Article

Using Interpretable Machine Learning for Differential Item Functioning Detection in Psychometric Tests

Applied Psychological Measurement
2024, Vol. 48(4-5) 167–186

DOI: 10.1177/01466216241238744
journals.sagepub.com/home/apm


Elisabeth Barbara Kraus[1], Johannes Wild[2], and Sven Hilbert[2]

Abstract

This study presents a novel method to investigate test fairness and differential item functioning combining psychometrics and machine learning. Test unfairness manifests itself in systematic and demographically imbalanced influences of confounding constructs on residual variances in psychometric modeling. Our method aims to account for resulting complex relationships between response patterns and demographic attributes. Specifically, it measures the importance of individual test items, and latent ability scores in comparison to a random baseline variable when predicting demographic characteristics. We conducted a simulation study to examine the functionality of our method under various conditions such as linear and complex impact, unfairness and varying number of factors, unfair items, and varying test length. We found that our method detects unfair items as reliably as Mantel–Haenszel statistics or logistic regression analyses but generalizes to multidimensional scales in a straight forward manner. To apply the method, we used random forests to predict migration backgrounds from ability scores and single items of an elementary school reading comprehension test. One item was found to be unfair according to all proposed decision criteria. Further analysis of the item's content provided plausible explanations for this finding. Analysis code is available at: https://osf.io/s57rw/?view_only=47a3564028d64758982730c6d9c6c547.

Keywords

psychometrics, machine learning, interpretable machine learning, random forest, test fairness, differential item functioning

[1]LMU Munich, Germany
[2]University of Regensburg, Germany

Corresponding Author:
Elisabeth Barbara Kraus, Chair for Computational Modeling in Psychology, LMU Munich, Akademiestr. 7, München 80799, Germany.
Email: e.kraus@psy.lmu.de

Introduction

Fairness has received considerable attention in the social sciences (Mashek & Hammer, 2011). It is conceptually divided into measurement and decision fairness (Kuppler et al., 2022), which are defined in different ways. The latter—also called algorithmic fairness—tackles the question of how personal attributes should be related to a target score gained from individual data which is then used in a decision about individual access to resources (Corbett-Davies & Goel, 2018). In contrast, the definition of fairness in the context of measurement refers to validity—in particular, convergent, discriminant, and consequential validity (Jacobs & Wallach, 2021). Thus, fairness in measurement refers to the relationship between the value of a real-world attribute and its associated measured score. To understand how fairness and validity intertwine, it is necessary to embed our understanding of fairness into measurement theory.

Measurement Concepts and Fairness

In the social sciences, especially in psychology and education, the concept of latent constructs is established. A construct refers to a personal attribute that can represent an ability, a trait, or a state (Malik, 2013; Spearman, 1904). In the context of psychometric testing, latent constructs are called abilities. This translates to the fact that a correct response to a test item requires the ability represented by the latent variable and further influences that are not identified or named. These further influences are therefore conceptualized as residual variance. The residual variance can stem from other latent constructs, such as other abilities, but also from traits such as a preference for guessing (Attali & Bar-Hillel, 2003, Penfield & Camilli, 2006). Finally, other influences also comprise random error. Test fairness within this framework arises from the characteristics of the residual variances. In psychometric models, it is assumed that one or more latent constructs are measured by the test and that the residual variance is due to random error. However, this assumption can be wrong and residual variance may stem from one or more additional latent constructs and be related to demographics. This is called a bias (Penfield & Camilli, 2006) or differential item functioning (DIF). On a mathematical level, DIF results in multidimensionality of a test. The latent constructs introducing the DIF act as additional latent constructs, that systematically influence the item response behavior. With multiple additional latent constructs, they may also be intertwined and have interactive effects on the test item responses.

Fairness in testing is a manifold construct and depends on the testing situation, social factors, but also the test itself. Sticking to this last notion, fairness can be connected to the distributions of these additional latent constructs (which are not intended to be measured by a test): A psychometric test is called *biased/multidimensional* if some of its items measure other abilities, which are *evenly distributed* among population groups. In contrast, a test is *unfair* if some of its items require the use of other abilities (not intended to be measured by the test) that are *unevenly distributed* among populations (Dorans & Cook, 2016). Unevenly distributed can refer to mean differences between population groups on the other ability but also all other functional relationship between the additional latent construct and group membership. Thus, an unfair test is always biased/multidimensional but a biased/multidimensional test does not need to be unfair.

Test Fairness is a Matter of Test Validity

Aspiring test fairness is crucial in test development because test fairness is a matter of validity (Dorans & Cook, 2016). Validity refers to the meaningfulness of a test (Messick, 1995). Moreover, validity comprises multiple facets of which content, convergent, discriminant, and consequential validity are particularly important because a deficit in these facets directly threatens the implications of a test score (Bryant, 2000).

As described above, an unfair test lacks content validity because group differences in the residual variance add undesirable multidimensionality to the test. In consequence, the test measures two abilities as the same time instead of measuring a single intended to be measured ability. In addition, an unfair test lacks convergent validity. This means that the unfair test score is (also statistically) less strongly related to unbiased tests of similar latent variables—let us call them convergent tests—than expected. The lack of convergent validity, again, is rooted in the additional latent construct (not intended to be measured by the test). Imagine that this additional latent construct influences the items of the unfair test but not the items of the convergent test. Now imagine that one population group has higher values on the additional latent construct on average. In consequence, the test items' values would also be higher in this group, whereas they would be lower in the second group. These performance differences would change the order of the scores of the test takers in the unfair test and therefore diminish the correlation between the two and, in turn, reduce the metric for the convergent validity. This becomes even more relevant, if multiple, possibly interacting additional latent constructs impact the test item responses and distort the expected order of test scores in a non-linear manner.

The same holds for discriminant validity. Unfairness may result in at least one additional latent variable, which is inseparable from the original one within a given group. For example, imagine taking a biology test in a language you do not know. Then your test score would be completely dominated by your language ability and consequently, your biology ability could not be separated from your foreign language ability—even though language ability and biology ability are unlikely to be related on a conceptual level. In consequence, reduced convergent and divergent validity can lead to false conclusions on an individual level, when convergent and divergent tests are used in diagnosing individuals, but also when using these measures in research (Carlson & Herdman, 2012).

Finally, when scores derived from unfair tests are used to inform decisions, consequential validity of the scores is at stake. Consequential validity refers to the implications test scores have in terms of actual or expected consequences. These consequences can result from expectations based on the test scores (e.g., high educational achievement expectations resulting from high IQ scores) but also from decisions based on test scores (e.g., hiring a person based on an aptitude test). This is where measurement fairness and decision fairness meet (Messick, 1989).

Standard Practices to Handle Unfair Test Items

Standard practices to handle unfair test items during test revision are to identify them by different kind of item difficulty analyses (e.g., Mantel–Haenszel (MH)-statistics, (Holland & Thayer, 1988)), differential item functioning (DIF) analyses during psychometric modeling, and logistic regression on single item responses using the remaining item responses, total scores as well as demographic variables as predictors (Dorans, 2004; Magis et al., 2010).

In MH statistics, the performance of two or more groups is compared on each test item, after controlling for overall ability level. This is done by creating a contingency table for each item, with rows for the two or more groups and columns for the two possible responses (e.g., correct or incorrect). The cells of the table represent the number of individuals in each group who gave each response. MH statistics then compute a weighted average of the odds ratios from these contingency tables, where the weights are proportional to the sample sizes of the groups being compared. If the odds ratio is significantly different from 1, it is indicated that the item is functioning differently for the two groups, even after controlling for ability level.

In logistic regression, single item responses are explained by the total score, group membership, and the interaction of the two. The idea behind using logistic regression to identify DIF is

to test the interaction effect for significance and thereby to test if item response probabilities are different in different groups, even when controlled for the overall ability level.

However, all the above-mentioned techniques come with different shortcomings when using them to identify single unfair items. First, item difficulty comparisons as well as graphical model tests in the context of IRT-modeling cannot account for valid ability differences between population groups. MH statistics and logistic regressions additionally require a lot of hypotheses testing, which leads to a need for Type I error correction and in consequence to reduced power. Also results may vary according to different choices of the criterion representing the overall ability level (Clauser & Mazor, 1998), and most importantly standard procedures do not readily generalize to the multidimensional case (Clauser et al., 1996).

New Trends in Psychometrics and DIF

In the last decade, modern data analysis techniques like Machine Learning (ML) entered the field of psychology and education (Hilbert et al., 2021), as well as psychometrics and DIF. Not only have they been shown to be usable to find the number of factors in exploratory factor analysis (Goretzko & Bühner, 2020) and help in identifying cheating behavior (Man et al., 2019), but also have they been shown to help identifying DIF.

Bauer et al. for instance, used regularization in a logistic regression approach to DIF items, by regularizing all regression coefficients for the interaction effects simultaneously (Bauer et al., 2020; Belzak, 2022; Belzak & Bauer, 2020; Tutz & Schauberger, 2015). Others have used recursive partitioning methods, like decision trees for DIF identification (Strobl et al., 2015). In their model, changes in item parameters of IRT models are tested for significance after recursive partitioning of the sample with respect to different demographic attributes. Thereby, relevant demographic attributes introducing DIF can be identified. Their DIF trees can account for complex relationships between item response patterns and various demographic attributes and overcome the problem of intensive hypotheses testing by using ML procedures. Yet, the advanced DIF techniques presented here are of exploratory nature and mostly capture DIF as a characteristic of the whole test. Their focus is not to identify single items, that cause the DIF. Consequently, they should preface the localization of test unfairness by answering the question of whether there is multidimensionality or unfairness in measurement, rather than answering the question of where potential unfairness might come from.

Proposition of a Novel Method

In this article we propose a novel method to identify DIF by determining unfair items in psychometric tests. Our approach also uses ML methods, namely, random forests (RF) and permutation variable importance measures.

In comparison to single decision trees, RF is an ML method that combines multiple decision trees. The method was proposed by Breiman in 2001 and builds a set of decision trees to make predictions by combining the results from each tree. The decision tree algorithm is a nonparametric method that partitions the sample into subgroups by splitting predictors in a binary manner, grouping subjects with similar properties together. This process aims to maximize the difference between the impurity in the sub-nodes and the impurity in the parent node. Impurity in classification trees is commonly defined as the misclassification rate in the nodes. The decision tree building process ends when a minimum size of subjects is left in a node or when a minimum change in the impurity measure after a split is achieved (Probst et al., 2019). Randomness enters the decision trees on two levels: (1) for every split, the best feature and the best splitting value of the feature are selected from a randomly drawn subsample of the features and (2) individual trees

are a grown for repeated bootstrap samples of the observations. Thereby, an RF provides a relatively conservative and precise estimate of the error rate, known as the out-of-bag error (OOB) rate (Breiman, 1996). It is computed by passing the unsampled observations through the trees and using their misclassification rates to estimate the generalization error. The final prediction is achieved by averaging the results from each tree (e.g., Breiman, 2001).

In our method, we use RF to predict demographic variables like socioeconomic status, or migration background from the set of test items, a random variable, and the test score. Its advantage over existing procedures is that many ML algorithms, such as RF, can identify even nonlinear and complex relationships without having to test a large number of models. We argue, that if a demographic variable is complexly related to the item responses, there are response patterns that are not captured by the latent variable but related to the demographic variable. Such response patterns would be expected in multidimensional tests because then an additional latent construct would influence item responses and produce specific response patterns. Finally, an association of this additional latent construct with the demographic variable would indicate test unfairness. Therefore, we evaluate the variable importance of every single predictor in comparison to a random variable. Variable importance measures can be used to shed light on what would otherwise be considered black box models and to learn something about the structure of item responses (Molnar, 2020). They can be used to identify important variables in the prediction by measuring the impact of permuting each feature on the model's accuracy. This process is known as permutation variable importance. It shuffles the values of a single feature and refits the RF model. The decrease in model accuracy after refitting is used as a measure of the importance of the feature. Thus, high relative variable importance values mean a high decrease in accuracy and speak for high relevance of the predictor variable associated. To obtain a more accurate estimate of the variable importance and the estimation error, this procedure is repeated multiple times and the average and standard deviation of the results are used as estimators.

In our method we investigate the permutation variable importance of the test items and the random variable. After all, unfair items will show high permutation variable importance, as they are related to the demographic variable beyond their relation to the latent variables. But fair items and the random variable will produce low variable importance scores because they are unrelated to possible group differences in underlying unfairness provoking latent concepts.

By application of variable importance measures to the proposed ML models, unfair items can be identified. As a valid relationship between the latent variable and the demographic variable may exist in terms of *impact* (Gipps & Stobart, 2009), we investigate, if the latent variable scores as additional predictors help to express justified differences between demographic groups. To showcase our procedure, we conducted a simulation study and applied the method to a reading comprehension test, whose fairness to test takers with a migration background was in question.

Simulation Study

The following simulation study was conducted to empirically investigate the statistical properties of the novel DIF detection method. All simulations were run in R (R Core Team, 2022), using the packages *mirt* (Chalmers, 2012), *randomForest* (Liaw & Wiener, 2002), *mvtnorm* (Genz et al., 2021), *iml* (Molnar et al., 2018), *data.table* (Dowle & Srinivasan, 2021), *tidyr* (Wickham, 2021), *ggplot2* (Wickham, 2016), *ggpubr* (Kassambara, 2020), and *difR* (Magis et al., 2010).

Datasets with item responses following different response patterns were created and analyzed. By predicting a group variable with RF and evaluating variable importance scores in reference to a random variable, the proposed method was examined. The aim was to investigate if the proposed method could identify items that were simulated to have an error component related to unequally distributed additional constructs. Finally, results were compared to the standard DIF-methods MH and logistic regression.

Experimental Conditions and Simulation Setup

Data Generation Models. Item responses were simulated according to a multiple factor structure composed of one or two latent constructs (depending on the condition). DIF was introduced by assigning some items to one additional factor, not intended to be measured by the test. These DIF items were chosen randomly. All latent factors were simulated to be uncorrelated. The model equation for the simulation of the item responses was

$$p\left(y_i = 1;\, \theta_p\right) = \frac{e^{a_i' \theta_p + d_i}}{1 + e^{a_i' \theta_p + d_i}}$$

$p\left(y_i = 1;\, \theta_p\right) = probability\ to\ solve\ item\ i$
$\theta_p = vector\ of\ one, two, or\ three\ person\ parameters$
$a_i' = transposed\ loading$
$d_i = difficulty\ parameter$

(Van der Linden, 2016). Person and difficulty parameters were drawn from a standard normal distribution. Loading parameters were set to two to assign items to factors, or to zero—for items unrelated to that factor. Second, the group membership variable was simulated according to different underlying relationships with the person parameters. The conditions with no additional factor constituted the null conditions. There were three null conditions (1–3), that differed in terms of the relationship between the person parameters of the intended factors and the group membership, which were modeled to be either random, linear, or quadratic with respect to the centered logits of group membership probability. By assuming a linear or quadratic relationship, impact conditions were realized. Impact means that one group performs better than the other on the intended latent construct. Figure 1 visualizes the four possible resulting models with either one, two, or three dimensions.

When simulating multidimensional response data, person parameters of the not-intended, additional, factor were used to model different relationships with the group membership variable. In the fair conditions no relationship between the DIF items and the group membership was simulated (conditions 4–6). In the unfair conditions, two variations of the relationship between the person parameters of the additional factor and the group membership variable were modeled: linear and quadratic (conditions 7–12). This resulted in a total of 12 different relationships between person parameters and group membership shown in Table 1.

Data Generation Procedure. The experiment was set up with person parameter values simulated according to independent multivariate normal distributions with 1000 observations per data frame. Each data frame comprised the items, a Bernoulli random variable, the group variable, and estimated person parameter variables. Furthermore, the simulation study comprised different conditions resulting from variations of design elements sought to induce robustness. Specifically, the test length (8 items, 20 items), the number of multidimensional/DIF items loading on the additional factor (0, 2, 4) as well as the number of intended factors (1, 2) were varied and resulted in a 3 (impact) × 4 (fairness) × 3 (additional items) × 2 (number of factors) = 144 conditional design. As however conditions without additional factor could not be separated according to fairness, 2 (test length) × 3 (fairness) × 3 (impact) × 2 (number of factors) = 36 conditions were deleted from the design, resulting in a total of 108 conditions. Every condition was replicated 50 times in the 8-item and 30 times in the 20-item condition.

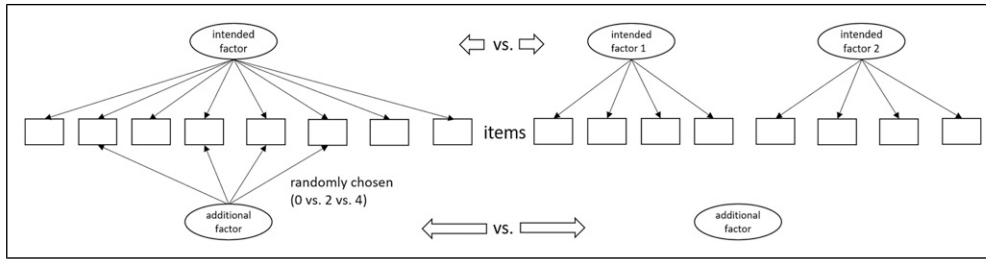


Figure 1. Data generation models.

Table 1. Simulation of Response Data Structure with respect to the Group Membership Variable.

Multi-dimensionality (θ_{ad})	Impact (θ)	Name of condition	Formula for all conditions: $x_{i\,group} \sim Ber(p)$, *for all i*
No	No	(1) Null condition	$p = 0.5$
No	Linear	(2) Null condition impact linear	$p = \frac{e^{\theta}}{1 + e^{\theta}}$
No	Quadratic	(3) Null condition impact quadratic	$p = \frac{e^{\theta+\theta^2-\mathbb{E}(\theta+\theta^2)}}{1 + e^{\theta+\theta^2-\mathbb{E}(\theta+\theta^2)}}$
Fair	No	(4) Null condition fair	$p = 0.5$
Fair	linear	(5) Fair impact linear	$p = \frac{e^{\theta}}{1 + e^{\theta}}$
Fair	Quadratic	(6) Fair impact quadratic	$p = \frac{e^{\theta+\theta^2-\mathbb{E}(\theta+\theta^2)}}{1 + e^{\theta+\theta^2-\mathbb{E}(\theta+\theta^2)}}$
Unfair linear	No	(7) Unfair linear	$p = \frac{e^{\theta_{ad}}}{1 + e^{\theta_{ad}}}$
Unfair linear	Linear	(8) Unfair linear impact linear	$p = \frac{e^{\theta_{ad}+\theta-\mathbb{E}(\theta_{ad}+\theta)}}{1 + e^{\theta_{ad}+\theta-\mathbb{E}(\theta_{ad}+\theta)}}$
Unfair linear	Quadratic	(9) Unfair linear, impact quadratic	$p = \frac{e^{\theta_{ad}+\theta+\theta^2-\mathbb{E}(\theta_{ad}+\theta+\theta^2)}}{1 + e^{\theta_{ad}+\theta+\theta^2-\mathbb{E}(\theta_{ad}+\theta+\theta^2)}}$
Unfair quadratic	No	(10) Unfair quadratic	$p = \frac{e^{\theta_{ad}+\theta_{ad}^2-\mathbb{E}(\theta_{ad}+\theta_{ad}^2)}}{1 + e^{\theta_{ad}+\theta_{ad}^2-\mathbb{E}(\theta_{ad}+\theta_{ad}^2)}}$
Unfair quadratic	Linear	(11) Unfair quadratic impact linear	$p = \frac{e^{\theta_{ad}+\theta_{ad}^2+\theta-\mathbb{E}(\theta_{ad}+\theta_{ad}^2+\theta)}}{1 + e^{\theta_{ad}+\theta_{ad}^2+\theta-\mathbb{E}(\theta_{ad}+\theta_{ad}^2+\theta)}}$
Unfair quadratic	Quadratic	(12) Unfair quadratic impact quadratic	$p = \frac{e^{\theta_{ad}+\theta_{ad}^2+\theta+\theta^2-\mathbb{E}(\theta_{ad}+\theta_{ad}^2+\theta+\theta^2)}}{1 + e^{\theta_{ad}+\theta_{ad}^2+\theta+\theta^2-\mathbb{E}(\theta_{ad}+\theta_{ad}^2+\theta+\theta^2)}}$

After simulation of the response data, the factor models were specified according to the structure of the intended factors given in the data creation process, but ignoring possible additional factors. Intended loadings were freely estimated with the EM algorithm (Bock & Aitkin, 1981) from the *mirt* package (Chalmers, 2012). Most models had at least acceptable model fit (all 1st quartile $CFI_{8\text{-item}} > .94$; $CFI_{20\text{-item}} > .97$, all 3rd quartile $RMSEA_{8\text{-item}} < .08$; $RMSEA_{20\text{-item}} < .06$). Reliability of the person parameters ranged between .65 and .77 in the 8-item conditions and between .79 and .88 in the 20-item conditions. Person parameters were gained via WLE estimation (Warm, 1989). Finally, a random Bernoulli distributed variable with $p = .5$ was added to the dataset to serve as the baseline for the consequent RF analyses.

Random Forests and Variable Importance. The created datasets were then used to predict their group membership variable with RF applying the randomForest() function from the *RandomForest* package (Liaw & Wiener, 2002). Each RF was fit once using the item responses and the random variable as predictors only and once using the item responses, the random variable and estimated person parameters as predictors. RF were not tuned using an mtry parameter of $\sqrt{p}$ (p denoting the total number of variables in the model), a bootstrap sample of observations, 500 trees per forest, and unconstrained tree depth. Subsequently, ten iterations of permutation variable importance scores and confidence intervals for all predictors were computed and evaluated according to the variable importance of the random variable.

Criterion Variables

To assess the functionality of the novel method, the following evaluation criteria were recorded and assessed. The distributions of importance scores of items in the permutation variable importance analyses were analyzed graphically, conditional on properties of the characteristics of the datasets (number of intended factors, number of additional items, and item difficulty), and in reference to a random baseline variable. Therefore, results were combined over conditions.

After these descriptive evaluations false-positive-rates (fpr) and true-positive-rates (discovery rates) were calculated based on the random variable criterion. Consequently, conditions were collapsed over number of factors and number of additional items. Whenever a single item's variable importance was higher than the random variable's importance with non-overlapping confidence intervals, items were coded as unfair DIF items.

Benchmark Experiment

In a final analysis, false positive and discovery rates of our novel method were compared to fpr and discovery rate of MH statistics and logistic regression. Therefore, these statistics were computed on the same simulated datasets, as described and implemented by Magis et al. (2010).

Simulation Study Results

Visual inspection of the distributions of importance scores showed higher variable importance for unfair items in the unfair test conditions compared to fair items and the random baseline variables. As displayed in Figure 2, there was no apparent difference in variable importance between fair items that loaded on the additional factor (multidimensional items) and such that did not load on the additional factor. In comparison, multidimensional items had higher importance scores in the unfair test conditions.

Interestingly, whenever included, the person parameters (thetas) were found to have consistently high variable importance. This may be due to the impact conditions, but also due to the relatively high correlation among items, which is caused by underlying common latent variable(s). This high correlation allows items to be easily replaced by others in the single decision trees, resulting in a smaller decrease in accuracy when their values are shuffled during variable importance analyses (Strobl et al., 2008).

There was no relevant relationship ($r = -.03, p < .001$) between the decision accuracy and the number of intended factors, and no relevant relationship between decision accuracy and the number of additional items ($r = -.06, p < .001$). Finally, binned item difficulty was plotted against the accuracy, and fpr, and a slight curvilinear relationship between item difficulty and the fpr could be observed (see Figure 3).

Finally, discovery rates and fpr were determined and compared to MH and logistic regression criteria applied to the same datasets. Conditions were therefore combined across the variations in number of factors and number of items loading on the additional factor. The three remaining

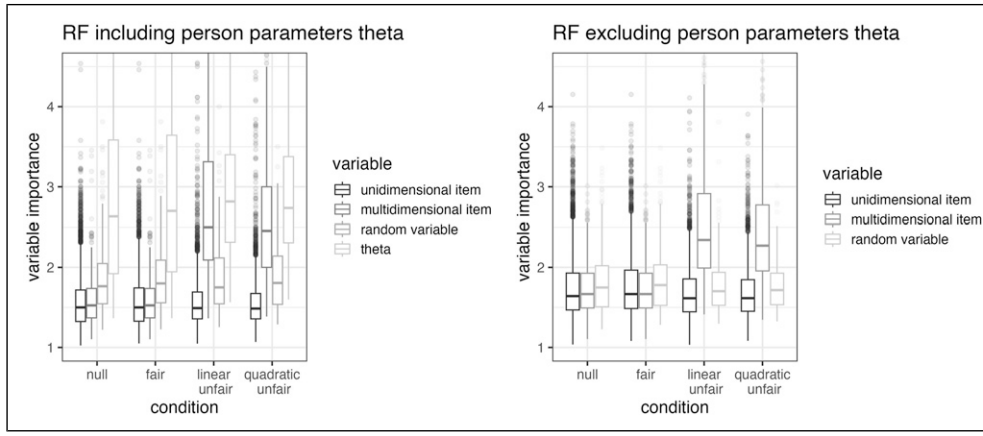


Figure 2. Distributions of variable importance scores. Note. Theta = person parameter of intended factors; RF = random forest; boxes show the 25th, 50th, and 75th percentiles.

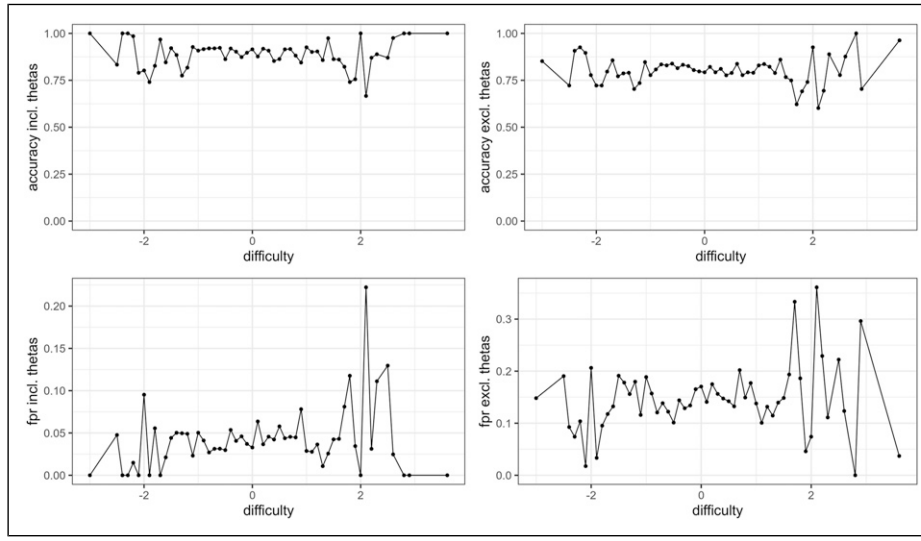
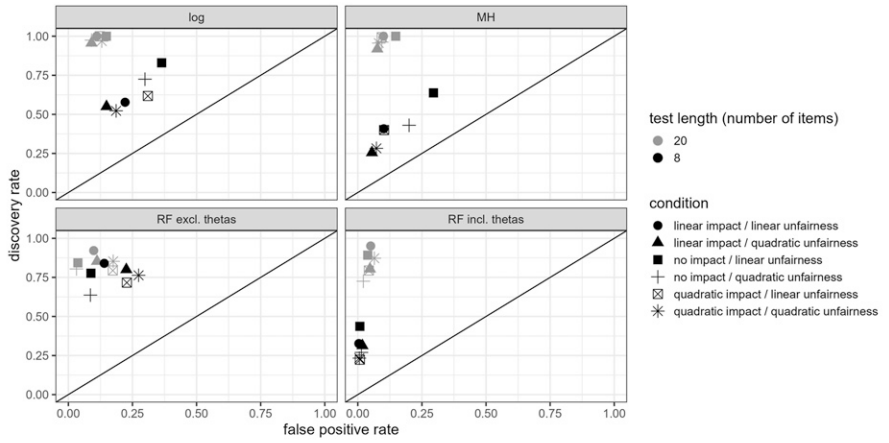


Figure 3. Relationships between item difficulty, accuracy, and false-positive-rates (fpr).

variations had a 12-factorial design, with variations in test length, impact and the test being unidimensional, fair, unfair with linear relationships, or unfair with quadratic relationships between the additional latent construct and the group variable. Then, an item was defined as an unfair DIF item in the unfair test condition, whenever its importance exceeded the importance of the random variable. Exceeding was defined as the confidence interval of the variable importance having no overlap with the random variable. The discovery rate was defined as the rate of detecting unfair items in the unfair test conditions and is presented in Figure 4.

The fpr was computable in all test conditions. The fpr was determined by the rate of unidimensional or multidimensional but fair items falsely identified as unfair items. It is presented in Figures 4 and 5. The RF including thetas yielded the lowest fpr in most conditions but fell short in terms of discovery rates in the 8-item condition, while discovery rates in the 20-item condition



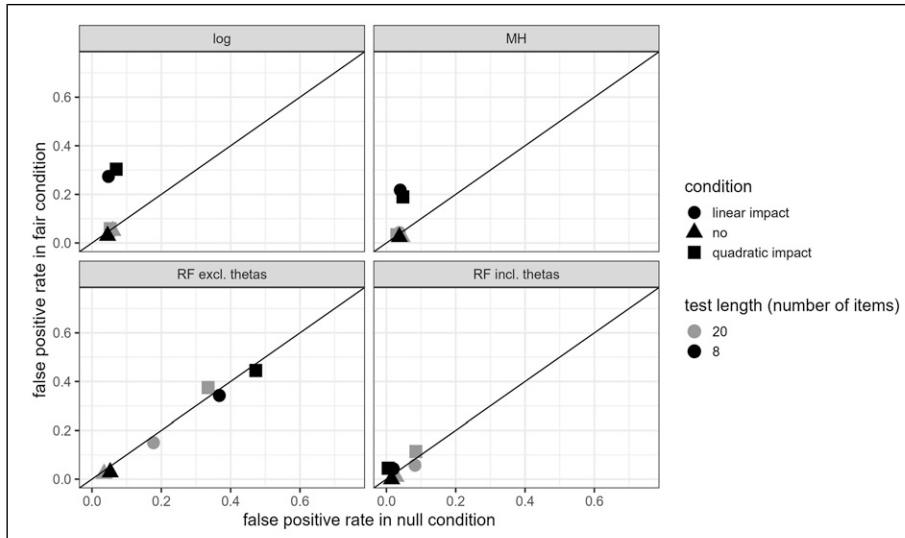
conditions			RF incl. thetas		RF excl. thetas		MH		Logistic Regression	
impact	*unfairness*	*number of items*	*fpr*	*dr*	*fpr*	*dr*	*fpr*	*dr*	*fpr*	*dr*
no	linear	20	0.04	0.89	0.04	0.84	0.15	1.00	0.15	1.00
linear	linear	20	0.05	0.95	0.10	0.92	0.10	1.00	0.11	1.00
quadratic	linear	20	0.04	0.79	0.17	0.79	0.09	0.99	0.12	1.00
no	quadratic	20	0.02	0.73	0.03	0.80	0.09	0.96	0.09	0.97
linear	quadratic	20	0.05	0.80	0.11	0.85	0.08	0.92	0.09	0.96
quadratic	quadratic	20	0.06	0.87	0.17	0.85	0.08	0.96	0.13	0.97
no	linear	8	0.01	0.44	0.09	0.78	0.29	0.64	0.36	0.83
linear	linear	8	0.00	0.33	0.14	0.84	0.10	0.41	0.22	0.58
quadratic	linear	8	0.01	0.22	0.23	0.72	0.10	0.40	0.31	0.62
no	quadratic	8	0.01	0.27	0.09	0.64	0.20	0.43	0.30	0.72
linear	quadratic	8	0.02	0.31	0.23	0.80	0.06	0.26	0.15	0.55
quadratic	quadratic	8	0.01	0.23	0.27	0.76	0.07	0.28	0.19	0.52

Figure 4. Discovery (dr) and false-positive-rates (fpr) for random forest (RF) including and excluding thetas and benchmark methods of Mantel–Haenszel (MH) and logistic regression (log) in unfair conditions.

were all higher than .73. RF excluding thetas was least sensitive to test length but could not reach as high discovery rates as logistic regression or MH in the 20-item condition. The fpr in null and fair conditions were generally lower in the 20-item conditions, which was especially pronounced for logistic regression and MH in fair impact conditions. RF excluding thetas showed the highest fpr of all methods.

Simulation Study Discussion

Overall, the proposed method worked well across varying conditions. Unfair items had substantially higher variable importance than fair or unidimensional items. In fair tests, multidimensional and unidimensional items descriptively showed comparable variable importance. Person parameters showed consistently high variable importance, which might be due to the nature of RF artificially preferring numeric features in variable importance analyses (Strobl et al., 2009). The method



conditions		RF incl. thetas		RF excl. thetas		MH		Logistic Regression	
impact	*number of items*	*fpr_null*	*fpr_fair*	*fpr_null*	*fpr_fair*	*fpr_null*	*fpr_fair*	*fpr_null*	*fpr_fair*
no	20	0.03	0.01	0.04	0.02	0.04	0.03	0.06	0.05
linear	20	0.08	0.06	0.18	0.15	0.04	0.04	0.05	0.04
quadratic	20	0.09	0.11	0.34	0.38	0.03	0.03	0.05	0.06
no	8	0.01	0.00	0.05	0.03	0.04	0.03	0.04	0.03
linear	8	0.02	0.04	0.37	0.34	0.04	0.22	0.05	0.27
quadratic	8	0.01	0.04	0.47	0.45	0.05	0.19	0.07	0.30

Figure 5. False-positive-rates (fpr) for random forest (RF) including and excluding thetas and benchmark methods of Mantel–Haenszel (MH) and logistic regression (log) in null- and fair conditions.

showed varying discovery rates across the tested conditions and the fpr exceeded the desirably threshold of five percent, when person parameters were included. However, higher confidence levels in the confidence intervals for the variable importance would be a solution to this because they raise the decision boundary. In fact, the confidence intervals for the variable importance are one tuning factor and allow for one's own weighing of fpr and discovery rates according to the application setting at hand. Fpr and discovery rates were unrelated to the number of factors or number of multidimensional items but slightly related to item difficulty. Therefore, it could be reasonable to change the confidence level of the confidence intervals for variable importance to account for item difficulty with respect to the item difficulty of the specific items at hand. Yet, increased fpr for items with medium difficulty is a common problem for most standard DIF-procedures (Mazor et al., 1998), so this aspect is not exclusive to our method.

Further, our method slightly outperformed MH and logistic regressions in some unfair test conditions but yielded worse (RF excluding thetas) or better (RF including thetas) results in the null and fair conditions. All compared methods were insensitive to multidimensionality in the 20-item condition (see Figure 5) and yielded comparable fpr in the fair and null conditions. In the 8-item conditions all but RF including thetas were somewhat sensitive to multidimensionality with increased fpr in the fair conditions. However, the RF analyses excluding thetas showed the most pronounced differences between discovery rates and fpr in the short test length condition. This might illustrate the problem that undesired variance from unfair and unintended additional latent

constructs might be absorbed in the estimation of intended factors and consequently hard to be identified by the RF analysis. Apparently, in the impact conditions, the RF were only partly able to implicitly learn the intended factor structure without adding the person parameters explicitly to the predictor set. This can be understood by the fact that impact conditions had much higher fpr than no impact conditions for the RF excluding person parameters (see Figure 5).

Application Study

In the following section we present an application of the novel method to a real-life dataset. We predicted the migration status of 4,141 test takers of a reading comprehension test and analyzed potentially unfair items on a content level. To underpin content-related hypotheses about possible sources of unfairness in reading comprehension tests a short introduction to the construct of reading comprehension is given in the following section.

Application to the Construct of Reading Comprehension

To understand which parts of the reading comprehension items' variance components should be related to the latent variable and which parts should not, a short introduction to the theory of reading comprehension is given in the following section. Following prominent large-scale studies such as Progress in International Reading Literacy Study (PIRLS) or Programme for International Student Assessment (PISA), reading comprehension is conceptualized in five successive levels of competence (Mullis et al., 2015). Mastery of these levels requires the ability to decode (Level I), information extraction (Level II), sentence-level inferencing (Level III), section-level inferencing and integration of general knowledge (Level IV), and building a mental model including metaknowledge about texts (Level V). In addition, it has been shown that other constructs—such as vocabulary and background knowledge—influence reading comprehension without being a part of the latent construct of reading comprehension. This is summarized, for example, in the direct and inferential mediation (DIME) model (Cromley & Azevedo, 2007). So, vocabulary and background knowledge are constructs through which unfairness can enter the reading comprehension test items (see Figure 6). For example, text topics may be chosen to be associated with cultural heritage or cultural knowledge that test takers with a migration background are less likely to have. The same is true for specific vocabulary that is uncommon in everyday language use (Becker et al., 2013). If items containing these characteristics show high variable importance, the test is to be considered unfair.

Since cultural knowledge and vocabulary are assumed to converge between persons with and without migration background along with the number of migration generations (Lüdemann & Schwerdt, 2013), the comparison of different generations of migration backgrounds can show the robustness of the analyses.

In summary, using our method to detect unfair items in a reading comprehension test regarding migration background is expected to reveal items that are linked to cultural knowledge or rarely used vocabulary. It is further expected that the sets of unfair items revealed for test takers with a second-generation and first-generation migration background overlap with a stronger influence for those with a first-generation migration background.

Application Study Methods

Sample. The sample comprised 4,141 students and was collected during a longitudinal reading intervention study between 2018 and 2021. The sample was semi-random, as school directors applied for participation in the study, but neither teachers nor students (test takers) did. The test takers' grade levels ranged from second grade to seventh grade. Approximately half of the sample

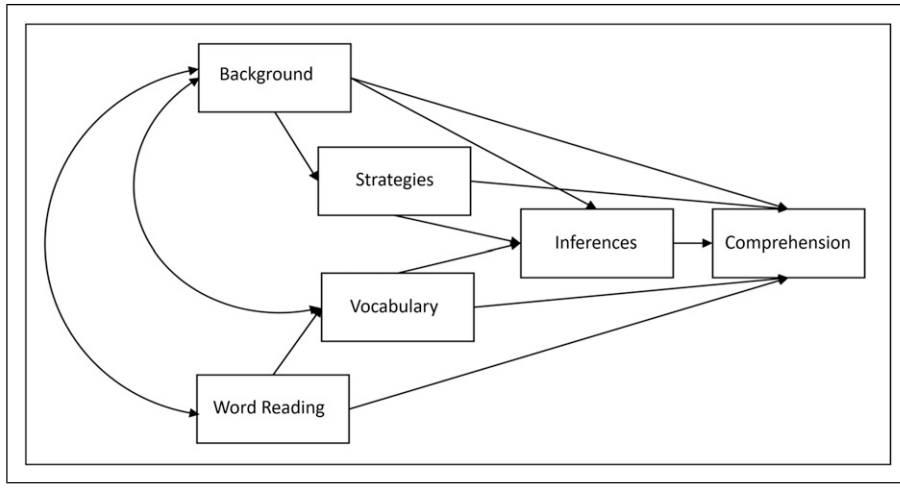


Figure 6. DIME model (direct and inferential mediation model; Cromley & Azevedo, 2007).

was male (N_{male} = 2,075; $N_{missing}$ = 18), and 514 children had a first-generation migration background. This means that the students were not born in Germany. In total, 665 test takers had a second-generation migration background, which means that they were born in Germany, but both of their parents were born somewhere else.

Measures. Reading comprehension was assessed with the Bayerische Lesetest (BYLET, Kraus, 2022). It consists of four text passages with four questions each and four final additional questions, representing the reading levels II-V of the PIRLS-reading ability model. Level I (decoding) was seen as a prerequisite for the test and not operationalized by items. All items were multiple-choice items. Latent scores for reading ability were obtained from a four-factored compensatory MIRT-Model, with one factor for every reading level. Item parameters were estimated with the Metropolis-Hastings-Robinson-Monroe (MHRM) algorithm. Latent variables were scaled to zero mean and unit standard deviation. Person parameters were obtained by the maximum a-posteriori (MAP) method (Reckase, 2009). The model showed good fit indices (CFI = .99, $RMSEA$ = .02) and mostly adequate loadings with exception of Level III loadings being very small.

Demographic data about sex and migration background were reported via a questionnaire, filled out once at the beginning of the study. The model structure and parameters are shown in Figure 7.

Analyses. All analyses were performed using R (R Core Team, 2022). Next to descriptive analyses of item difficulties, four RF were fitted, using first- and second-generation migration background as dichotomous target variables and either the random variable and single items or the random variable, single items, and latent variable scores as predictors. They were estimated with the *randomForest* package (Liaw & Wiener, 2002). Neither items nor latent variable scores were rescaled. The number of trees was set to 500 without constraints to the tree depth. Mtry, the size of the set of variables to be chosen from at each split, was set to $\sqrt{p}$ (p denoting the total number of variables in the model). Hyperparameters (like tree depth or mtry) were not tuned and the OOB served as the optimization criterium. Model performance was assessed by classification errors (CE) and by the area under the curve (AUC). AUC indicates the benefit of a model-based decision

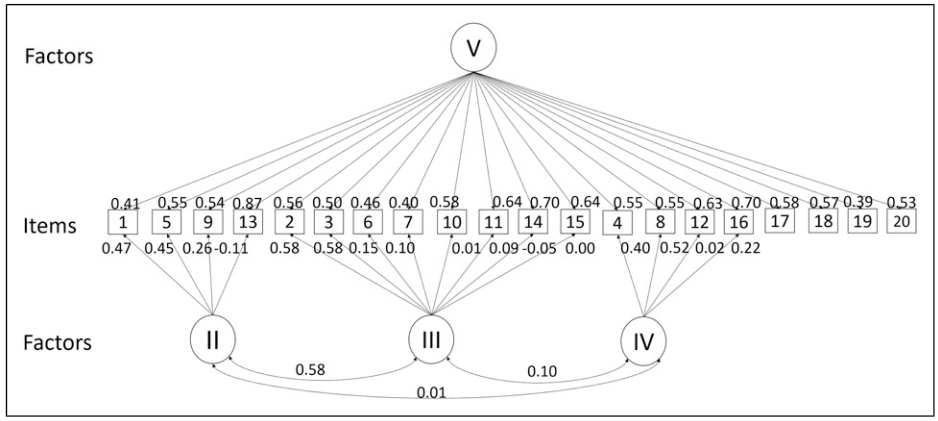


Figure 7. The psychometric model. Note. Roman numbers indicate levels of reading competence. Level I (decoding = reading single words) was not covered explicitly by the test.

criterion compared to deciding to the baseline model. It is derived from the area under the receiver-operator curve, which computes the true-positive-rate and false-positive-rate for all potential thresholds of a decision criterion and visualizes them as a curve. Decision criteria are usually, such as deciding randomly according to the proportion of positive outcomes of the target variable (here proportion of test takers with migration status) or a model-based decision criterion, such as a certain threshold of a prediction score from an ML model. In consequence, the AUC can be used to quantify the benefit of not deciding randomly by the additional area under the curve gained by adopting the model-based decision criterion.

A random seed was set to make analyses numerically replicable. Variable importance plots using the decrease in prediction accuracy after variable permutation were constructed. The increase in classification error was used for the loss function and the process was repeated ten times to assess the variability of the measure. In addition, the inclusion of the random Bernoulli distributed variable indicated a baseline in the variable importance analyses and was used as decision criterion, as introduced in the simulation study. The robustness of the influencing predictors was assessed by descriptively comparing the results for the ML models using different target variables (first- and second-generation migration background). Finally, an MH analysis was performed with respect to the different target variables, using the default setting of the *DifR* package (Magis et al., 2010) and the overall reading competence person parameter as matching criterion.

Application Study Results

Data Preparation and Descriptive Results. The dataset was split into three disjoint subsamples according to the test takers' migration backgrounds. The first subsample represented all children, that were born outside of Germany ($N = 514$) irrespective of their parents' birthland. The second subsample represented all children, who were born in Germany, but whose parents were born outside of Germany ($N = 665$). The last subsample comprised all children who were born in Germany and had parents who were also born in Germany ($N = 2{,}962$). Out of this last subsample, test takers were sampled and then combined with one of the other two subsamples to create two datasets with an even distribution in the migration background target and a stratification by age. The two datasets comprised 1,028 and 1,330 test takers for the

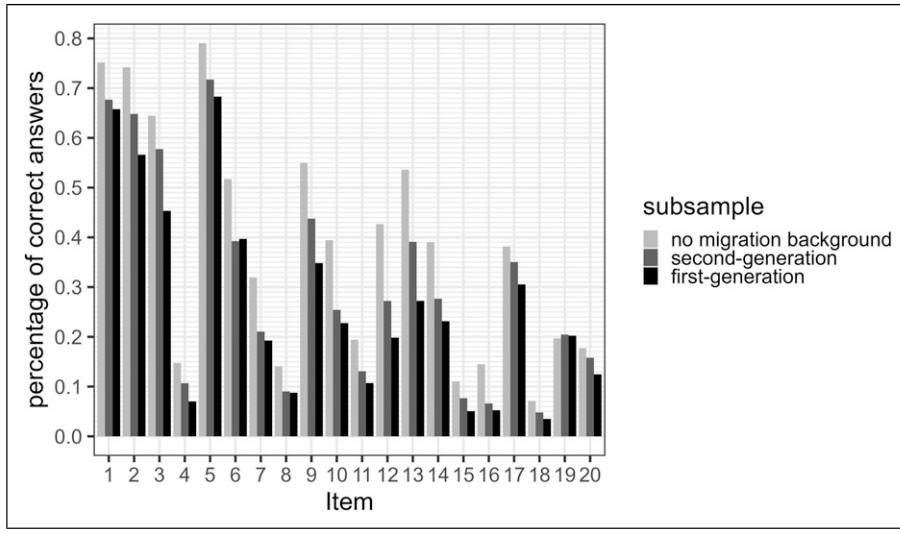


Figure 8. Percentages of correct responses conditional on the migration background.

first-generation and second-generation migration background analyses, respectively. The percentages of correct item responses for the different subsamples are presented in Figure 8.

Random Forests. RF indicated that first- and second-generation migration background could be predicted from the test data with moderate accuracy ($OOB_{\text{first no theta}} = 32.78\%$, $OOB_{\text{first with theta}} = 31.23\%$, $OOB_{\text{second no theta}} = 36.24\%$, $OOB_{\text{second with theta}} = 33.91\%$). The confusion matrices indicated balanced classification errors for both classes ($CE_{\text{mb\_first\_notheta}} = .27$, $CE_{\text{nomb\_first\_notheta}} = .39$; $CE_{\text{mb\_first\_withtheta}} = .32$, $CE_{\text{nomb\_first\_notheta}} = .30$, $CE_{\text{mb\_second\_notheta}} = .37$, $CE_{\text{nomb\_second\_nottheta}} = .35$, $CE_{\text{mb\_second\_withtheta}} = .35$, $CE_{\text{nomb\_second\_withtheta}} = .35$) and the AUCs indicated RF prediction was better than guessing ($AUC_{\text{first\_notheta}} = .72$, $AUC_{\text{first\_withtheta}} = .73$, $AUC_{\text{second\_notheta}} = .69$, $AUC_{\text{second\_withtheta}} = .72$). Furthermore, first-generation migration backgrounds could be predicted with slightly higher accuracy.

It was revealed that the variable importance patterns were very similar with respect to the two target variables and the different predictor sets. When included, the latent score of overall reading competence (Level V) was most influential in the prediction of the first-generation migration background, followed by the random variable and the reading competences associated with the lower levels (Level II, Level III, and Level IV). In the second-generation migration background model, the random variable was the best predictor, followed by some latent scores for the reading competences (Level II, Level IV, and Level V). Single items consistently reached higher variable importance when latent variable scores were not part of the predictors, underpinning the finding of the simulation study of higher sensitivity for unfairness in models without latent variables.

In terms of single unfair items, no item was identified as unfair in the RF including the person parameters for the second-generation migration background and item 12 was identified for the first-generation migration background. But items 6, 10, 12, 13, and 17 were identified in the RF without person parameters for both targets and 2, 3, and 9 were identified to predict the first-generation migration background without person parameters (see Figure 9).

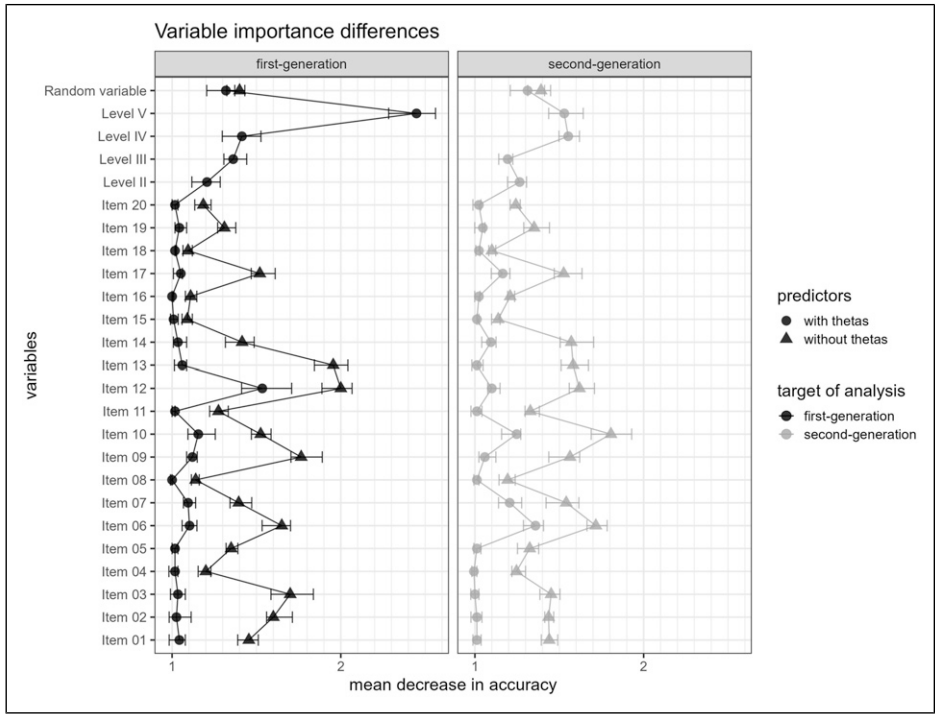


Figure 9. Comparison of variable importance. Note. Error bars represent 95% confidence intervals.

MH Analysis. The MH analysis identified items 12 and 17 as DIF items for test takers with a first-generation migration background and items 10, 12, 16, 17, and 19 as DIF items for test takers with a second-generation migration background. MH results therefore had an overlap with the results from the RF analysis.

Application Study Discussion

Results showed that similar variables were useful in predicting a first- versus a second-generation migration background using single items and psychometric latent variable scores of a reading comprehension test as possible predictors. In accordance with theory, test takers with a migration background descriptively scored lower on all single items. In consequence, the latent ability scores were most predictive when included in the ML models. The two targets (first- and second-generation migration background) were conceptually close. Test takers with a second-generation and a first-generation migration background should differ from test takers without a migration background in similar ways. Language use, language proficiency, and cultural heritage for example are expected to differ conditional on the migration background. Therefore, similar variables should be and were identified as important, when predicting different generations of migration background. The fact that latent ability scores were the most influential predictors was expected. This does however not imply that the test was unfair. On the contrary, as defined in the theory section, a test is unfair, if residual variance—thus variance *not* captured by the latent ability—has a high predictive power in distinguishing between population groups. In our models, this phenomenon would have been represented by single items reaching high variable importance. In fact, interpreting the models, which included person parameters, just one item met the decision criterion for unfairness.

When interpreting the RF models excluding the person parameters a total of five items was associated with the demographic attribute for both targets. Hence, if at all, the fairness of these items must be put in question. As, item 12 had the highest variable importance in all analyses, it was analyzed on a content level for illustrative and educational purposes and was found to contain rarely used vocabulary. These analyses are presented in the Appendix B and need to be viewed with caution because even though they seem reasonable, they are only post-hoc explanations. Results from RF analysis were partly supported by MH analysis, which also identified item 12 among others as an unfair item which speaks to robustness of the finding.

General Discussion

The presented manuscript introduced a novel method. But it can also be seen as a framework that allows to investigate local unfairness or DIF in psychometric tests. We showed by simulation that the method works as reliably as standard procedures in a diverse set of unfair conditions but was slightly less powerful in null and fair impact conditions. In application of the method to a real-world dataset we showed that results are compatible to standard procedures as well as to content based post-hoc explanations. However, the idea of using interpretable ML to investigate unfairness opens a broader perspective on the investigation of test fairness. We specifically chose RF as the ML models and permutation importance as a measure for variable importance. But surely our idea transfers to other ML models, that might also be chosen due to prior knowledge about possible sources of test unfairness. Also, we chose a setting where the number of population groups was two. In other contexts, it can be imaginable, that the population group variable has more than two values (like second language) or is continuous (e.g., age or educational level of the parents). These specifications will lead to different prediction task requirements. Finally, we explored just one baseline criterion, namely the variable importance of a Bernoulli distributed random variable, while random variables from different distributions might lead to fruitful decision criteria as well, or even outperform the current baseline criterion. Therefore, our studies are just a first step into the huge field of possibilities following the general theoretical framework of explaining test unfairness through the influence of unevenly distributed, unintended additional latent constructs on item responses and a resulting complex relationship between item responses and demographic attributes. Before applying the method, we therefore suggest that its specific properties in the application situation be investigated in simulations. Given these first simulation results, we suggest applying the RF including person parameters for longer tests, but RF excluding person parameters for short tests. Also, it must be considered that ML approaches, just as any complex statistical procedure, require large sample sizes to produce reliable results. As for the practical implications of the results, the following section gives some suggestions on how to proceed after identifying unfair items.

Implications for Application

Having identified DIF statistically, the question arises at which level variable importance scores reach a practically relevant level. In our applied analyses most variable important scores for single items were lower than the variable importance score for the random baseline variable. Also, most variable important scores had very low absolute values. Therefore, we conclude that the influence of most items (except for the one discussed above) is probably quite negligible.

Conclusion

In conclusion, our approach shows a novel possibility for identifying single potentially unfair test items. Simulation study results and a model comparison between first-generation and second-generation migration backgrounds as targets showed robustness in prediction and variable importance scores. Regarding the statistical properties of the method, further simulations will be useful. Regarding single potentially unfair items in an application setting, further qualitative analyses need to follow-up in order to unravel the processes leading to the identified unfair measurement properties.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Bayerisches Staatsministerium für Bildung und Kultus, Wissenschaft und Kunst (III.1- BS6200–4b.21994).

ORCID iD


Elisabeth Barbara Kraus https://orcid.org/0000-0001-8007-0321

Supplemental Material

Supplemental material for this article is available online.

References

Attali, Y., & Bar-Hillel, M. (2003). Guess where: The position of correct answers in multiple-choice test items as a psychometric variable. *Journal of Educational Measurement*, *40*(2), 109–128. https://doi.org/10.1111/j.1745-3984.2003.tb01099.x

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

Becker, B., Klein, O., & Biedinger, N. (2013). The development of cognitive, language, and cultural skills from age 3 to 6: A comparison between children of Turkish origin and children of native-born German parents and the role of immigrant parents' acculturation to the receiving society. *American Educational Research Journal*, *50*(3), 616–649. https://doi.org/10.3102/0002831213480825

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

Belzak, W. C. (2022). The multidimensionality of measurement bias in high-stakes testing: Using machine learning to evaluate complex sources of differential item functioning. *Educational Measurement: Issues and Practice*, *42*(1), 24–33. https://doi.org/10.1111/emip.12486

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

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140. https://doi.org/10.1007/BF00058655

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Bryant, F. B. (2000). Assessing the validity of measurement. In L. G. Grimm, & P. R. Yarnold (Eds.), *Reading and understanding MORE multivariate statistics* (pp. 99–146). American Psychological Association.

Carlson, K. D., & Herdman, A. O. (2012). Understanding the impact of convergent validity on research results. *Organizational Research Methods*, *15*(1), 17–32. https://doi.org/10.1177/1094428110392383

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice*, *17*(1), 31–44. https://doi.org/10.1111/j.1745-3992.1998.tb00619.x

Clauser, B. E., Nungester, R. J., Mazor, K., & Ripkey, D. (1996). A comparison of alternative matching strategies for DIF detection in tests that are multidimensional. *Journal of Educational Measurement*, *33*(2), 202–214. https://doi.org/10.1111/j.1745-3984.1996.tb00489.x

Corbett-Davies, S., & Goel, S. (2018). *The measure and mismeasure of fairness: A critical review of fair machine learning*. arXiv preprint arXiv:1808.00023. https://doi.org/10.48550/arXiv.1808.00023

Cromley, J. G., & Azevedo, R. (2007). Testing and refining the direct and inferential mediation model of reading comprehension. *Journal of Educational Psychology*, *99*(2), 311–325. https://doi.org/10.1037/0022-0663.99.2.311

Dorans, N. J. (2004). Using subpopulation invariance to assess test score equity. *Journal of Educational Measurement*, *41*(1), 43–68. https://doi.org/10.1111/j.1745-3984.2004.tb01158.x

Dorans, N. J., & Cook, L. L. (Eds.). (2016). *Fairness in educational assessment and measurement*. Routledge. https://doi.org/10.4324/9781315774527

Dowle, M., & Srinivasan, A. (2021). data.table: Extension of `data.frame`. R package version 1.14.2. https://CRAN.R-project.org/package=data.table

Genz, A., Bretz, F., Miwa, T., Mi, X., LeischScheipl, F. F., & Hothorn, T. (2021). mvtnorm: Multivariate normal and t distributions. R package version 1.1-2. https://CRAN.R-project.org/package=mvtnorm

Gipps, C., & Stobart, G. (2009). Fairness in assessment. In J. J. Wyatt-Smith, & C. M. Cumming (Eds.), *Educational assessment in the 21st century*. Springer. https://doi.org/10.1007/978-1-4020-9964-9_6

Goretzko, D., & Bühner, M. (2020). One model to rule them all? Using machine learning algorithms to determine the number of factors in exploratory factor analysis. *Psychological Methods*, *25*(6), 776–786. https://doi.org/10.1037/met0000262

Hilbert, S., Coors, S., Kraus, E., Bischl, B., Lindl, A., Frei, M., Wild, J., Krauss, S., Goretzko, D., Stachl, C., & Stachl, C. (2021). Machine learning for the educational sciences. *The Review of Education*, *9*(3), Article e3310. https://doi.org/10.1002/rev3.3310

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel–Haenszel procedure. In H. Wainer, & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Erlbaum. https://doi.org/10.4324/9780203056905-19

Jacobs, A. Z., & Wallach, H. (2021). Measurement and fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 375–385). https://doi.org/10.1145/3442188.3445901

Kassambara, A. (2020). Ggpubr: 'ggplot2' based publication ready plots. R package version 0.4.0. https://CRAN.R-project.org/package=ggpubr

Kraus, E. (2022). Diagnostische Entscheidungen mit dem treatment decision model – ein entscheidungstheoretischer Ansatz auf Basis von Evaluationsstudiendaten. (Dissertation).

Kuppler, M., Kern, C., Bach, R. L., & Kreuter, F. (2022). From fair predictions to just decisions? Conceptualizing algorithmic fairness and distributive justice in the context of data-driven decision-making. *Frontiers in Sociology*, *7*(883999), 1–18. https://doi.org/10.3389/fsoc.2022.883999

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*(3), 18–22.

Lüdemann, E., & Schwerdt, G. (2013). Migration background and educational tracking. *Journal of Population Economics*, *26*(2), 455–481. https://doi.org/10.1007/s00148-012-0414-z

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods*, *42*(3), 847–862. https://doi.org/10.3758/brm.42.3.847

Malik, J. A. (2013). Latent variable. In M. D. Gellman, & J. R. Turner (Eds.), *Encyclopedia of behavioral medicine*. Springer. https://doi.org/10.1007/978-1-4419-1005-9_758

Man, K., Harring, J. R., & Sinharay, S. (2019). Use of data mining methods to detect test fraud. *Journal of Educational Measurement*, *56*(2), 251–279. https://doi.org/10.1111/jedm.12208

Mashek, D., & Hammer, E. Y. (2011) *Empirical research in teaching and learning: Contributions from social psychology* (Vol. 4). John Wiley & Sons. https://doi.org/10.1002/9781444395341

Mazor, K. M., Hambleton, R. K., & Clauser, B. E. (1998). Multidimensional DIF analyses: The effects of matching on unidimensional subtest scores. *Applied Psychological Measurement*, *22*(4), 357–367. https://doi.org/10.1177/014662169802200404

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

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

Molnar, C. (2020). *Interpretable machine learning*. Lulu.com.

Molnar, C., Bischl, B., & Casalicchio, G. (2018). iml: An R package for interpretable machine learning. *Journal of Open Source Software*, *3*(26), 786. https://doi.org/10.21105/joss.00786

Mullis, I. V., Martin, M. O., Kennedy, A. M., Trong, K. L., & Sainsbury, M. P. (2015). *Assessment frameworks*. TIMMS and Pirls International Study Center. https://pirls2021.org

Penfield, R. D., & Camilli, G. (2006). 5 differential item functioning and item bias. *Handbook of Statistics*, *26*, 125–167. https://doi.org/10.1016/s0169-7161(06)26005-x

Probst, P., Boulesteix, A. L., & Bischl, B. (2019). Tunability: Importance of hyperparameters of machine learning algorithms. *Journal of Machine Learning Research*, *20*(1), 1934–1965.

R Core Team. (2022). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Reckase, M. D. (2009). Multidimensional item response theory models. In *Multidimensional item response theory* (pp. 79–112). Springer. https://doi.org/10.1007/978-0-387-89976-3_4

Spearman, C. (1904). General intelligence, objectively determined and measured. *Objectively Determined and Measured. American Journal of Psychology*, *15*(2), 201–293. https://doi.org/10.2307/1412107

Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*(1), 307–311. https://doi.org/10.1186/1471-2105-9-307

Strobl, C., Kopf, J., & Zeileis, A. (2015). Rasch trees: A new method for detecting differential item functioning in the Rasch model. *Psychometrika*, *80*(2), 289–316. https://doi.org/10.1007/s11336-013-9388-3

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, *14*(4), 323–348. https://doi.org/10.1037/a0016973

Tutz, G., & Schauberger, G. (2015). A penalty approach to differential item functioning in Rasch models. *Psychometrika*, *80*(1), 21–43. https://doi.org/10.1007/s11336-013-9377-6

Van der Linden, W. J. (2016). In *Handbook of Item Response Theory: Volume 1: Models*. CRC Press.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450. https://doi.org/10.1007/bf02294627

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer. https://doi.org/10.1007/978-0-387-98141-3

Wickham, H. (2021). tidyr: Tidy messy data. R package version 1.1.3. https://CRAN.R-project.org/package=tidyr