simulees based on diagnostic assessments without DIF (Gonzalez, 2021). However, these results expand upon previous findings, illustrating the strengths of each approach on different types of classification measures (e.g., sensitivity versus specificity).

Table 1. Effect sizes and marginal means: main effects and interactions for classification outcomes in conditions with no DIF.

	Partial η^2					
	Accuracy	Sensitivity	Specificity	Precision	NPV	F1 Score
Method	1.00	0.98	0.98	0.96	0.98	0.99
Sample Size (N)	0.21	0.13	–	–	0.13	0.27
Number of Items (#)	0.09	–	–	–	–	–
Prevalence (p)	0.68	0.30	–	0.72	0.30	0.70
$N \times$ Method	0.40	0.21	–	0.07	0.21	0.43
$p \times$ Method	0.72	0.28	0.05	0.81	0.28	0.76
	Marginal Means (SD)					
	Accuracy	Sensitivity	Specificity	Precision	NPV	F1 Score
IRT	.96 (.01)	.96 (.02)	.96 (.03)	.99 (.01)	.96 (.02)	.97 (.01)
RF	.96 (.01)	.97 (.02)	.90 (.06)	.98 (.01)	.97 (.02)	.97 (.01)

Note: Partial η^2 = effect size on the given outcome; DIF = differential item functioning. Effect sizes < 0.06 (medium effect; Cohen, 2009) were omitted from this table for simplicity.

3.2 DIF Conditions without DTF

Conditions that contained DIF without DTF were analyzed, but no meaningful differences were found between the pattern of results for these DIF conditions

and no DIF conditions. However, the inclusion of DIF did result in more unstable performance across replications. As such, results of conditions with DIF but no DTF are not included here but can be found in the Supplementary Materials.

3.3 DIF Conditions with DTF

As we were interested in examining potential differences in classification performance across methods (IRT versus RF) in simulees in the focal group versus the reference group, ANOVAs were conducted as follows for each outcome. First, we fit a model using group (focal versus reference) and all manipulated data features (total sample size, ratio of focal to reference group, number of items, prevalence rate, proportion of items containing DIF, whether DIF was balanced, and strength of both threshold and slope DIF) as main effects. Variables with partial $\eta^2 \geq .06$ (medium effect; Cohen, 2009) were deemed influential. Then, to create final models, two- and three-way interactions were probed that contained at least one influential variable. Interaction effects were deemed influential if they had partial $\eta^2 \geq .06$ (medium effect; Cohen, 2009). The classification method was found to be influential for all outcomes. Effect sizes of all influential effects and the mean performance of each classification method can be found in Table 2.

Effects of Data Characteristics We first examined the influential effects of data characteristics, as presented in Figure 5. We found that higher prevalence rates, or more balanced class sizes, led to lower precision in the focal group (5A) but no change in the reference group. The specificity of focal group classifications was also found to decrease as more items contained DIF (5B), while specificity of reference group classifications slightly increased. Sensitivity of reference group classifications decreased as the proportion of focal group members increased (5C). The same pattern was observed for NPV. This decline likely reflects a shift in the classification model's estimates toward the focal group, as model training was based on the combined sample.

Precision remained stable across varying levels of threshold DIF strength when the focal group contained 50% of the total sample. However, when the focal group represented a smaller portion of the sample, precision declined as the strength of threshold DIF increased (5D). Across all sample compositions, the F1 score decreased as threshold DIF strength increased, with the steepest decline occurring when the focal group comprised half of the overall sample (5E). Lastly, the overall accuracy of all models decreased as more items contained DIF (5F).

Effects of Classification Method. Next, we shift our focus to the effects resulting from method type (IRT vs. RF) on each classification metric.

Accuracy. The proportion of the sample that consisted of the focal group, or the proportion of simulees in the sample with DIF, did not influence the accuracy of RF models but did influence the accuracy of IRT models such that a larger focal group led to lower accuracy (Figure 6A). Increases in threshold DIF strength,

Table 2. Effect sizes and marginal means: main effects and interactions of all classification metrics in conditions that contained DIF, resulting in DTF.

	Partial η^2					
	Accuracy	Sensitivity	Specificity	Precision	NPV	F1 Score
Method	0.16	0.24	–	–	0.24	0.12
Sample Size (N)	–	–	–	–	–	–
Prevalence (p)	–	–	–	–	–	–
FR Ratio (FR)	0.06	0.26	0.17	0.11	0.26	0.07
DIF Group (g)	–	0.35	0.45	0.40	0.35	**0.06**
DIF Rate (d)	**0.09**	–	–	–	–	–
b Strength (b)	0.18	0.11	0.26	0.21	0.11	0.14
Method × FR	0.06	0.20	0.06	–	0.20	0.07
Method × g	–	0.16	0.09	0.07	0.16	–
Method × b	0.10	0.10	0.09	0.07	0.10	0.08
b × FR	–	0.17	0.07	0.06‡	0.17	0.07‡
b × g	–	0.14	0.24	0.18	0.14	–
g × d	–	–	0.06‡	–	–	–
g × FR	–	0.17	0.15	0.11	0.17	–
g × p	–	–	–	0.13‡	–	–
b × Method × FR	–	0.16	0.06	–	0.16	–
b × Method × g	–	0.08	0.08	0.06	0.08	–
FR × g × b	–	0.11	–	–	0.11	–
FR × g × Method	–	–	0.06	–	–	–
	Marginal Means (SD)					
	Accuracy	Sensitivity	Specificity	Precision	NPV	F1 Score
IRT						
Full Sample	.89 (.08)	.89 (.11)	.90 (.11)	.97 (.04)	.89 (.11)	.92 (.07)
Reference Group	.86 (.17)	.82 (.22)	.98 (.03)	.99 (.01)	.82 (.22)	.88 (.18)
Focal Group	.92 (.08)	.99 (.02)	.68 (.36)	.92 (.09)	.99 (.02)	.95 (.05)
RF						
Full Sample	.95 (.02)	.96 (.02)	.88 (.07)	.97 (.02)	.96 (.02)	.97 (.02)
Reference Group	.95 (.02)	.95 (.03)	.92 (.06)	.98 (.01)	.95 (.03)	.96 (.02)
Focal Group	.95 (.03)	.99 (.02)	.78 (.17)	.95 (.04)	.99 (.02)	.97 (.02)

Note. Partial η^2 = effect size on the given outcome; DIF = differential item functioning; DTF = differential test functioning. Effect sizes < 0.06 (medium effect; Cohen, 2009) were omitted. Bolded values indicate influential main effects that are not part of an influential interaction effect. A model with only main effects was first examined for each outcome. We then probed all possible two- and three-way interactions that contained at least one variable with an influential main effect. In such cases when a main effect or interaction was not found to be influential on any classification measure, it was omitted from this table. Effects were determined to be influential if they were greater than or equal to 0.06 (medium effect; Cohen, 2009). ‡ Indicates influential two-way interactions that are not part of an influential three-way interaction.

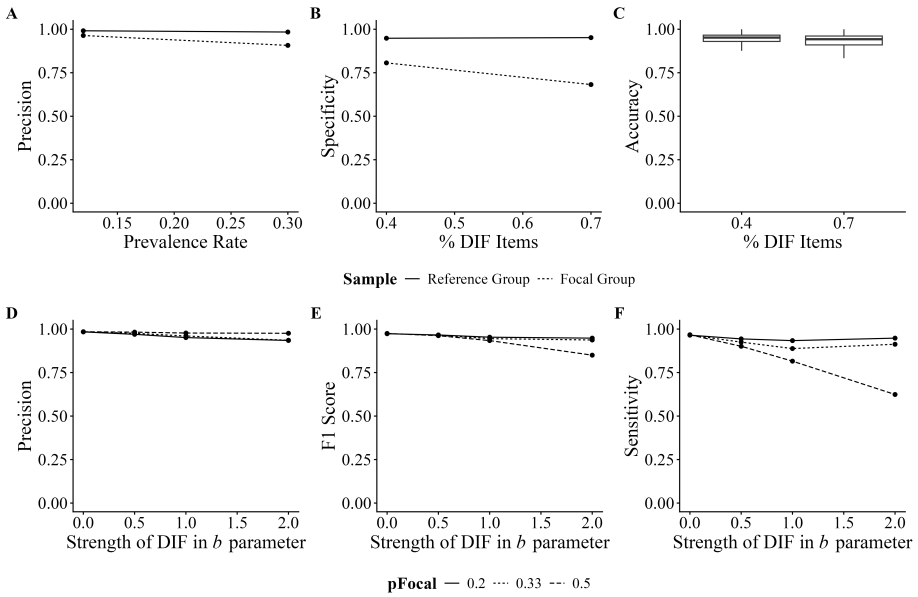

Figure 5. Influential effects in differential test functioning conditions that do not contain the method as a predictor. *Note:* pFocal = proportion of the total sample that consisted of the focal group; NPV = negative predictive value; % DIF Items = proportion of total items containing DIF. Panel A displays the interaction between prevalence rate and sample on precision. Panel B displays the interaction between % DIF items and sample on specificity. Panel C displays the main effect of % DIF items on accuracy. Panel D, E, and F display the interaction between pFocal and the strength of *b*-parameter DIF on precision, F1 score, and sensitivity, respectively.

or how much more severe symptoms of simulees in the focal group must be compared to the reference group, led to lower accuracy in IRT (6B), whereas RF models were relatively robust to changes in threshold DIF.

Sensitivity. Across all threshold DIF strengths and all examined focal group percentages, RF models had higher sensitivity than IRT models (Figure 7A). The sensitivity of IRT models decreased as the strength of threshold DIF increased, with the strongest effect occurring when at least 50% of the sample was in the focal group. Increases in threshold parameter DIF did not affect the sensitivity of RF models but led to decreases in sensitivity for the reference group of IRT models (7B). No visible differences were observed in the sensitivity of focal group classifications across RF and IRT models. However, since reference group sensitivity was worse in IRT models than RF models, the overall sensitivity of RF models was greater than that of IRT models.

Specificity. Influential method effects on specificity are presented in Figure 8. The specificity of IRT models was influenced by both the strength of threshold

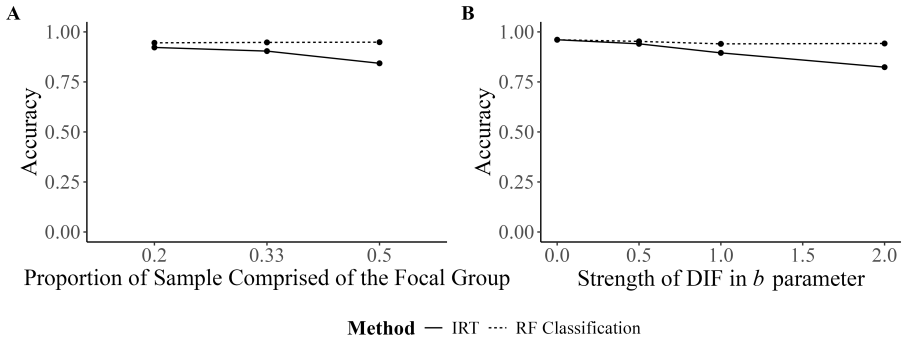

Figure 6. Influential interactions for accuracy. *Note:* Panel A shows accuracy as a function of the proportion of the sample that consisted of the focal group, with values ranging from 0.2 to 0.5. Panel B shows accuracy as a function of the strength of DIF in the *b*-parameter. In both panels, RF classification (dotted line) maintains relatively stable accuracy, while IRT (solid line) shows declining accuracy as the proportion of the focal group increases (Panel A) or as DIF strength increases (Panel B). Both methods demonstrate high accuracy levels (above 0.75) across all conditions tested, with RF classification consistently outperforming IRT, particularly in more challenging measurement conditions.

DIF and the size of the focal group (8A). When the focal group constituted 50% of the overall sample, IRT models were robust to varying threshold DIF strength and exhibited superior specificity compared to RF models. However, when the focal group comprised 20 or 30 percent of the sample, IRT model specificity declined as threshold DIF strength increased. In cases where threshold parameter DIF exceeded one and a half, RF models outperformed IRT models in terms of specificity.

This effect is further illustrated in Panel 8B, where we see that the specificity of the reference group remained relatively stable across varying threshold DIF strengths for both IRT and RF models, with minor increases observed as the strength increases. Conversely, focal group specificity was substantially affected by threshold DIF strength, and this relationship diverged between the two modeling approaches. For IRT models, specificity consistently declined with increases in threshold parameter DIF. RF models also exhibited decreasing specificity up to a threshold DIF strength of one, after which the specificity stabilized. Consequently, under high *b*-parameter only DIF conditions, RF models yielded higher specificity for focal group members and the overall sample than IRT models. In contrast, when threshold DIF strength was low, IRT models maintained higher specificity.

Panel 8C illustrates the varying effect the subsample and focal group proportion had on specificity across model types. Specifically, reference group specificity increased as the focal group proportion increased in both IRT and RF models. This likely resulted from consequential sample size increases, with IRT models always producing higher reference group specificity than focal group specificity.

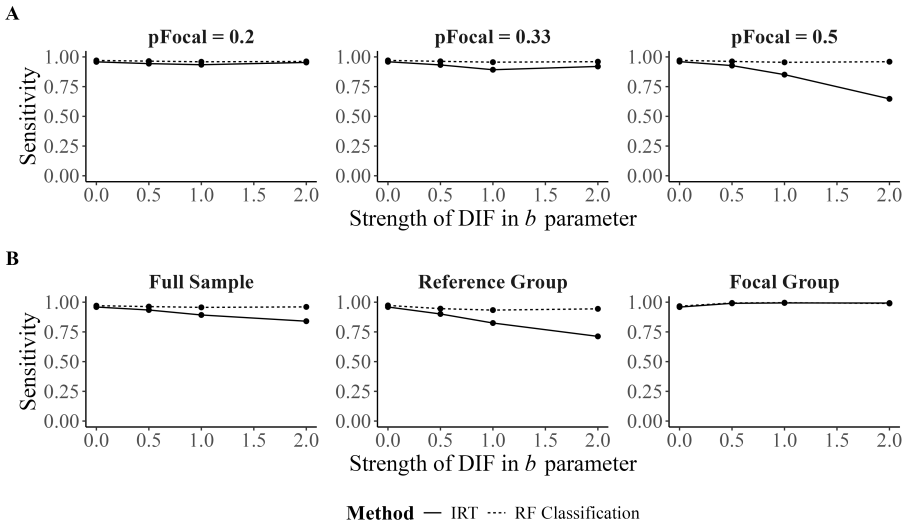

Figure 7. Influential interactions for sensitivity. *Note:* IRT = item response theory; RF Classification = random forest classification; pFocal = proportion of sample that consisted of the focal group. Panel A displays three scenarios with varying focal group sizes showing how accuracy changes with increasing strength of DIF in the *b*-parameter. Panel B presents the interaction between sub-sample (full, reference, or focal) and strength of DIF in the *b*-parameter. In all panels, RF demonstrates consistently higher and more stable sensitivity compared to IRT, particularly as DIF strength increases. The performance gap between methods widens under more challenging conditions, specifically with larger focal group proportions (pFocal = 0.5) and the reference group. Both methods maintain accuracy above 0.75 across most conditions, though IRT performance notably decreases with higher DIF strength values (as expected), while RF classification remains relatively robust.

When the sample contained equal numbers of reference and focal group simulees, IRT models resulted in higher focal group specificity than RF models, but if the focal group was less than half the total sample, RF models had higher focal group specificity than IRT models. If examining the overall sample only, RF and IRT models yielded equal specificity values, on average, until the focal group comprised at least 40% of the overall sample, when IRT began to yield better sensitivity than RF models.

Precision. The reference group precision of both IRT and RF models was robust to changes in the strength of threshold DIF, with IRT slightly outperforming RF models. Both RF and IRT models had worse focal group precision as the strength of threshold DIF increased, with strength of threshold DIF having a stronger effect on IRT models than RF models, such that as threshold parameter DIF strength reached or exceeded one, RF outperformed IRT in the focal group. This result held in the full sample results as well.

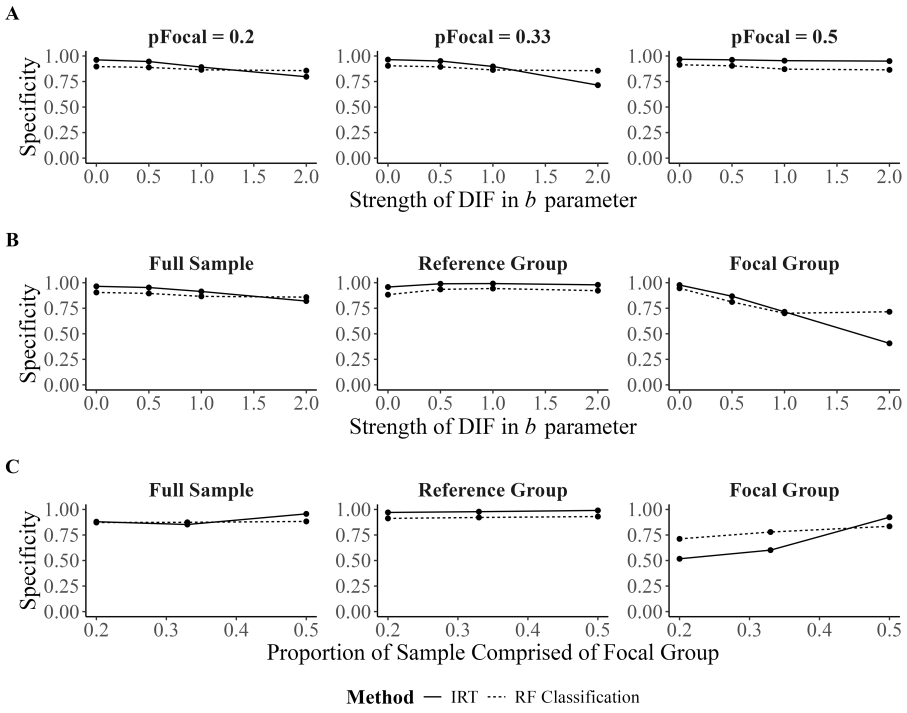

Figure 8. Influential interactions for specificity. *Note:* IRT = item response theory; RF Classification = random forest classification; pFocal = proportion of the sample that consisted of the focal group. Panel A shows specificity as a function of DIF strength in the *b*-parameter for three different focal group proportions. Panel B displays specificity across DIF strength for the full sample, reference group, and focal group separately. Panel C presents specificity as a function of the proportion of the the sample that consisted of the focal group for three different sample types: full sample, reference group, and focal group. RF classification generally maintains lower, though more stable, specificity across conditions, while IRT performance tends to decline with increasing DIF strength (as expected) and higher focal group proportions.

NPV. Regardless of sample type, RF models had higher NPV values than IRT models, and the NPV of RF models was robust to changes in focal group proportion (Figure 9A). However, for IRT models, the NPV of the reference group decreased as more of the sample contained simulees experiencing DIF, which in turn decreased the NPV of the full sample in IRT models. We saw a similar influence of threshold DIF strength (9B), where RF models were robust to changes in threshold DIF strength, regardless of the sub-sample, while IRT models had lower NPV values as threshold parameter DIF strength increased.

F1 Score. RF models were found to produce higher F1 scores than IRT models, regardless of either threshold DIF strength or the proportion of the the sample that

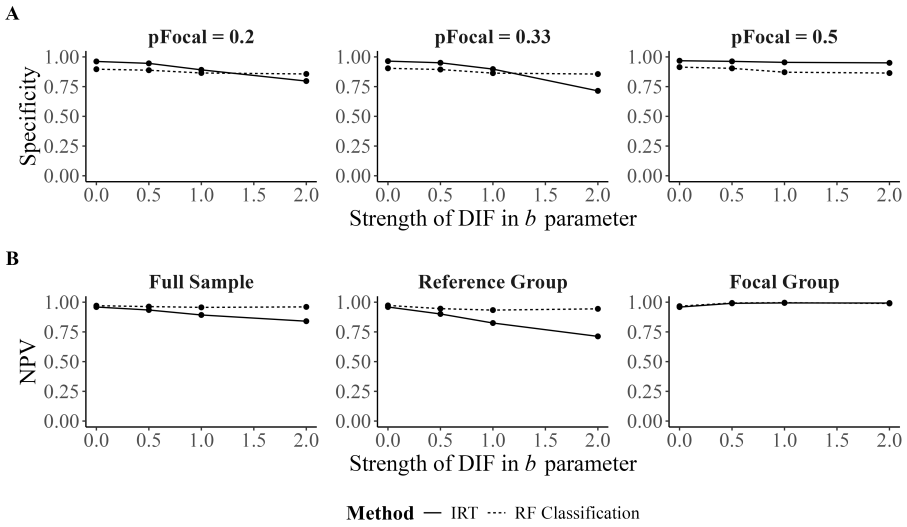

Figure 9. Influential interactions for NPV. *Note:* IRT = item response theory; RF Classification = random forest classification; NPV = negative predictive value. Panel A illustrates the effect of focal group proportion across sample and method type. Panel B displays the effect of the strength of threshold DIF across sample and method type. In general, RF classification yields models with higher and more stable NPV values than IRT models.

consisted of the focal group. In addition, threshold DIF strength and proportion of the the sample that consisted of the focal group did not affect the F1 score of the RF models, whereas increases in either led to worse F1 scores in IRT models.

3.4 Impact Conditions

As we were interested in examining potential differences in classification performance across methods (IRT-based versus RF-based) in simulees in the focal group versus the reference group, ANOVAs were conducted as follows for each outcome. First, we fit a model using group (full, focal, and reference) and all manipulated data features (total sample size, ratio of focal to reference group, number of items, prevalence rate, proportion of items containing DIF, whether DIF was balanced, strength of both threshold and slope DIF, and impact) as main effects. Variables with partial $\eta^2 \geq .06$ (medium effect; Cohen, 2009) were deemed influential. Then, to create final models, two- and three-way interactions were probed that contained at least one influential variable. Interaction effects were deemed influential if they had partial $\eta^2 \geq .06$ (medium effect; Cohen, 2009). Effect sizes of all influential effects and the mean performance of each classification method can be found in Table 3. The classification method was

Table 3. Effect sizes and marginal means: main effects and interactions of all classification metrics in conditions that contained impact.

	Partial η^2					
	Accuracy	Sensitivity	Specificity	Precision	NPV	F1 Score
Method	0.08	**0.08**	–	–	**0.08**	0.06
Sample Size (N)	–	–	–	–	–	–
Prevalence (p)	–	–	–	**0.12**	–	**0.07**
FR Ratio (FR)	–	–	–	–	–	–
DIF Group (g)	–	**0.11**	0.22	**0.13**	**0.11**	–
DIF Rate (d)	**0.07**	–	0.07‡	–	–	–
b Strength (b)	0.13	–	0.10	–	–	0.10
Impact (Imp)	–	–	0.08	–	–	–
Method × b	0.08	–	–	–	–	0.06
$g \times d$	–	–	0.11	–	–	–
$g \times b$	–	–	0.13‡	–	–	–
$g \times$ Imp	–	–	0.15‡	–	–	–
	Marginal Means (SD)					
	Accuracy	Sensitivity	Specificity	Precision	NPV	F1 Score
IRT						
Full Sample	.93 (.06)	.95 (.05)	.93 (.08)	.98 (.03)	.93 (.08)	.94 (.08)
Reference Group	.91 (.11)	.94 (.10)	.90 (.13)	.99 (.01)	.90 (.13)	.97 (.03)
Focal Group	.94 (.06)	.96 (.05)	.97 (.03)	.96 (.08)	.97 (.03)	.81 (.29)
RF						
Full Sample	.95 (.02)	.97 (.01)	.97 (.02)	.97 (.01)	.97 (.02)	.89 (.06)
Reference Group	.95 (.02)	.97 (.02)	.96 (.03)	.98 (.01)	.96 (.03)	.91 (.07)
Focal Group	.95 (.03)	.97 (.02)	.98 (.02)	.96 (.04)	.98 (.02)	.82 (.18)

Note. Partial η^2 = effect size on the given outcome; DIF = differential item functioning; DTF = differential test functioning. Effect sizes < 0.06 (medium effect; Cohen, 2009) were omitted. Bolded values indicate influential main effects that are not part of an influential interaction effect. A model with only main effects was first examined for each outcome. We then probed all possible two- and three-way interactions that contained at least one variable with an influential main effect. In such cases when a main effect or interaction was not found to be influential on any classification measure, it was omitted from this table. Effects were determined to be influential if they were greater than or equal to 0.06 (medium effect; Cohen, 2009) ‡ Indicates influential effects that were not observed in the non-impact conditions (seen in Table 2).

found to be influential for all outcomes except for specificity and precision. Note that this is the same as results in non-impact conditions (Table 2).

As many effects were the same between Table 2 and Table 3, we focus our discussion here on the three unique effects in Table 3. We begin with the two effects that did not contain impact. First, it was observed that as more items contained DIF, the specificity of the model decreased (Figure 10A). Second, as the strength of DIF in the b-parameter increased, the specificity of both the full

sample and focal group decreased, with this effect being stronger in the focal group (10B). However, the specificity of the reference group improved as the strength of DIF in the b-parameter changed from zero to one and then began to decrease slightly once DIF strength surpassed one. This disordinal interaction is interesting and could be the result of the relationship seen between method and b-parameter DIF in Figure 7A. Lastly, when impact was either zero (focal and reference group ability come from the same underlying distribution) or positive (focal group $\mu = 1$, reference group $\mu = 0$), the specificity of the models appears to be relatively unaffected (10C). However, when the impact is negative (focal group $\mu = -1$, reference group $\mu = 0$), the specificity of the focal group across methods is notably lower (mean = 0.51) than that of the full sample (mean = 0.90) and reference group (mean = 0.95). Note that no influential interaction effect on specificity was found between sample type, impact, and method (partial $\eta^2 = 0.01$), indicating that both RF and IRT have poor focal group specificity when impact is negative (i.e., the mean ability of the focal group is notably lower than that of the reference group).

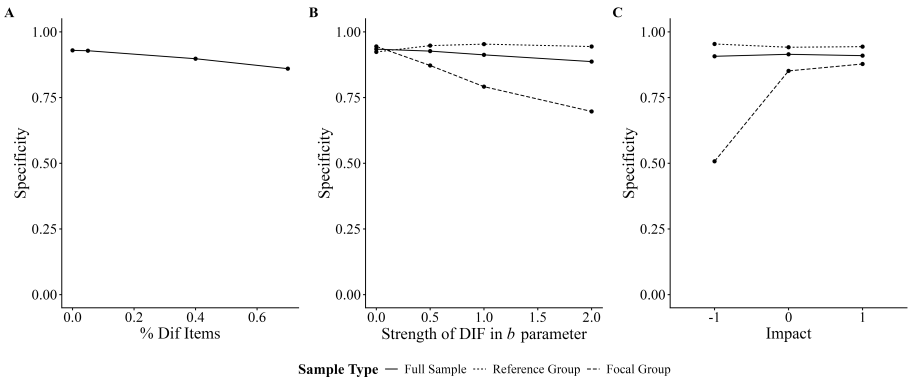

Figure 10. Influential interactions for conditions containing impact. *Note:* Panel A illustrates the effect of the proportion of items containing DIF on specificity. Panel B illustrates the effect of the strength of DIF in the b-parameter on specificity across sample types. Panel C illustrates the effect of strength and direction of impact across sample types (full, reference, focal) on specificity.

4 Empirical Example

To illustrate the practical application of IRT and RF classification methods in the presence of DIF, we applied both approaches to data collected via the University of Oklahoma's Psychology Research Pool. The purpose of the study was to predict misophonia diagnosis based on item responses to the DMQ (Rosenthal et al.,

2021). As misophonia is often co-diagnosed with autism, we hypothesized that DIF may occur across gender in misophonia, as is the case in autism (Schiltz & Magnus, 2020). Informed consent was obtained electronically, and all procedures were approved by the Institutional Review Board. After removing participants who failed attention checks, were under 18 years old, or reported a gender other than cisgender man or cisgender woman, the final sample included 725 participants (80.3% female, 75.6% White, M_{age} = 18.8, SD_{age} = 1.28). Male participants were less likely to pass attention checks than female participants (β = -0.54, SE = 0.22, z = -2.49, $p < .05$). No missing data were present in the sample. For more on this sample, see Bain et al. (2025).

4.1 Measures

DMQ The 12-item impairment subscale of the DMQ (Rosenthal et al., 2021), which assesses the extent to which auditory triggers interfere with daily functioning, was used in this study. Items use a 5-point Likert-type scale ranging from 0 (Not at all) to 4 (Extremely). Gender-based DIF was identified in four of these 12 items (see Supplementary Figure 2 for trace lines).

Misophonia Questionnaire (MQ) The MQ severity question (Wu et al., 2014) asks participants to rate their sound sensitivity from 1 (minimal) to 15 (very severe). As is traditionally done, this item was used as a self-report gold standard to split participants into diagnostic groups based on a threshold cutoff value of seven or above.

Classification Models Both IRT and RF models were evaluated using the DMQ impairment items as predictors and the MQ severity-based diagnosis as the outcome. For the IRT model, θ estimates were obtained using the GRM, and diagnostic classifications were made based on ROC-optimized thresholds using the Youden Index. For the RF model, classification was performed using a Bayes classifier. All models were evaluated in a 50% hold-out (i.e., test) dataset using the classification metrics described previously.

4.2 Results

Classification Outcomes Classification performance for IRT and RF models is shown in Figure 11 across the total sample and by gender. Overall, RF outperformed IRT on key metrics including accuracy, F1 score, sensitivity, and NPV, suggesting RF more effectively identified both diagnosed and non-diagnosed simulees. These advantages held across the total sample as well as the female and male subgroups. In contrast, IRT yielded higher specificity and precision, indicating a more conservative approach with fewer false positives but also reduced sensitivity, particularly among females. This trade-off mirrors patterns observed in the simulation study, where IRT underperformed in the presence of DIF. These results highlight the strength of RF in diagnostic contexts where sensitivity is prioritized or DIF may be unaccounted for.

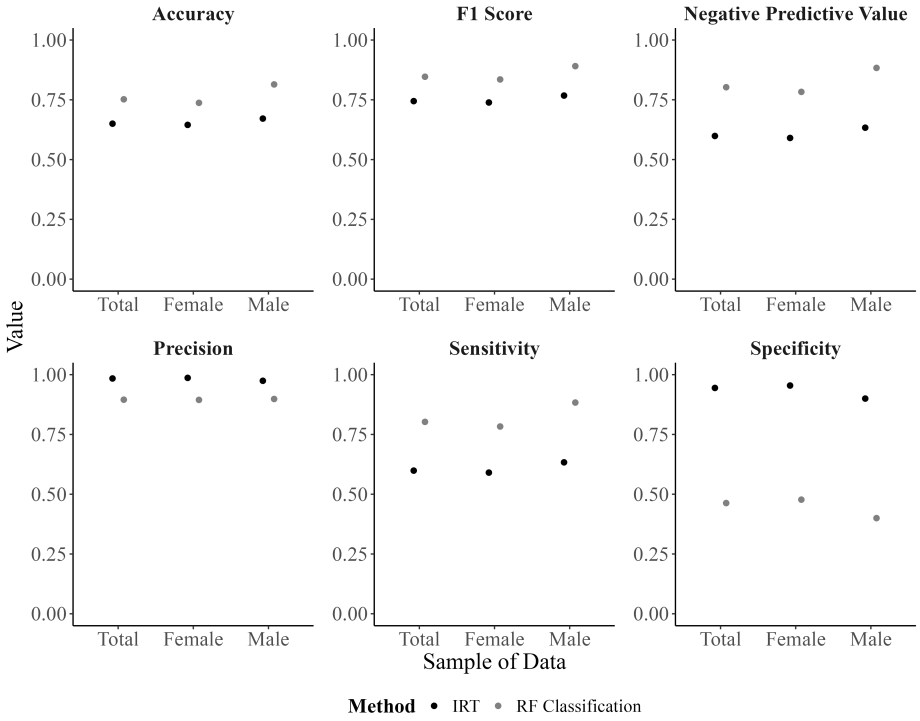

Figure 11. Mean classification performance metrics across models using the impairment items from the DMQ, separated by sample group (Total, Female, Male). *Note:* IRT = item response theory; RF Classification = random forest classification. Each panel represents a distinct evaluation metric (accuracy, F1 score, negative predictive value, precision, sensitivity, and specificity) with outcome values plotted on the y-axis. Points indicate model performance in the 50% holdout sample. Higher values indicate better classification performance.

Variable Importance Although RF does not provide the same, clearly interpretable relationship between the classifications and item responses as IRT, RF does output variable importance scores. One such variable importance score is the mean decrease in accuracy, which reflects how much the model's overall prediction accuracy decreases when that variable is systematically altered (i.e., randomly permuted) to remove its association with the outcome variable. We examined the relative importance of DIF items versus non-DIF items in the RF model and found that the two most important variables (which together contribute 47.38% of the total variable importance) in the RF model were DIF items. All item importance scores can be found in Supplemental Table 4.

5 Discussion

The current study sought to evaluate how DIF impacts classification accuracy using parametric (IRT) and nonparametric (RF) models. Accurate classification in psychology is paramount, yet classification accuracy may decrease when DIF is present but unmodeled, leading to erroneous group comparisons, compromised construct validity, and issues with replication. Given the increasing use of machine learning in psychological science, this study investigated whether RF maintains more robust classification performance than single-group IRT when DIF is present but not explicitly modeled. The empirical results from the misophonia dataset mirrored these simulation-based trends, indicating that, the advantages of RF observed in the simulations generalize to real-world diagnostic contexts. Notably, RF achieved greater classification stability across gender groups, whereas IRT demonstrated lower sensitivity among females, suggesting that unmodeled DIF can meaningfully reduce classification accuracy in practice.

5.1 Parametric Assumptions and Classification Under Non-DIF Conditions

When DIF was absent, both IRT and RF showed comparable average performance across classification metrics. These findings are consistent with prior research showing that parametric and nonparametric classifiers perform similarly when assumptions are met (Gonzalez, 2021). However, IRT models make a number of critical assumptions (e.g., monotonicity, local independence, item invariance, and unidimensionality). Violations of these assumptions, particularly in applied data where multidimensionality or local dependence is likely, can reduce the classification performance of IRT-based approaches. RF, which does not rely on these assumptions, showed marginally higher accuracy, sensitivity, NPV, and F1 scores, especially as sample size increased. These differences, while marginal, point to a broader tradeoff. IRT provides a theory-driven framework for modeling latent traits and item functioning, but its parametric assumptions must be reasonably met for accurate inference. RF offers a flexible, data-driven alternative with advantages when data characteristics substantially violate IRT assumptions. However, RF does come with the cost of reduced interpretability, which is potentially problematic when explanation and transparency are priorities. It is important to reiterate that our study compares RF to single-group IRT. More complex IRT-based approaches, including multiple-group IRT (when the cause of DIF is known) and mixture IRT (when the cause of DIF is unknown and impact is present), can explicitly model DIF but cannot always reasonably be implemented. Our findings therefore reflect the performance trade-offs between RF and single-group IRT in scenarios where DIF exists but is ignored or cannot be modeled.

5.2 The Influence of DIF: Classification Stability and Robustness

When DIF was present but not strong enough to produce DTF, both IRT and RF maintained performance close to what was seen in conditions with no

DIF. However, greater variability was observed across simulation replications, particularly for IRT. This suggests that unmodeled DIF, even when not resulting in DTF, can reduce the stability of classification decisions.

Under more severe DIF conditions involving DTF, IRT's classification performance decreased (as expected). Decreases in performance were systematic and not limited to the focal group, reflecting broader model misspecification. RF, in contrast, maintained relatively stable performance despite increases in DIF item count, magnitude, or focal group size. This robustness suggests that RF's flexible modeling of item interactions allows it to maintain classification performance when DIF is present but not modeled. These simulation findings were further supported in the empirical example, where gender-based DIF in the DMQ items reduced IRT sensitivity but had a negligible impact on RF classification. Despite clear item-level DIF, RF maintained high diagnostic accuracy and F1 scores across subgroups, consistent with the simulated evidence that RF is robust to latent sources of DIF and DTF. This consistency across simulated and empirical data further demonstrates RF's practical robustness to DIF effects.

RF's classification robustness under DIF conditions may stem from its ability to adaptively partition the response space. The fairness literature in ML suggests that ensemble methods like RF can learn proxy variables for latent group membership, thereby partially mitigating the impact of unmeasured DIF (Classe & Kern, 2024; Kraus, Wild, & Hilbert, 2024). RF can maintain classification performance by partitioning on combinations of item responses that reflect underlying group differences without requiring explicit identifications of those groups. The present data offer some evidence for this. In the empirical example, the two items with the highest variable importance scores were both DIF items, which is consistent with the idea that RF prioritized items containing group-specific signal to maintain classification performance. However, this finding comes from a single dataset and does not constitute a formal assessment of what aspects of the RF algorithm allow it to maintain classification performance. Whether RF is explicitly partitioning on item combinations that proxy DIF group membership remains an open question.

It is worth noting that this finding may raise concerns when the training data contains biased or noisy labels (e.g., Garb, 2021). If the gold standard diagnosis reflects systematic misclassification (e.g., underdiagnosis in marginalized populations), RF may learn and reproduce these patterns in its classifications. This introduces the risk of reifying diagnostic bias, particularly if the model is used without adequate scrutiny or interpretation tools. As such, we recommend that researchers carefully evaluate the quality and potential bias of gold standard classifications and utilize post hoc interpretation methods in contexts where model decisions have direct implications for diagnosis, treatment, or resource allocation.

5.3 The Role of Impact on Classification Performance

The introduction of impact (i.e., group-level mean differences in the latent trait) had influences on classification performance that extend beyond those observed

under non-impact conditions alone. Notably, the presence of negative impact (i.e., when the focal group's mean trait level is lower than that of the reference group) led to a pronounced decrease of focal group specificity, regardless of whether IRT or RF was used. This finding echoes concerns in the fairness literature, where group-based disparities in base rates or latent trait distributions can produce disproportionate false positive or false negative rates, even in the absence of overt bias in the model itself (Garb, 2021). From a practical standpoint, this indicates that more focal group members were incorrectly classified as having the condition, which could result in overdiagnosis in real-world settings. Importantly, this pattern did not interact significantly with the classification method, suggesting that both IRT and RF are affected by negative impact. This finding represents a limitation to RF's robustness: while it maintains classification performance in the presence of unmodeled DIF, it cannot fully compensate for true group-level differences in trait distributions. In other words, RF can maintain classification performance despite DIF, but disparities that result from true trait differences rather than item-level bias will still be present in RF classifications. This underscores the need to distinguish between bias (e.g., DIF) and true differences (i.e., impact) when evaluating classification fairness and utility.

5.4 Interpretability, Fairness, and Clinical Implications

While RF demonstrated stronger classification performance under DIF and DTF conditions, its limited interpretability should prompt careful consideration when used in clinical settings. In clinical contexts, where classification outcomes may inform diagnosis or treatment, stakeholders often need to understand how a decision was reached, not just whether it was accurate. RF's flexibility enables it to model complex response patterns and adapt to unobserved sources of bias, but the internal logic of ensemble methods is relatively unclear. As a result, it may be difficult to determine whether maintained classification performance reflects appropriate adaptation to complex response patterns or a reliance on spurious associations in the data.

As RF is an ensemble method, the decision rules for creating model predictions are unknown, and RFs are not directly interpretable. Metrics like variable importance measures can be used to identify which variables (i.e., items) the model is most reliant on when creating predictions. In the empirical study, the two most influential items for diagnostic prediction were DIF items. This may indicate that these items, despite exhibiting DIF, capture aspects of misophonia most strongly predictive of diagnosis. However, this was only one circumstance for identifying how RF uses DIF versus non-DIF items in diagnostic classification. Future simulation research is needed to investigate the use of variable importance measures to determine the relative influence of DIF versus non-DIF items on diagnostic classification in RF. In applied settings, researchers should be sure to investigate whether the most important variables align with theoretical expectations or clinical relevance.

The inclusion of impact in our simulation design also reinforces the need for interpretability when group-level differences exist. Misclassification due to

true group differences could have real-world consequences, particularly in clinical assessment, where over- or under-diagnosis may lead to inappropriate treatment or missed intervention opportunities. Post hoc interpretability tools and fairness evaluations are critical to understanding whether classification errors are distributed equitably across groups. In the context of parametric modeling, approaches like mixture IRT may offer a promising solution by explicitly modeling both DIF and impact within a unified framework, allowing researchers to understand whether classification errors stem from measurement bias versus true trait differences (Sajobi et al., 2022). Simulation-based calibration or reweighting strategies could also be explored in future work to mitigate the effects of impact when ground-truth differences are known or suspected (Naderalvojoud, Curtin, Asch, Humphreys, & Hernandez-Boussard, 2025).

When grouping variables are known and measured, the appropriate approach is to test for DIF and use approaches like multi-group IRT or mixture IRT for scoring and classification. Our findings are not intended to replace these approaches. Rather, they address scenarios where traditional DIF modeling cannot be implemented—when relevant grouping variables are unknown, unmeasured, or too complex to model. In such cases, our results suggest RF may maintain more robust classification performance than single-group IRT and therefore may be a reasonable alternative approach to classification.

While RF may offer practical advantages in research contexts where maintaining classification performance under model misspecification is prioritized over interpretability (e.g., early-stage screening, exploratory identification of risk or symptom groups, large-scale accuracy-focused classification), its use in high-stakes clinical decision-making warrants caution. For example, when decisions must be justified, when treatment planning requires transparency, or when fairness and accountability are critical. Conclusions about clinical utility should acknowledge that simulation studies, while useful for testing specific conditions, cannot fully reflect the complexity of real-world diagnosis, including comorbidity, diagnostic ambiguity, and contextual influences.

5.5 Limitations and Future Directions

As with any study, there are limitations to consider. First, our simulations focused exclusively on unidimensional data structures. Although this serves as a necessary first step in isolating the effects of DIF on classification accuracy, many constructs in diagnostic assessment are multidimensional. Future work should examine the impact of DIF in multidimensional IRT and ML contexts to assess generalizability. Second, only a single ML approach (e.g., RF) was implemented in this study. While RF is widely used for its accuracy (especially in subsequent or new samples), examining only one nonparametric method restricts the scope of this simulation and may overlook important comparisons with other algorithms. We hypothesize that RF's ability to model interactions between items may be critical. DIF creates complex patterns where the relationship between items and diagnosis depends on response patterns across multiple items. RF's recursive partitioning may capture these patterns (i.e., interactions between items)

without researchers having to explicitly specify interaction terms, as would need to be done in models like regularized logistic regression. Future research should systematically compare RF to other machine learning methods (e.g., regularized logistic regression, elastic net) to determine which specific algorithmic features maintain classification robustness under unmodeled DIF. For example, future research could examine the performance of decision trees that leverage the same strengths regarding item interactions but are more interpretable than RF models.

Third, the simulation used a two-group DIF structure, which does not fully capture the applied scenarios that motivated this study, where the source of DIF may be complex. The two-group design was chosen because it represents the simplest case, and the observed performance of RF relative to single-group IRT-based classifications may be a lower bound: as DIF structure becomes more complex, the bias in θ estimates from the misspecified single-group IRT approach would almost certainly be worse. Future work should examine how these classification approaches perform when the complex scenario is directly simulated, that is to say, DIF should occur across multi-group, intersectional, or other complex groups, to determine whether the present findings generalize to these increasingly complex scenarios.

Fourth, although RF models can be partially interpreted through tools like variable importance scores or partial dependence plots, these do not guarantee actionable insight. For example, an item exhibiting DIF may appear less important in the RF model, but whether this reflects the model appropriately minimizing its influence or a confounding interaction remains unclear. Future research could evaluate variable importance within simulation frameworks to better understand how RF maintains classification accuracy when DIF is present. Lastly, future work should explore interpretable ML strategies, such as shallow decision trees, to help understand which response patterns RF uses to maintain classification under DIF conditions. For instance, one could train a decision tree with a maximum depth of two to capture simple interaction patterns in the data. The resulting terminal node memberships could then be used as indicators of latent grouping structures or proxy variables influencing DIF. These derived groups could subsequently be incorporated into a traditional IRT framework to model DIF more explicitly, offering a novel way to combine RF's demonstrated classification performance with IRT's interpretability.

Our comparison focused on single-group IRT and on polytomous data analyzed with the GRM. While the GRM is the most widely used IRT model for Likert-type scales in psychology, other IRT models (e.g., 1PL, 2PL, 3PL) are commonly used in educational assessment and for some psychological measures with binary response formats. Given that the GRM reduces to the 2PL when the number of response options is two and the 1PL is often considered a more constrained version of the 2PL, it is likely that the GRM represents a more complex and potentially challenging scenario for model fitting. Examining the GRM specifically serves as a reasonable starting point for this line of inquiry and provides a proof-of-concept for the approach. However, additional work is needed to generalize the results of the current study to dichotomous models, as the nature of DIF and the

performance characteristics of both IRT and RF may differ when working with alternative IRT models. It would also be useful for future simulation studies to examine whether our pattern of results holds for other response formats that play a significant role in psychological assessment (e.g., forced-choice questionnaires such as the MMPI).

When DIF is suspected but groups are unknown, mixture IRT models and latent class analysis offer psychometric alternatives to model latent group structure. These approaches would be particularly relevant for comparison in impact conditions where groups differ in trait distributions, as class separation may be more identifiable under such conditions. Future research should compare RF to these mixture modeling approaches and examine whether our findings extend to other IRT models to provide a more complete picture of available methods for handling unmodeled DIF. This would help clarify whether RF's robustness reflects a limitation of single-group IRT specifically or represents benefits relative to all available psychometric approaches. It is also worth noting that comparisons with mixture IRT would be particularly informative in the complex multi-group DIF scenarios not examined here. The two-group findings reported in the present study should be treated as initial evidence for rather than a complete demonstration of RF's relative performance under unmodeled DIF.

Additionally, more research is needed to evaluate the generalizability of these findings for more data contexts that are representative of a larger number of psychological scales. For example, many real-world screeners are brief (e.g., 10–20 items). Although our empirical example uses a short scale (i.e., 12 items), future simulation work should investigate the performance of these methods when using shorter instruments. Future work should consider incorporating domain-specific interpretation methods to ensure that ML models align with the substantive research goals in psychology. For example, a domain-specific interpretation method could involve aligning the RF model's item importance scores with clinically meaningful symptom dimensions, as determined by clinicians or existing diagnostic frameworks. Items identified as highly important by the model could be examined to determine whether they correspond to core features of the psychological construct being assessed (e.g., anhedonia in depression). This alignment would help researchers to interpret the model's classification decisions in a manner consistent with diagnostic theory and practice, enhancing the clinical validity and utility of the ML outputs.

6 Conclusion

These findings contribute to the growing body of literature on nonparametric ML methods in psychometric contexts (Classe & Kern, 2024; Gonzalez, 2021; Kraus et al., 2024; Orrù, Gemignani, Ciacchini, Bazzichi, & Conversano, 2020; Quan & Wang, 2026). Our results demonstrate that RF models maintain stable classification performance under adverse data conditions, even when DIF is ignored. These findings suggest that RF models may serve as a valuable complement to parametric frameworks, particularly in applied contexts where

fairness, generalizability, and classification reliability are critical. While RF offers empirical flexibility, it is not a substitute for IRT or more traditional psychometric models, which remain essential for applications requiring latent trait estimation, theoretical alignment, and construct interpretation. Rather than advocating for the replacement of theory-driven models, this study supports a context-sensitive and pluralistic approach to classification, drawing on the strengths of both parametric and nonparametric methods. As psychological assessment increasingly confronts complex and heterogeneous populations, aligning methodological choices with both theoretical and empirical demands will be crucial to ensuring validity, fairness, and replicability in diagnostic classification.

Notes

Correspondence concerning this article should be addressed to Catherine M. Bain and Patrick D. Manapat, Department of Psychology, University of Oklahoma, 455 W. Lindsey Street, Dale Hall Tower, Room 705, Norman, OK 73019, United States. Email: cbain1@ou.edu; pmanapat@ou.edu

Article Information

Conflict of interest disclosures No potential conflicts of interest were reported by the authors.

Funding The authors reported there is no funding associated with the work featured in this article.

Ethical principles The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Data availability statement The simulation R code relevant to this manuscript is published on GitHub: https://github.com/cbain1/difml_open. Empirical data are available upon reasonable request to the corresponding author.

Supplementary Materials The supplementary materials for this article are published on GitHub: https://github.com/cbain1/difml_open.

Acknowledgments

The authors would like to thank Lauren E. Ethridge, Jordan E. Norris, and Alex Conley for their help with data collection. Monte Carlo simulations were performed at the OU Supercomputing Center for Education & Research (OSCER) at the University of Oklahoma (OU). Horst Severini, the Associate Director for Remote & Heterogeneous Computing, and Thang Ha, former Research Computing Facilitator, provided valuable technical expertise on installation of the R-bundle-CRAN module on OSCER.

References

American Psychiatric Association. (2022). *Diagnostic and statistical manual of mental disorders* (DSM-5-TR ed.). American Psychiatric Association Publishing. https://psychiatryonline.org/doi/book/10.1176/appi.books.9780890425787 doi: https://doi.org/10.1176/appi.books.9780890425787

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. doi: https://doi.org/10.1002/jclp.70028

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

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. doi: https://doi.org/10.1136/bmjqs-2023-016996

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

Brown, C., Tollefson, N., Dunn, W., Cromwell, R., & Filion, D. (2001). The adult sensory profile: Measuring patterns of sensory processing. *The American Journal of Occupational Therapy*, *55*(1), 75–82. doi: https://doi.org/10.5014/ajot.55.1.75

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, Calif. London: 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. doi: https://doi.org/10.1016/j.ejpsy.2019.05.002

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. doi: https://doi.org/10.1177/0013164415584576

Classe, F., & Kern, C. (2024). Detecting differential item functioning in multidimensional graded response models with recursive par-

titioning. *Applied Psychological Measurement*, *48*(3), 83–103. doi: https://doi.org/10.1177/01466216241238743

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

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

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. doi: https://doi.org/10.1111/j.1751-9004.2009.00194.x

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Florkowski, C. M. (2008). Sensitivity, specificity, receiver-operating characteristic (roc) curves and likelihood ratios: communicating the performance of diagnostic tests. *The Clinical Biochemist. Reviews*, *29 Suppl 1*(Suppl 1), S83-87.

Garb, H. N. (2021). Race bias and gender bias in the diagnosis of psychological disorders. *Clinical Psychology Review*, *90*, 102087. doi: https://doi.org/10.1016/j.cpr.2021.102087

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*. doi: https://doi.org/10.1111/jnp.12409

Gibbons, R. D., Hooker, G., Finkelman, M. D., Weiss, D. J., Pilkonis, P. A., Frank, E., ... Kupfer, D. J. (2013). The cad-mdd: A computerized adaptive diagnostic screening tool for depression. *The Journal of clinical psychiatry*, *74*(7), 669–674. doi: https://doi.org/10.4088/JCP.12m08338

Golay, P., Abrahamyan Empson, L., Mebdouhi, N., Conus, P., & Alameda, L. (2023). A better understanding of the impact of childhood trauma on depression in early psychosis: A differential item functioning approach. *Schizophrenia Research*, *261*, 18–23. doi: https://doi.org/10.1016/j.schres.2023.09.001

Gonzalez, O. (2021). Psychometric and machine learning approaches for diagnostic assessment and tests of individual classification. *Psychological Methods*, *26*(2), 236–254. doi: https://doi.org/10.1037/met0000317

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. doi: https://doi.org/10.1037/pas0000938

Graham, A. K., Trockel, M., Weisman, H., Fitzsimmons-Craft, E. E., Balantekin, K. N., Wilfley, D. E., & Taylor, C. B. (2019). A screening tool for detecting eating disorder risk and diagnostic symptoms among college-age women. *Journal of American College Health*, *67*(4), 357–366. doi: https://doi.org/10.1080/07448481.2018.1483936

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

Hill, C. (2004). *Precision of parameter estimates for the graded item response*

model [unpublished master's thesis] (Doctoral dissertation, . The University of North Carolina at Chapel Hill). https://catalog.lib.unc.edu/catalog/UNCb4556245

Jacobucci, R., Grimm, K. J., & Zhang, Z. (2023). *Machine learning for social and behavioral research*. New York, NY: The Guilford Press.

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. doi: https://doi.org/10.1177/01466216241238744

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

Liu, R., Huggins-Manley, A. C., & Bulut, O. (2018). Retrofitting diagnostic classification models to responses from irt-based assessment forms. *Educational and Psychological Measurement*, *78*(3), 357–383. doi: https://doi.org/10.1177/0013164416685599

Liu, X. (2012). Classification accuracy and cut point selection. *Statistics in Medicine*, *31*(23), 2676–2686. doi: https://doi.org/10.1002/sim.4509

Lord, F. M. (2012). *Applications of item response theory to practical testing problems*. Hoboken: Taylor and Francis.

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. doi: https://doi.org/10.1177/00131644211063453

Manapat, P. D., Edwards, M. C., MacKinnon, D. P., Poldrack, R. A., & Marsch, L. A. (2021). A psychometric analysis of the brief self-control scale. *Assessment*, *28*(2), 395–412. doi: https://doi.org/10.1177/1073191119890021

Naderalvojoud, B., Curtin, C., Asch, S. M., Humphreys, K., & Hernandez-Boussard, T. (2025). Evaluating the impact of data biases on algorithmic fairness and clinical utility of machine learning models for prolonged opioid use prediction. *JAMIA Open*, *8*(5), ooaf115. doi: https://doi.org/10.1093/jamiaopen/ooaf115

Ohiri, S. C., Momoh, M., Christopher, O., Ikeanumba, I., & Benedict, C. (2024). Differential item functioning detection methods: An overview. *International Journal of Research Publication and Reviews*, *5*(2), 1555–1564. doi: https://doi.org/10.55248/gengpi.5.0224.0505

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*. doi: https://doi.org/10.3389/fmed.2019.00319

Parker-Guilbert, K. S., Leifker, F. R., Sippel, L. M., & Marshall, A. D. (2014). The differential diagnostic accuracy of the ptsd checklist among men versus women in a community sample. *Psychiatry Research*, *220*(1), 679–686. doi: https://doi.org/10.1016/j.psychres.2014.08.001

Patel, T. A., Robison, M., & Cougle, J. R. (2024). Item response theory analysis and differential item functioning of the social appearance anxiety scale. *Assessment*, 10731911241306370. doi: https://doi.org/10.1177/10731911241306370

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. doi: https://doi.org/10.1177/0146621619858675

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. doi: https://doi.org/https://doi.org/10.1111/bmsp.70041

R Core Team. (2024). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

Raskin, R., & Terry, H. (1988). A principal-components analysis of the narcissistic personality inventory and further evidence of its construct validity. *Journal of Personality and Social Psychology*, *54*(5), 890–902. doi: https://doi.org/10.1037//0022-3514.54.5.890

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. doi: https://doi.org/10.3389/fpsyg.2021.709928

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. doi: https://doi.org/10.1007/s11136-022-03169-0

Samejima, F. (1997). Graded response model. In W. J. Van Der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (p. 85–100). New York, NY: Springer New York. doi: https://doi.org/10.1007/978-1-4757-2691-6_5

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. doi: https://doi.org/10.1016/j.rasd.2020.101669

Schröder, A., Vulink, N., & Denys, D. (2013). Misophonia: Diagnostic criteria for a new psychiatric disorder. *PLoS ONE*, *8*(1), e54706. doi: https://doi.org/10.1371/journal.pone.0054706

Sen, S., & Cohen, A. S. (2024). An evaluation of fit indices used in model selection of dichotomous mixture irt models. *Educational and Psychological Measurement*, *84*(3), 481–509. doi: https://doi.org/10.1177/00131644231180529

Shen, W., Kiger, T. B., Davies, S. E., Rasch, R. L., Simon, K. M., & Ones, D. S. (2011). Samples in applied psychology: Over a decade of research in review. *Journal of Applied Psychology*, *96*(5), 1055–1064. doi:

https://doi.org/10.1037/a0023322

Sims, R., Michaleff, Z. A., Glasziou, P., & Thomas, R. (2021). Consequences of a diagnostic label: A systematic scoping review and thematic framework. *Frontiers in Public Health*, *9*, 725877. doi: https://doi.org/10.3389/fpubh.2021.725877

Smits, N., Smit, F., Cuijpers, P., & De Graaf, R. (2007). Using decision theory to derive optimal cut-off scores of screening instruments: an illustration explicating costs and benefits of mental health screening. *International Journal of Methods in Psychiatric Research*, *16*(4), 219–229. doi: https://doi.org/10.1002/mpr.230

Spann, D. J., Cicero, D. C., Straub, K. T., Pellegrini, A. M., & Kerns, J. G. (2024). Examining measures of schizotypy for gender and racial bias using item response theory and differential item functioning. *Schizophrenia Research*, *272*, 120–127. doi: https://doi.org/10.1016/j.schres.2024.08.015

Strobl, C., Hothorn, T., & Zeileis, A. (2009). Party on! a new, conditional variable importance measure available in the party package. *The R Journal*, *2*, 14–17.

Terry-McElrath, Y. M., & Patrick, M. E. (2018). Simultaneous alcohol and marijuana use among young adult drinkers: Age-specific changes in prevalence from 1977 to 2016. *Alcoholism: Clinical and Experimental Research*, *42*(11), 2224–2233. doi: https://doi.org/10.1111/acer.13879

Wakefield, J. C. (2010). False positives in psychiatric diagnosis: implications for human freedom. *Theoretical Medicine and Bioethics*, *31*(1), 5–17. doi: 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.

Wardell, J. D., Cunningham, J. A., Quilty, L. C., Carter, S., & Hendershot, C. S. (2020). Can the audit consumption items distinguish lower severity from high severity patients seeking treatment for alcohol use disorder? *Journal of Substance Abuse Treatment*, *114*, 108001. doi: https://doi.org/10.1016/j.jsat.2020.108001

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*. doi: https://doi.org/10.3389/fpsyg.2022.897901

Woods, C. M., Cai, L., & Wang, M. (2013). The langer-improved wald test for dif testing with multiple groups: Evaluation and comparison to two-group irt. *Educational and Psychological Measurement*, *73*(3), 532–547. doi: https://doi.org/10.1177/0013164412464875

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. doi: https://doi.org/10.1002/jclp.22098

Yavuz Temel, G. (2023). A simulation and empirical study of differential test functioning (dtf). *Psych*, *5*, 478–496. doi:

https://doi.org/10.3390/psych5020032

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. doi: https://doi.org/10.1093/jpepsy/jst062

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