MEASUREMENT NON-INVARIANCE IN MACHINE LEARNING: AN INTERSECTION OF MACHINE LEARNING BIAS AND TEST BIAS

Honoka Suzuki

A thesis submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Arts in the Department of Psychology and Neuroscience (Quantitative Psychology).

Chapel Hill
2023

Approved by:

Oscar Gonzalez

Daniel J. Bauer

Kathleen M. Gates

ABSTRACT

Honoka Suzuki: Measurement Non-invariance in Machine Learning: An Intersection of Machine Learning Bias and Test Bias
(Under the direction of Oscar Gonzalez)

Algorithmic and machine learning bias have stirred concern in society as machine learning continues to channel into sensitive and high-stakes applications, including in healthcare, hiring, and criminal justice. While research surrounding *machine learning bias* may be relatively new, psychometricians have for decades researched a closely paralleled topic of *test bias* in psychological and educational testing. Leveraging the connection between these two fairness domains, this thesis studies the problem of machine learning bias from a measurement perspective, specifically focusing on measurement non-invariance in outcome variables as a source of machine learning bias. A framework is introduced, which conceptualizes machine learning bias in a psychometric sense and allows for tests of measurement invariance in machine learning. Using a Monte Carlo simulation study, the consequences of measurement bias on machine learning bias are demonstrated, as well as the effectiveness of a proposed bias mitigation technique to address these effects of measurement bias, which also follows from the proposed framework. The application of the proposed methods is illustrated with data from a large-scale health survey. Broader implications of the relevance of *fairness in measurement* for *fairness in machine learning* are discussed.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

LIST OF ABBREVIATIONS AND SYMBOLS

CART	Classification and regression trees
FNN	Feedforward neural network
MEPS	Medical Expenditures Panel Survey
MG-CFA	Multiple-groups confirmatory factor analysis
MLMI	Machine learning measurement invariance
RMSE	Root mean squared error
Θ_ϵ	Item residual (co)variance matrix
$\tilde{G}$	Fair dummy
$\hat{Y}$	Machine learning model prediction
$\hat{d}$	Discriminator in fair dummies model fitting procedure
$\hat{f}$	Predictive model in fair dummies model fitting procedure
$\hat{r}$	Predictive model in fair dummies test procedure
$\hat{\eta}$	Factor score estimate
ℓ	Loss function
Λ	Factor loading matrix
Σ	Observed item covariance matrix
Ψ	Factor (co)variance matrix
B	Number of iterations in fair dummies model fitting procedure
D	Discrepancy function
G	Grouping variable
I	Set of observation indices
J	Cost function

K	Number of iterations in fair dummies test procedure
M	Number of indictaors / observed proxy outcome variables
N	Sample size
P	Probability distribution
R	Correlation matrix
T	Test score
X	Predictor in machine learning task
Y	Observed outcome variable in machine learning task
b	Number of gradient descent steps per iteration in fair dummies model fitting procedure
c	Number of simulation conditions flagged in simulation part 1
i	Index for observation
k	Index for iteration in fair dummies test procedure
m	Index for indictaor / observed proxy outcome variable
p	P-value
r	Correlation
η	Target latent variable
θ	Vector of (measurement or machine learning) model parameters
κ	Factor mean vector
λ	Factor loading
μ	Observed item mean vector
τ	Item intercept vector
φ	Trade-off parameter in fair dummies model fitting procedure

CHAPTER 1: INTRODUCTION

In a popular article published in *Science*, Obermeyer and colleagues (2019) describe a commercial machine learning model used in a health system to identify patients with the highest health needs for enrollment in a high-risk care management program. The model was trained on patients' past-year insurance claims data to predict their current year's medical expenses—the model's outcome variable, chosen by the model-creators as a proxy for health needs. The rationale here was that by selecting patients with the highest predicted medical expenses, patients with the highest projected health needs, and therefore patients who will benefit the most from enrolling in the program, were being selected. However, possible evidence of racial bias in the model surfaced when the authors formulated a different proxy for health needs: a patient's number of chronic medical conditions. Given the same level of model-predicted medical expenses, black patients had more chronic conditions than white patients. The authors attributed this disparity to racial differences in access to healthcare. With similar chronic conditions but different access to care, patients would have different expenses (i.e., care received) recorded in the system, from which the machine learning model was trained. As such, by enrolling patients for care based on model-predicted expenses, patients with high care needs but limited care access were disadvantaged, thereby perpetuating health disparities.

This *Science* article is just one of many to spotlight algorithmic fairness and bias in machine learning—topics gaining vast attention as machine learning increasingly informs consequential decision-making in a myriad of disciplines (Mehrabi et al., 2021). Psychology is no exception: machine learning informs hiring pipelines in industrial-organizational psychology

(Liem et al., 2018), intelligent tutoring systems in educational psychology (Koedinger et al., 2015), and mental health screening in clinical psychology (Graham et al., 2019). While algorithmic decision-making has been touted for its seemingly objective nature, grounded in data-driven reasoning rather than subjective human judgment, there is emerging evidence that machine learning models are vulnerable to societal and human biases (O'Neil, 2017). Decisions based on models and predictions tainted by such biases can unintentionally discriminate against and amplify inequities across groups (e.g., race, sex), resulting in machine learning bias.

Responding to these concerns, the computer science community has proposed various definitions of fairness in machine learning, such that machine learning models may be audited for evidence of possible biases (Mehrabi et al., 2021; Mitchell et al., 2021). For example, *sufficiency* refers to the conditional independence of group membership and the observed value of the outcome variable given a certain level of machine learning predictions. Unfortunately, no such single definition can identify all instances of machine learning bias. In fact, the *Science* article mentions that given any level of model-predicted expenses, white and black patients had similar levels of observed expenses (Obermeyer et al., 2019), meaning sufficiency was satisfied in this situation, despite the evidence of racial bias found otherwise.

While in this article, the authors uncovered the bias by examining the relationship of model predictions to the number of chronic conditions, we propose another way to view this problem, by framing it as a problem of *measurement*. We could think of medical expenses, the model's proxy outcome variable, as having been a biased indicator of health needs, a target latent variable which the model-creators ideally intended to predict. In other words, medical expenses, the observed variable, did not measure health needs, the latent construct, equivalently across

racial groups. This framing motivates this thesis to examine machine learning bias from the psychometric lens of *measurement bias* or *non-invariance.*

This perspective draws from the psychometric literature on test bias in psychological and educational testing. In this body of literature, test scores (e.g., SAT scores) are examined for bias to ensure fairness in their use across groups, such as those defined by race, sex, or nationality (Zwick, 2019). These fairness considerations of the testing domain closely parallel those of the machine learning domain—a connection that has been recognized in previous machine learning fairness literature (Barocas et al., 2019; Hutchinson & Mitchell, 2019). However, to date, this machine learning-test fairness connection has primarily focused on the analogy between machine learning fairness and the psychometric practice of *predictive* invariance testing. Predictive invariance testing assesses for predictive bias in test scores by examining whether a test score predicts some future, observed criterion (e.g., college grades) equivalently across groups (Millsap, 2007; Millsap, 2011). *Measurement* invariance testing, on the other hand, checks for measurement bias by assessing whether test items or sub-test scores measure or relate to the intended latent construct (e.g., math achievement) in the same way across groups (Millsap, 2007; Millsap, 2011).

Extending the machine learning-test fairness connection that has previously focused on predictive invariance, we aim to contribute a framework to study machine learning bias from the viewpoint of measurement invariance. Motivated by the *Science* article of Obermeyer et al. (2019), we will specifically focus on measurement non-invariance in outcome variables as a source of machine learning bias. Especially in social and behavioral science applications, what a machine learning task intends to predict may be a latent construct (e.g., job performance, depression), but it is not uncommon for one observed indicator to be chosen as a proxy to serve

as the outcome variable. There is much emphasis on defining predictors (features) in machine learning, as feature selection and feature engineering are routine steps in machine learning workflows (Domingos, 2012). However, as demonstrated in the *Science* article, careful consideration in defining and selecting valid outcome variables can be just as critical, if not more (Goretzko & Israel, 2022). In light of these considerations, this thesis explores the following interconnected questions: *How can we test for measurement invariance of outcome variables in machine learning?; How can machine learning bias be conceptualized in a measurement framework?; What are the effects of measurement bias on machine learning bias, as conceptualized in this framework?*; and *How can these of effects of measurement bias be addressed?*

There are three main components to this thesis. First, we introduce a simple framework to operationalize the true target variable in a machine learning task as a latent variable, to allow for tests of measurement invariance with respect to a grouping variable. This also leads to a definition of machine learning fairness that aligns with the psychometric concept of measurement invariance, providing a useful notion of machine learning bias from a measurement perspective. Second, in part 1 of a two-part Monte Carlo simulation study, we investigate the effects of measurement bias on the fulfilment of this definition under various conditions, to demonstrate the role of measurement bias as a source of machine learning bias, as defined in this framework. We additionally introduce a bias mitigation technique, which follows from the proposed framework, to address measurement bias in machine learning by combining methods from measurement modeling with a bias mitigation technique from the machine learning fairness literature. We evaluate the effectiveness of this proposed technique in part 2 of the simulation

study. Third, we demonstrate the application of these proposed methods on an empirical health dataset.

The rest of Chapter 1 is organized as follows. In Section 1.1, we provide a high-level overview of machine learning, how machine learning differs from statistical techniques traditionally used in psychological research, and why machine learning is becoming increasingly important and relevant in psychological research. In Section 1.2, we provide a more detailed background on fairness research in machine learning, possible sources of machine learning bias, and current research efforts in the computer science community around tackling such biases. In Section 1.3, we review relevant psychometric research on predictive and measurement invariance testing. We then draw a more thorough connection between machine learning fairness and test fairness, as well as past research efforts that have drawn this connection, upon which this thesis builds. Following Chapter 1, we introduce the proposed framework in Chapter 2. We then describe and execute the two-part Monte Carlo simulation study in Chapters 3 and 4, and we demonstrate a relevant use case of the proposed methods with an applied example in Chapter 5. Finally, we conclude with a discussion of the broader contributions of this thesis in Chapter 6.

Section 1.1: Machine Learning and Psychology

Machine learning

Broadly, machine learning refers to the automatic detection or learning of patterns in data (Domingos, 2012; Dwyer et al., 2018; Liem et al., 2018). In supervised machine learning, the goal is to learn patterns among provided examples (i.e., training data) of predictor and outcome variable values, such that given new instances (i.e., test data) of predictor values, the learned mapping of predictors to the outcome variable generates accurate predictions of the outcome variable (Jordan & Mitchell, 2015). As such, a key focus of machine learning is on accurate

prediction of unseen outcomes and generalizability of the model beyond the training data used to fit it (Domingos, 2012; Orrù et al., 2020).

Machine learning differs from statistical methods traditionally used in psychological research (e.g., multiple regression, structural equation modeling) due to this heavier emphasis on prediction rather than explanation (Yarkoni & Westfall, 2017). A core aim of psychological research is to understand and explain human behavior, and it is largely theory-driven. In conducting statistical analyses in psychological research, relationships among predictors and outcome variables are specified *a priori*. Then, statistical models are fit accordingly to test the significance and magnitude of such relationships in the observed data by examining model parameter estimates, which carry interpretable, substantive meaning. In contrast, machine learning methods are suited for largely data-driven research. For example, in classic applications of machine learning, such as spam filtering or fraud detection, researchers may gather a large number of predictors without prior theory and observed instances of an outcome variable. Then, the machine learning model automatically (i.e., without explicit specification from the researcher) detects patterns from this data, which may be highly complex and uninterpretable. Rather than to examine parameter estimates or to test theories that explain the relationships or causal mechanisms underlying these variables, the interest lies in empirically identifying relationships among the variables to fit a model, explainable or not, for future application on data where the outcome variable is not yet observed, such that predictions can be generated for those future instances. In the machine learning context, the fitted model is evaluated with respect to the accuracy of its predictions on unseen (by the model) data. In the statistical analysis context, a model is evaluated in terms of its goodness of fit to the observed data used to fit it (Yarkoni & Westfall, 2017).

Machine learning applications in psychology

Despite these fundamental differences, the psychological sciences have seen a recent surge of machine learning applications (Jacobucci & Grimm, 2020). This may in part be due to the growing availability of high-dimensional datasets and new or complex data types (e.g., text, video) that have made machine learning more applicable and relevant in psychological research (Adjerid & Kelley, 2018). For example, advances in computer technology have made possible the large-scale collection of data generated by users of digital services, such as social media platforms, which are increasingly used in the area of personality assessment (Bleidorn & Hopwood, 2019). Machine learning is a promising tool for the analysis of such digital records and footprints, including profile pictures, status updates, and follower networks, given its ability to empirically learn complex patterns among a large number of variables for which there is limited prior theory (Bleidorn & Hopwood, 2019).

Another likely driver of machine learning applications in psychology is the growing recognition of the utility of prediction and the need for reproducibility in psychological research. For example, Dwyer et al. (2018) discussed the practical utility of machine learning approaches in informing clinical care (e.g., diagnosis, prognosis, treatment decisions), given its focus on predictive accuracy and generalizability at the individual patient-level, over statistical methods traditionally used in clinical psychology and psychiatry that focus on group differences. Further, Yarkoni and Westfall (2017) discussed practices in machine learning that could be useful for improving replicability of findings in psychological research. For example, taking steps to avoid a model from overfitting, or learning the idiosyncrasies of, the training data is routine practice in machine learning (e.g., cross-validation, regularization), given that it focuses on generalizability to the test data. With the replication crisis emerging as a serious concern in psychological

research (Shrout & Rodgers, 2018), such practices from machine learning could serve as useful guides, even if machine learning algorithms themselves may not be as applicable to the types of substantive research questions sought in traditional psychological research.

In a similar vein, machine learning techniques can not only be applied to directly answer substantive research questions, but also instrumentally in enhancing tools for psychological measurement. For example, machine learning models, such as decision trees, can be used to select items from a full psychological questionnaire to create a short-form or tailored test to alleviate respondent burden (Gonzalez, 2021). Techniques common in machine learning, such as lasso regularization and recursive partitioning, have also been used to develop new methods for selecting anchor items and detecting differential item functioning in psychological scales (Belzak & Bauer, 2020; Strobl et al., 2015).

In sum, machine learning presents a powerful set of tools with which researchers can leverage to automatically uncover patterns from large amounts of data, work with new and complex data types, and build models that are generalizable to external data. Despite possible tensions between the theory-guided, explanation-focused nature of psychology and the data-driven, prediction-focused nature of machine learning, recent research occurring at the intersection of these two fields has underscored the potential of machine learning approaches to augment, not replace, traditional approaches to quantitative research in psychology.

Section 1.2: Machine Learning Bias

In considering the far-reaching utility of machine learning, we must also recognize the possible harms and dangers that come with its use. In many applications of machine learning, model predictions are used to inform selection and decision-making about individuals, such as whether to approve a loan applicant, whether to enroll a patient in a health service program, or

whether to select a candidate for a job interview. Especially when predictions are used in sensitive, high-stakes applications that are consequential to individuals' lives, it is crucial that the resulting decisions are not skewed towards and do not discriminate against any groups based on sensitive attributes, such as race, sex, or nationality. These dangers may be heightened by the complex and black-box nature of most machine learning algorithms or the lack of transparency in their use and deployment in practice, making it difficult to assess possible causes or mechanisms of models' discriminatory behavior (O'Neil, 2017).

One example of machine learning bias that has received significant media attention and has stimulated much conversation around algorithmic fairness in recent years is the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), a recidivism risk assessment tool used by some U.S. courts to aid in judges' decision-making, such as pretrial release decisions. An investigation into the COMPAS tool revealed possible concerns of racial bias, such that among those who did not re-offend, black defendants were more likely to be falsely predicted as being high-risk compared to white defendants (Angwin et al., 2016). Given the wide-spread use of machine learning in high-stakes domains such as criminal justice and the consequent need for fairness considerations in designing and deploying machine learning models, the study of algorithmic fairness has become a growing research area of its own in recent years (Mitchell et al., 2021). The topic has also been discussed in various influential outlets, including a recent White House report (Executive Office of the President, 2016).

Sources of machine learning bias

There are many ways in which bias can enter into a machine learning system. Mehrabi et al. (2021) described possible entry points within a feedback loop between data used to train the model, the model, and users' interactions with the output of the model. First, the training data

may be biased in a number of ways. One such possible way, which will be the focus of this thesis, is if the variables in the data exhibit measurement bias or more generally contain differential measurement error across groups. The training data could also suffer from sampling bias, such that it is not representative of or generalizable to the population for which predictions are to be made (Mehrabi et al., 2021; Mitchell et al., 2021; Suresh & Guttag, 2021). Even if data are not collected erroneously or in a non-random manner, the training data can still be biased, in the sense that it reflects societal biases due to historical or existing structural inequities (Suresh & Guttag, 2021). Because the goal of machine learning is to learn patterns in data (Domingos, 2012; Dwyer et al., 2018; Liem et al., 2018), it is problematic to feed biased training data to a model, as it can lead to the reproduction and exacerbation of unfair patterns. Besides the training data, several analytical decisions go into building a machine learning model which may induce bias, such as the functional form of the model, optimization function, and regularizations (Mehrabi et al., 2021; Mitchell et al., 2021). If biased predictions are generated, it is also possible that human interactions with such model outputs lead to the generation of more training data that confirm or reinforce biases, upon which future models may be built. For example, consider a company that is using an algorithmic hiring tool that is biased against female candidates for STEM-related roles and consequently has disproportionately fewer female employees in such roles. If the company then decides to update the model using data from its current employees, the model will be further trained on data which are a reflection of its own biased hiring decisions. Within the cycle of data generation, model building, and output usage, there is a clear danger of a feedback loop that can amplify any biases that enter into this cycle (O'Neil, 2017).

Machine learning fairness definitions

Given these concerns, fairness research in machine learning has produced various technical definitions of fairness, such that machine learning models and their outputs may be evaluated for possible biases. The following are three definitions of fairness that are expressed as joint distributions of the grouping variable G, observed outcome variable Y, and model prediction $\hat{Y}$ (Barocas et al., 2019).

Independence refers to the independence of the prediction and group membership, or $\hat{Y} \perp G$ (Barocas et al., 2019), and is also commonly referred to as *demographic parity*. This definition requires groups to have equal distributions of machine learning predictions, such that groups are equal in their chances of selection or access to resources. While simple, this notion may be limited in utility. In cases where Y is correlated with G, imposing this constraint can hinder the predictive accuracy of the model (Hardt et al., 2016). A related concern is that independence can be satisfied by a model whose performance is excellent in one group but poor in the other, as long as the distribution of predictions is equal across groups. It may be odd to consider a model "fair" when errors (e.g., rejecting a qualified individual or selecting an unqualified individual) are made at a higher rate in one group compared to another.

Separation refers to the conditional independence of the prediction and group membership given the outcome variable, or $\hat{Y} \perp G \mid Y$ (Barocas et al., 2019), and is also referred to as *equalized odds* (Hardt et al., 2016). This definition requires groups to have equal distributions of model predictions across groups within each level of the observed value of the outcome variable. By conditioning on Y, the idea is to require independence only among individuals with the same level of "merit" or "success", as indicated by the outcome variable. However, this assumes that the outcome variable is indeed a valid reflection of merit, which is

not always the case in practice (and is what this thesis aims to address). Separation can also be specifically expressed for cases where $\hat{Y}$ is a binary decision, rather than a continuously measured prediction. For example, *equal opportunity* refers to the equality $P(\hat{Y} = 1 \mid Y = 1, G) = P(\hat{Y} = 1 \mid Y = 1)$, which posits that the probability of selection given "success" on the outcome variable is equal across groups.

Sufficiency refers to the conditional independence of the outcome variable and group membership given the model prediction, or $Y \perp G \mid \hat{Y}$ (Barocas et al., 2019). This definition requires that groups have equal distributions of the outcome variable within each level of model predictions (Barocas et al., 2019). This definition may be considered to reflect the perspective of the model-creators, as individuals are stratified based on their placement on model predictions rather than on the observed outcome variable (Mitchell et al., 2021). Sufficiency can also appear in more specific forms for binary predictions. For example, *predictive parity* refers to the equality $P(Y = 1 \mid \hat{Y} = 1, G) = P(Y = 1 \mid \hat{Y} = 1)$, such that the probability of "success" on the outcome variable is the same across groups among those selected (Chouldechova, 2017).

Given these various fairness definitions, researchers have developed methods to train machine learning models that satisfy these criteria and mitigate possible biases. These methods can be broadly grouped into three categories: those that pre-process the training data to "de-bias" it; those that impose constraints or modifications during the model training process to satisfy some fairness definition; and those that post-process or adjust the model outputs to "de-bias" predictions (Suresh & Guttag, 2021; Barocas et al., 2019). Various toolkits and software have also been developed that allow machine learning researchers and practitioners to measure and mitigate bias issues in their models using these definitions and different mitigation techniques,

including Aequitas (Saleiro et al., 2018), IBM's AI Fairness 360 (Bellamy et al., 2019), and pymetrics' Python package audit-AI.[1]

Other broader notions of fairness have also been proposed besides the above (conditional) independence definitions. *Fairness through unawareness* refers to the notion that a machine learning model is fair if the grouping variable was not explicitly used as a predictor in the model. While simple, this approach can be problematic if other predictors are included that correlate with or contain information that can be used to approximate group membership (Barocas et al., 2019; Kusner et al., 2017). Furthermore, this approach may be inappropriate when the use of group membership is legitimate and provides important information for accurate prediction, such as in a medical application where symptoms for diagnosis of illnesses may differ across groups (Suresh & Guttag, 2021). *Individual fairness* refers to the notion that similar individuals with respect to the outcome variable should be treated similarly across groups (Berk et al., 2017), where equality is considered at the individual-level rather than aggregated at the group-level (Mehrabi et al., 2021). Finally, *counterfactual fairness* examines fairness through a causal framework. In studying causal pathways between predictors, the grouping variable, and the outcome variable, this framework aims to assess how model predictions would differ for an individual in a counterfactual world where their group membership was set to a different value (Kusner et al., 2017; Mitchell et al., 2021).

There is no one universally accepted definition of machine learning fairness, nor any agreement on which definition is preferred in what situation (Mehrabi et al., 2021). In fact, some definitions of fairness have been shown to be inconsistent with one another, such that they cannot be simultaneously satisfied except under stringent and/or unrealistic conditions (Barocas

[1] Available at https://github.com/pymetrics/audit-ai.

et al., 2019; Chouldechova, 2017; Friedler et al., 2016; Mitchell et al., 2021). As the machine learning fairness domain is still nascent, synthesizing these definitions into a more unified framework remains an open challenge (Mehrabi et al., 2021). It is also important to recognize that these technical notions of fairness alone do not address the complexity of the issues surrounding machine learning fairness. However, their utility is non-trivial, as their use can facilitate clear and explicit articulation of the objectives, assumptions, and values in terms of fairness that are implied within model predictions (Hutchinson & Mitchell, 2019; Mitchell et al., 2021).

Section 1.3: Machine Learning Bias and Test Bias

Test bias

While machine learning fairness may be a relatively new domain, the issue of fairness is no stranger to the field of psychometrics. During the late 1960s and into the 1970s, there was a surge in interest in the fairness of educational and psychological tests and their use in selection decisions, such as school admissions and personnel selection. This was largely prompted by the civil rights movement and the women's rights movement, which heightened scrutiny around test scores and their meaning across different groups (Cole & Zieky, 2001). Fairness remains a prominent topic today, and the most recent edition of the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014) discusses its importance as a foundational consideration in testing.

From a purely technical standpoint, fairness in testing has been studied through conceptualizations of test bias based primarily on two forms of invariance: predictive invariance and measurement invariance (Borsboom et al., 2008; Millsap, 2007; Millsap, 2011).

Predictive invariance

Predictive invariance (or a lack of predictive bias) involves some equality across groups in the relationship between test scores and a criterion variable to be predicted by the test scores (Millsap, 2007). In many testing settings, test scores are used for selection because they are deemed to predict a future outcome upon which decision-makers are interested in basing their selection decisions. A classic example is SAT scores. When SAT scores are used in college admissions to select applicants with the highest potential for success in college, it may be of interest to examine how SAT scores predict a relevant criterion, such as college grades or degree completion, across groups.

Such studies of differential prediction in testing most commonly use regression lines of the criterion on the test score to examine for equivalence in regression slopes and intercepts across groups (AERA, APA & NCME, 2014; Zwick, 2019). This regression model of test bias was proposed by Cleary (1968) and is based on the notion that if group differences exist in slopes and/or intercepts, the use of a common regression line will lead to systematic errors in prediction of the criterion for one or both groups. If the regression lines are equivalent, a test can be fairly used to select individuals predicted to perform highest on the criterion. Throughout the early 1970s, many other definitions of test bias were explored, although not necessarily advocated, by different psychometricians. These notions typically involved setting a decision threshold on test scores to determine selection versus rejection, such that certain selection outcomes or proportions were equated across groups. For example, Thorndike (1971) proposed the constant ratio model, where the ratio of the proportion of individuals selected to the proportion of individuals successful (based on a cut score on the criterion) is equal across groups. Einhorn and Bass (1971) proposed that groups be equal in their probability of success at the

decision threshold of the test score. Cole (1973) explored the notion that groups should be equal in their conditional probability of selection given success, and Linn (1973) discussed the equality in the conditional probability of success given selection. Darlington (1971) presented four definitions in terms of correlations, such as a zero correlation between test scores and group membership and a zero partial correlation between test scores and group membership when the criterion is taken into account.

From the various definitions proposed during this period, no universally accepted definition of test bias emerged (Cole & Zieky, 2001; Zwick, 2019), and it was generally held that no single technical definition will be entirely satisfactory (Hunter & Schmidt, 1976; Linn, 1973). The limitations of these definitions have been raised from various viewpoints. Petersen and Novick (1976) pointed out internal inconsistencies *within* some of these definitions, such that following a definition will lead to different thresholds on test scores, depending on whether one is considering the probability of success and selection or its converse (i.e., probability of failure and rejection). For example, the "converse" constant ratio model would refer to the equality across groups in the ratio of the proportion rejected to the proportion unsuccessful, and this can yield different results from the regular constant ratio model (Petersen & Novick, 1976). Inconsistencies *between* definitions have also been raised, such that different definitions cannot be simultaneously satisfied except under stringent and/or unrealistic conditions (Darlington, 1971; Linn, 1973; Petersen & Novick, 1976; Thorndike, 1971). These revelations of contradictory definitions have led to the discussion of test bias notions and their reflections of different value judgments or "ethical positions" (Hunter & Schmidt, 1976; Sawyer et al., 1976). Other limitations regarding these definitions include the skepticism around the practice of setting different thresholds of selection for different groups (Zwick, 2019) and the disconnect between

these technical definitions and the general public's perceptions of fairness (Cole & Zieky, 2001). Another widely recognized limitation (e.g., Linn, 1973; Petersen & Novick, 1976; Thorndike, 1971) is that these definitions operate on the assumption that the criterion variable itself is an unbiased, reliable measure, which may be dubious in practice.

Predictive invariance and machine learning fairness

Much of the fairness research in machine learning closely resembles this line of research on predictive invariance in the testing domain (Hutchison & Mitchell, 2019). Hutchison and Mitchell (2019) detailed this correspondence between test bias and machine learning bias by drawing an analogy between tests and machine learning models and their respective outputs. Specifically, a test score can be considered analogous to a machine learning prediction, and a criterion variable in testing can be considered analogous to the observed outcome variable (i.e., the "ground truth") in machine learning.[2] From there, certain equivalencies between the technical definitions of fairness explored in the areas of testing and machine learning become clear. For example, Darlington's (1971) fourth definition is equivalent to the independence criterion in machine learning (under a bivariate normal distribution), Cole's (1971) definition is equivalent to the equal opportunity criterion, and Linn's (1973) definition is equivalent to the predictive parity criterion. In addition to equivalent definitions, both the testing and machine learning literatures have discussed impossibilities in simultaneously satisfying multiple competing definitions, as well as the definitions' reflections of different value judgments (Hutchinson & Mitchell, 2019).

[2] An interesting distinction between machine learning models and tests in this analogy is that in machine learning, the prediction is made precisely to reproduce the ground truth outcome variable, and therefore the outcome variable is "internal" to the model. In testing, on the other hand, a criterion variable is external to the test, and while test scores should correlate with the criterion, test scores are not necessarily made to reproduce the criterion, or at least that is not how a test is constructed.

These stark parallels between machine learning bias and test bias reveal much potential for psychometric perspectives to contribute towards the machine learning fairness literature, as the research area continues to expand. In fact, just as a connection can be drawn between machine learning fairness and predictive invariance, there is an important connection to be made between machine learning fairness and measurement invariance, to which we turn next.

Measurement invariance

The second form of invariance commonly studied in the context of test bias is measurement invariance (or a lack of measurement bias). Whereas predictive invariance focuses on the equality in the relationship between the test score and an observed criterion across groups, measurement invariance focuses on the equality in the relationship between the test score and the latent construct which the test score intends to measure (Millsap, 2007). Formally, measurement invariance can be expressed as

$$f(T \mid \eta, G) = f(T \mid \eta) \tag{1}$$

where $f(\cdot)$ denotes a probability density function, T is a test score,[3] η is the latent construct, and G is a grouping variable. When Equation 1 holds, test scores have the same meaning across groups, or the latent construct is measured by the test score in the same way across groups. This is because conditional on the latent construct, individuals are expected to receive the same test score across groups, or $T \perp G \mid \eta$.

While predictive invariance was the central focus of much of the research on test bias in the 1970s, measurement invariance has emerged as a "very serious competitor" to predictive invariance in contemporary psychometrics (Borsboom et al., 2008, p. 76). While predictive invariance may be more easily investigated given that it involves observed variables only, it has

[3] In some contexts, *T* may be a test item response rather than a test score, depending on whether invariance is considered at the item-level or at the scale- or test-level.

some limitations. First, the choice of the criterion variable, as well as the timing at which this criterion will be measured, are arbitrary choices, yet are influential determinants of the conclusions about the presence or absence of predictive bias. Further, different conclusions can be drawn depending on whether the prediction of the criterion from test scores is considered, or the prediction of test scores from the criterion is considered. Such ambiguities are avoided under measurement invariance testing because measurement invariance concerns the relationship of the test score to the latent construct at the time of testing, and the measurement model implies a clear causal direction of effects, where the latent construct gives rise to test scores (Borsboom et al., 2008). Because predictive invariance and measurement invariance have been shown to contradict one another (Millsap, 2007; Millsap 2011), measurement invariance may be preferred, if one had to choose between the two (Borsboom et al., 2008).

One way in which measurement invariance is investigated for continuous variables is through a multiple-groups confirmatory factor analysis (MG-CFA; Millsap & Olivera-Aguilar, 2012). A CFA model is a measurement model that connects observed variables, or items, to their underlying latent constructs, or factors (Brown & Moore, 2012). These item-factor relationships are inferred from the covariance structure of the items, such that a common factor(s) is theorized to explain the intercorrelations among the items. The parameters $\boldsymbol{\theta}$ of a CFA model include a factor loading matrix $\boldsymbol{\Lambda}$ and an item intercept vector $\boldsymbol{\tau}$, which define the regression equations predicting each item from the factors, a factor mean vector $\boldsymbol{\kappa}$, factor (co)variance matrix $\boldsymbol{\Psi}$, and item residual (co)variance matrix $\boldsymbol{\Theta}_{\boldsymbol{\epsilon}}$, whose diagonal elements are the portion of item variances unexplained by the set of factors (Brown & Moore, 2012). These parameters are estimated with the maximum likelihood estimator, with the goal of reproducing the observed mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ of the items as closely as possible, where

$$\boldsymbol{\mu}(\boldsymbol{\theta}) = \boldsymbol{\tau} + \boldsymbol{\Lambda\kappa} \quad (2)$$

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Lambda\Psi\Lambda}' + \boldsymbol{\Theta}_{\boldsymbol{\epsilon}} \quad (3)$$

with $\boldsymbol{\mu}(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ denoting the CFA model-implied mean vector and covariance matrix, respectively. The goodness of fit of CFA models are evaluated on how well $\boldsymbol{\mu}(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ match $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ respectively, using statistical tests and fit indices such as the model chi-square test, Tucker-Lewis Index (TLI), comparative fit index (CFI), and root mean square error of approximation (RMSEA; Brown & Moore, 2012).

When discrete groupings are present among the sample observations (e.g., males vs. females), a MG-CFA can be applied to study differences in measurement properties across those groups, serving as a foundation for measurement invariance testing. Specifically, a series of nested MG-CFA models are fit sequentially, each time fitting a more restrictive model according to the level of invariance being tested (Meredith, 1993; Millsap & Olivera-Aguilar, 2012). A typical workflow starts with testing for *configural invariance*, where the general factor structure (i.e., number of factors, pattern of free and fixed factor loadings) is restricted to be equal across groups. If the configural model fits adequately well, *weak factorial invariance* is tested next. Here, the factor loading matrix is restricted to be equal across groups. If these additional constraints do not worsen model fit significantly more than the configural model according to a likelihood ratio test for nested models, weak factorial invariance is said to hold, and *strong factorial invariance* is tested next. Strong factorial invariance restricts the factor loading matrix and item intercept vector to be equal across groups. If this model does not fit significantly worse than the weak factorial model, then strong factorial invariance is said to hold. Finally, *strict factorial invariance* restricts factor loading matrices, item intercept vectors, and residual

(co)variance matrices to be equal across groups. However, this is not typically tested in practice, and invariance testing generally stops at strong factorial invariance (Putnick & Bornstein, 2016).

Measurement invariance and machine learning fairness

The connection between machine learning fairness and measurement invariance may not be as pronounced as the connection between machine learning fairness and predictive invariance, where there was a clear analogy to be drawn.[4] However, discussions of measurement bias (and measurement considerations more generally) and machine learning bias have intersected in several avenues of research, highlighting a promising area for further advancement to machine learning fairness research. For example, Johnson et al. (2022) and Tay et al. (2022) recognized the use of predictors and/or outcome variables that exhibit measurement bias as a possible source of machine learning bias in automated scoring in educational testing and in personality assessment, respectively. Goretzko and Israel (2021), Landers and Behrend (2022), and Raghavan et al. (2020) discussed how choosing and defining an outcome variable that is an objective, reliable, and valid measure can be a particularly challenging task in algorithmic hiring applications, such as resume screening, video interview scoring, and pre-employment assessment. For example, the outcome of job performance could be based on an organization's definition of a perfect employee using multiple indicators, a single objective metric like sales revenue, or simply the outcome of a traditional (i.e., not using machine learning) selection decision, and there are advantages and weaknesses to each one (Goretzko & Israel, 2021).

[4] Tay et al. (2022) proposed a framework called "machine learning measurement bias" for conceptualizing bias in machine learning applications for psychological assessment. They made an analogy between an observed test score (from a psychological assessment) and a machine learning prediction and between a latent variable and the ground truth outcome variable, to motivate a machine learning analogue of measurement bias that examines for differential relationships between a machine learning prediction and the outcome variable. However, a key feature of the proposed framework in this thesis is the distinguishing of the latent construct from the observed outcome variable, to recognize that the outcome variable is not necessarily a "truth" and that it should not be analogized to the latent construct. This makes our proposed framework fundamentally distinct from that of Tay et al. (2022).

Relatedly, Jacobs and Wallach (2021) and Tay et al. (2020) discussed the limitations of a "ground truth" outcome variable when imperfect proxy measures are used as outcome variables in machine learning tasks to predict latent constructs, such as personality, recidivism risk, and teacher effectiveness. In such cases, outcome variables may not necessarily represent a "truth". Mismatches between the intended latent construct and observed proxies can jeopardize fairness in computational systems (Jacobs & Wallach, 2021).

This thesis builds upon these studies that have acknowledged measurement as a key consideration for machine learning fairness, as we propose a framework to assess for measurement invariance of outcome variables in machine learning and psychometrically conceptualize machine learning bias. While the importance of measurement for machine learning fairness has been recognized in previous literatures as discussed above, to our knowledge, no concrete methods have been proposed to practically address these considerations. Further, there is limited systematic evidence around the nature of the relationship of measurement bias to machine learning bias that empirically demonstrates the importance of measurement invariance in machine learning. As such, the aim of this thesis is to fill these gaps, such that researchers can be better aware of the effects of measurement bias and are better practically equipped to study and address machine learning bias from a measurement perspective.

Finally, in this thesis, we almost exclusively focus on technical notions of machine learning bias and test bias, and to that end, we use "fairness" synonymously with "a lack of bias". However, it should be noted that considerations of fairness in both domains of machine learning and testing extend beyond these narrow, statistical properties of machine learning models and tests, respectively. For example, fairness can be thought of as a construct informed by social and cultural values, personal experiences, and philosophy (Zwick, 2019), which does

not simply equate to an absence of bias in the technical sense. While it is beyond the scope of this thesis, an important extension of this work may be to study the machine learning-test fairness connection from a nontechnical standpoint.

CHAPTER 2: PROPOSED FRAMEWORK

Suppose that there is a machine learning task, where fairness of predictions with respect to a binary grouping variable G is of interest from a measurement perspective. In this framework, we posit the existence of a target latent variable η, which is the construct that the machine learning task ideally aims to predict, but is not directly observed. Reflective indicators of η are M candidates to serve as the model's proxy outcome variable ($Y_m, m = 1, 2, \dots, M$), which are observed variables. The goal is to gather multiple, candidate proxy outcome variables which are observed indicators of η, such that a unidimensional factor model may be built to infer η and examine the measurement properties of those observed outcome variables. Specifically, we may test for measurement invariance using a MG-CFA with G, as is done traditionally in a psychometric setting, and flag any outcome variables for a violation of measurement invariance. In doing so, it is assumed that all usual assumptions and best practices of a MG-CFA with maximum likelihood estimation have been considered, including having adequate sample sizes from both groups, continuity and normality in the outcome variables (or with appropriate adjustments for nonnormality), and model identification. Furthermore, before a MG-CFA is applied, adequate fit of a one-factor CFA model within each group separately should be confirmed using fit indices and the model chi-square test.

While the outcome variable is often referred to as the "ground truth" in machine learning tasks, this framework conceptually separates an observed outcome variable from its true underlying latent construct. This helps to avoid an unquestioning perception of the observed

outcome variable as necessarily being a “truth”, which is particularly important in social and behavioral science applications of machine learning where it is common for the true target of prediction to be latent.

Further, by introducing η, a useful notion of machine learning fairness arises from this framework that aligns with the psychometric concept of measurement invariance: the conditional independence of the machine learning prediction and group membership given the target latent variable, or $\hat{Y} \perp G \mid \eta$. We will refer to this as *machine learning measurement invariance* (MLMI) hereafter. This definition requires that given a level of the target latent variable, machine learning models produce equal distributions of predictions across groups. This extends the analogy between tests and machine learning models of Hutchinson and Mitchell (2019), as it simply replaces the test score (T) with a machine learning prediction ($\hat{Y}$) in the definition of measurement invariance in Equation 1. Especially in machine learning applications where a proxy outcome variable is used, a definition that conditions on η rather than Y, as in separation ($\hat{Y} \perp G \mid Y$), provides a more rigorous and useful check of machine learning fairness from a measurement perspective. Using this conceptualization of machine learning fairness in the psychometric sense, we investigate the role of measurement bias as a source of machine learning bias in simulation part 1, demonstrating the importance of testing for measurement invariance using the proposed framework.

CHAPTER 3: SIMULATION STUDY PART 1

Section 3.1: Methods

In the first part of the two-part simulation study, we investigated the consequences of measurement bias for machine learning bias, as defined by a violation of MLMI. In particular, we examined whether training a machine learning model on a non-invariant outcome variable results in predictions that violate MLMI, even when separation is satisfied. In doing so, we aimed to elucidate the simulation conditions under which non-invariance in the outcome variable may manifest as biased predictions, even when attempts to train a "fair" model with observed variables have otherwise been made. Because η is unobserved and one cannot test for MLMI in practice, examining this research question using a simulation (where η is simulated and known) demonstrates when and why there is added benefit to assessing for non-invariance in outcome variables in machine learning using the proposed framework, even beyond working with existing definitions of machine learning fairness. We focus here on separation out of the various fairness definitions, as it serves as a natural counterpart to MLMI involving observed variables only. See Table 1 for a summary of these fairness-related concepts. While not the central focus of the simulation, we additionally studied effects of measurement bias on the predictive performance of machine learning models.

Table 1

Summary of the three fairness-related concepts explored in the simulation study

Concept	Type	Definition
Measurement invariant outcome variable	Measurement property of training data	$Y \perp G \mid \eta$

Machine learning measurement invariance (MLMI)	Machine learning fairness definition (proposed)	$\hat{Y} \perp G \mid \eta$
Separation	Machine learning fairness definition (existing)	$\hat{Y} \perp G \mid Y$

Data generation

For each replication of a simulation condition, we generated a dataset using the below steps:

1. Simulate a binary grouping variable $G \in \{0,1\}$ as

$$G_i \sim \text{Bernoulli}(0.5) \tag{4}$$

for observation $i = 1, \ldots, N = 5{,}000$.

2. Simulate a $N \times (q = 8)$ matrix of predictors $\boldsymbol{X} = [X_1, \ldots, X_8]$ from a multivariate normal distribution with a specified mean vector $\boldsymbol{\mu}_G$ and correlation matrix $\boldsymbol{R}$, where

$$\boldsymbol{\mu}_{G=0} = [0, 0, 0, 0, 0, 0, 0, 0] \tag{5}$$

$$\boldsymbol{\mu}_{G=1} = [0, 0.1, 0.2, 0.4, 0.5, 0, 0.3, 0.15] \tag{6}$$

$$\boldsymbol{R} = \begin{bmatrix} 1.0 & & & & & & & \\ 0.22 & 1.0 & & & & & & \\ 0.12 & 0.12 & 1.0 & & & & & \\ 0.19 & 0.12 & 0.21 & 1.0 & & & & \\ 0.06 & 0.20 & 0.11 & 0.11 & 1.0 & & & \\ 0.23 & 0.13 & 0.28 & 0.23 & 0.08 & 1.0 & & \\ 0.02 & 0.19 & 0.29 & 0.25 & 0.09 & 0.06 & 1.0 & \\ 0.22 & 0.02 & 0.13 & 0.07 & 0.24 & 0.08 & 0.30 & 1.0 \end{bmatrix} \tag{7}$$

The elements of $\boldsymbol{\mu}_{G=1}$ were generated between 0 to 0.5 to create group mean differences in predictors of up to a medium effect size (Cohen, 1988). The off-diagonal elements of $\boldsymbol{R}$ were generated between 0 and 0.3 to create correlations among predictors of up to a medium effect size (Cohen, 1988).

3. Simulate the target latent variable η as a function of a subset of the predictors in $\boldsymbol{X}$, such that

$$\eta_i = 6X_{1i}X_{2i} + 5.5G_iX_{2i} + 8X_{3i} - 4{X_{4i}}^2 + 6X_{7i} + 1.5e^{X_{8i}} + \varepsilon_i \quad (8)$$

for observation $i = 1, \ldots, N$, where $\varepsilon_i \sim N(0, 2^2)$. Then, standardize η. The machine learning model will use predictors X_1 through X_6, such that X_5, X_6 are noise predictors, and X_7, X_8 are signal predictors that are unavailable to the analyst.

4. Simulate four observed indicators of η to serve as the candidate proxy outcome variables for the machine learning task.

Three of the observed outcome variables in Step 4 were simulated to be measurement invariant across groups, and the fourth outcome variable was simulated to be measurement non-invariant across groups. Specifically, a vector of observed outcome variables $\boldsymbol{Y}_i = [Y_{1i}, Y_{2i}, Y_{3i}, Y_{4i}]$ (indicators of η) was generated for observation *i* in group G_i as

$$\boldsymbol{Y}_i = \boldsymbol{\tau}_{G_i} + \boldsymbol{\Lambda}_{G_i}\eta_i + \boldsymbol{u}_{iG_i} \quad (9)$$

where $\boldsymbol{\tau}_{G=1} = [0, 0, 0, 0]$, $\boldsymbol{\Lambda}_{G=1} = [\lambda, \lambda, \lambda, \lambda + \lambda^*]$, $\boldsymbol{\tau}_{G=0} = [0, 0, 0, \tau^*]$, $\boldsymbol{\Lambda}_{G=0} = [\lambda, \lambda, \lambda, \lambda - \lambda^*]$, and $\boldsymbol{u}_{iG} \sim N(\boldsymbol{0}, \boldsymbol{1} - \boldsymbol{\Lambda}_G^2)$. This configuration implies that $\boldsymbol{\tau}_{G=0} = \boldsymbol{\tau}_{G=1}$ and $\boldsymbol{\Lambda}_{\mathrm{G=0}} = \boldsymbol{\Lambda}_{\mathrm{G=1}}$ for the first three outcome variables (Y_1, Y_2, Y_3), whereas the fourth outcome variable (Y_4) had loadings and/or intercepts that differ across groups according to the simulation condition, with magnitudes of group differences in parameters also varying by condition. Data were generated using R (R Core Team, 2023).

Simulation factors

There were four manipulated simulation factors: (1) inclusion or exclusion of G as a predictor when training the machine learning models (i.e., *fairness through unawareness*; see Section 1.2; two levels); (2) value of the factor loadings λ (three levels); (3) type of non-

invariance in Y_4 (two levels); and (4) magnitude of the non-invariance in Y_4 (three levels). These four simulation factors and their respective levels resulted in a total of 36 simulation conditions for simulation part 1. We ran 500 replications within each condition.

When G is included as a predictor, the predictor matrix $\boldsymbol{X}$ refers to $[X_1, \ldots, X_6, G]$. The values of λ investigated were 0.75, 0.65, and 0.55. Factor loadings in this range are expected in practice, including in several common scales in areas of health and psychology such as the eight-item Patient Health Questionnaire depression scale (PHQ-8), seven-item General Anxiety Disorder scale (GAD-7), and Quality of Life Scale (QOLS; Burckhardt et al., 2003; Spitzer et al., 2006). The types of non-invariance in Y_4 were non-invariant intercepts only ($\lambda^* = 0;\ \tau^* \neq 0$) and non-invariant loadings and intercepts ($\lambda^* \neq 0;\ \tau^* \neq 0$). A level for non-invariant loadings only ($\lambda^* \neq 0;\ \tau^* = 0$) was not considered, given that non-invariant loadings are typically accompanied by non-invariant intercepts (Millsap & Olivera-Aguilar, 2012). The levels of the magnitude of the non-invariance in Y_4 were small, medium, and large, each of which were defined in the following manner. For loadings, we used $\lambda^* = 0.05, 0.10, 0.15$ for a small, medium, and large difference, respectively. This creates a group difference in the loading of Y_4 of 0.10 (small), 0.20 (medium), and 0.30 (large), while maintaining an average loading of λ when pooling across groups. This ensures that despite the measurement non-invariance, Y_4 has the same average reliability as an indicator of η as the rest of the outcome variables, such that any comparisons between Y_4 and the other invariant outcome variables are isolated to the measurement bias. For intercepts, we used $\tau^* = -0.2, -0.5, -0.8$ for a small, medium, and large difference, respectively, to correspond to Cohen's effect sizes for standardized mean differences (Cohen, 1988). Because Y_4 has a variance of 1 and its scale is "meaningful and familiar", the intercept differences may be directly interpreted in this way (Millsap & Olivera-Aguilar, 2012, p.

384). The levels of these simulation factors and their resulting measurement parameters are summarized in Table 2.

Table 2

Summary of simulation factors and resulting measurement parameters

Type of non-invariance in Y_4	λ	Magnitude of non-invariance	$\boldsymbol{\Lambda}_{G=0}$	$\boldsymbol{\Lambda}_{G=1}$	$\boldsymbol{\tau}_{G=0}$
Non-invariant intercepts only	0.75	Small	$[0.75, 0.75, 0.75, 0.75]$	$[0.75, 0.75, 0.75, 0.75]$	$[0, 0, 0, -0.2]$
		Medium			$[0, 0, 0, -0.5]$
		Large			$[0, 0, 0, -0.8]$
	0.65	Small	$[0.65, 0.65, 0.65, 0.65]$	$[0.65, 0.65, 0.65, 0.65]$	$[0, 0, 0, -0.2]$
		Medium			$[0, 0, 0, -0.5]$
		Large			$[0, 0, 0, -0.8]$
	0.55	Small	$[0.55, 0.55, 0.55, 0.55]$	$[0.55, 0.55, 0.55, 0.55]$	$[0, 0, 0, -0.2]$
		Medium			$[0, 0, 0, -0.5]$
		Large			$[0, 0, 0, -0.8]$
Non-invariant loadings and intercepts	0.75	Small	$[0.75, 0.75, 0.75, 0.7]$	$[0.75, 0.75, 0.75, 0.8]$	$[0, 0, 0, -0.2]$
		Medium	$[0.75, 0.75, 0.75, 0.65]$	$[0.75, 0.75, 0.75, 0.85]$	$[0, 0, 0, -0.5]$
		Large	$[0.75, 0.75, 0.75, 0.6]$	$[0.75, 0.75, 0.75, 0.9]$	$[0, 0, 0, -0.8]$
	0.65	Small	$[0.65, 0.65, 0.65, 0.6]$	$[0.65, 0.65, 0.65, 0.7]$	$[0, 0, 0, -0.2]$
		Medium	$[0.65, 0.65, 0.65, 0.55]$	$[0.65, 0.65, 0.65, 0.75]$	$[0, 0, 0, -0.5]$
		Large	$[0.65, 0.65, 0.65, 0.5]$	$[0.65, 0.65, 0.65, 0.8]$	$[0, 0, 0, -0.8]$
	0.55	Small	$[0.55, 0.55, 0.55, 0.5]$	$[0.55, 0.55, 0.55, 0.6]$	$[0, 0, 0, -0.2]$
		Medium	$[0.55, 0.55, 0.55, 0.45]$	$[0.55, 0.55, 0.55, 0.65]$	$[0, 0, 0, -0.5]$
		Large	$[0.55, 0.55, 0.55, 0.4]$	$[0.55, 0.55, 0.55, 0.7]$	$[0, 0, 0, -0.8]$

Model building

With the generated dataset in each replication, we randomly split the data into a training (60%; $N_{train} = 3{,}000$) and test set (40%; $N_{test} = 2{,}000$). We trained machine learning models on the training set $(Y_{mi}, \boldsymbol{X}_i), i \in I_{train}$ and applied the trained models on the test set to obtain predictions $\hat{Y}_{mi}$, $i \in I_{test}$. The machine learning models were feedforward neural networks (FNN), for reasons discussed in a subsequent section. Briefly, a FNN is a type of machine learning model that passes data through a series of function compositions to generate predictions (Goodfellow et al., 2016; Urban & Gates, 2021). This allows the model to learn complex,

flexible, nonlinear, multi-step mappings between the predictors and the outcome variable that often result in high predictive accuracy. The function composition works such that a function is first applied to the predictors, or the input layer, generating a vector of intermediate outputs, called a hidden layer, which in turn becomes the inputs of a subsequent function, generating another hidden layer. This chained procedure is repeated until the last hidden layer is reached, at which point a final function is applied to generate a prediction, or the output layer. Each element, or node, of a layer is generated as a weighted sum of its inputs (i.e., the preceding layer's nodes) plus a bias term, which is then passed through an activation function. For instance, the a^{th} node of a hidden layer may be generated as

$$h_a = f(b_{a,0}^g + \sum_{j=1}^{p} w_{a,j}^g x_j) \tag{10}$$

where $f(\cdot)$ is an activation function, b_0 is a bias term, w_j and x_j are the weight and value, respectively, of the j^{th} node of the preceding layer (of size p nodes), and the subscript a and superscript g index the node and layer, respectively, to denote that there are unique weights and bias terms for each node per layer (Urban & Gates, 2021). A common choice for the activation is the rectified linear unit (ReLU), which refers to the function $f(z) = \max\{0, z\}$.

A forward pass of the data from the input layer to the output layer generates a prediction, whose fit is then evaluated by some cost function $J(\boldsymbol{\theta})$. The parameters of a FNN, $\boldsymbol{\theta}$ (i.e., the weights and bias terms that generate each layer's nodes, or $b_{a,0}^g$ and $w_{a,j}^g$ in Equation 10), are updated using optimization algorithms such as gradient descent (Goodfellow et al., 2016). The gradient, or partial derivative, of the cost function with respective to each parameter, is calculated using a technique called backpropagation, in a backward pass of the cost function from the output layer to the input layer. The negative of the calculated gradient informs the direction in which the cost function decreases the fastest, and $\boldsymbol{\theta}$ is updated accordingly in that

direction, with a learning rate controlling the size of the update, or step (Goodfellow et al., 2016). A FNN goes through many such steps during training.

The simulation design involved building machine learning models that satisfy separation and testing whether their predictions satisfy MLMI. To do so, we used the procedures described in Romano et al. (2020). The testing procedure was originally proposed as a hypothesis test for separation, called the fair dummies test, but it can easily extend to checking for other conditional independence relations (Romano et al., 2020), such as MLMI. Below, we describe the details of these procedures.

Methods of Romano et al. (2020): fair dummies test

The fair dummies test (Romano et al., 2020) works by sampling a "dummy" copy of the grouping variable G, denoted $\tilde{G}$, from the distribution $P_{G|Y}$. Because $\tilde{G}$ is generated without having seen $\hat{Y}$, it by construction satisfies separation, or $\hat{Y} \perp \tilde{G} \mid Y$,[5] hence the naming of "fair dummy". The test leverages this property to test the null hypothesis that $(\hat{Y}, G, Y)$ and $(\hat{Y}, \tilde{G}, Y)$ are equal in distribution and therefore that $\hat{Y} \perp G \mid Y$ holds. The test involves the following steps to compute a p-value:

1. Split the test set observations $(\hat{Y}_i, G_i, Y_i), i \in I_{test}$, into two disjoint subsets, I_1 (50%) and I_2 (50%).
2. Train a predictive model $\hat{r}(\cdot)$ that aims to predict $\hat{Y}$ given (G, Y) using the data subset $(\hat{Y}_i, G_i, Y_i)$, $i \in I_1$.[6] Note that if the null is true, $\hat{r}(\cdot)$ should have poor, chance-level performance.

[5] Note that here, Y refers to a single observed outcome variable, rather than a vector of multiple observed outcome variables.

[6] The model $\hat{r}(\cdot)$ is a random forest made up of 10 trees.

3. Apply the trained model from Step 2 on the data subset $\left(\hat{Y}_i, \mathrm{G}_i, Y_i\right)$, $i \in I_2$ to compute the mean squared error as the observed test statistic: $t^* = \frac{1}{N_{I_2}} \sum_{i \in I_2} \left(\hat{Y}_i - \hat{r}(G_i, Y_i)\right)^2$.

4. Repeat the following procedure *K* times to obtain *K* test statistics under the null to compute an empirical p-value:

    a. Sample a fair dummy from the distribution $P_{G|Y}$ for each observation in I_2: $\tilde{G}_i \sim P_{G|Y}(G_i \mid Y_i)$, $i \in I_2$. The conditional distribution $P_{G|Y}$ is estimated with the data (G_i, Y_i), $i \in I_1$ using Bayes' theorem and a linear kernel density estimate.[7]

    b. Using the fair dummy from Step 4a, compute a test statistic for the k^{th} iteration as: $t^{(k)} = \frac{1}{N_{I_2}} \sum_{i \in I_2} \left(\hat{Y}_i - \hat{r}\left(\tilde{G}_i, Y_i\right)\right)^2$.

5. After *K* iterations, use the resulting $(K + 1) \times 1$ vector of test statistics to compute a p-value for the null hypothesis that $\hat{Y} \perp G \mid Y$ holds: $p_v = \frac{1 + I\{t^* \geq t^{(k)}\}}{1 + K}$, or the proportion of times that a test statistic as small or smaller than the observed test statistic was obtained under the null.

The fair dummies test can be implemented using the Python package *fair_dummies* (Romano et al., 2020). To test whether MLMI $\left(\hat{Y} \perp G \mid \eta\right)$ holds, rather than separation, we applied the fair dummies test by substituting η for *Y* in the procedure (Steps 1-5) described above. We used *K* = 1,000 iterations.

Methods of Romano et al. (2020): model fitting technique for separation

In addition to the fair dummies test, Romano et al. (2020) proposed a bias mitigation method to train a "fair" machine learning model approximately satisfying separation that also

[7] Using Bayes' theorem, $P(G = g|Y = y) = \frac{P(Y=y|G=g)P(G=g)}{P(Y=y|G=g)P(G=g)+P(Y=y|G=g')P(G=g')}$. Terms of the form $P(Y = y|G = g)$ are approximated with linear kernel density estimates.

leverages the fair dummy, $\tilde{G}$. Conceptually, this is achieved by fitting a predictive model $\hat{f}(\cdot)$ that minimizes a specific cost function, such that

$$\hat{f}(\boldsymbol{X}) = \underset{f}{\operatorname{argmin}} \left\{ (1-\varphi) \frac{1}{N_{train}} \left[\sum\nolimits_i \ell(Y_i, f(\boldsymbol{X}_i)) \right] + \varphi D\left(\left(\hat{Y}, G, Y\right), \left(\hat{Y}, \tilde{G}, Y\right) \right) \right\} \tag{11}$$

The minimized cost function includes a term for the usual loss function $\ell(\cdot)$ that penalizes for prediction error (e.g., squared error) and a regularization term for a discrepancy function $D(\cdot)$ which quantifies the distinction between two probability distributions, meant to penalize for a violation of separation. Because $\left(\hat{Y}, \tilde{G}, Y\right)$ satisfies separation by construction, the aim is to make $D\left(\left(\hat{Y}, \mathrm{G}, Y\right), \left(\hat{Y}, \tilde{G}, Y\right)\right)$ small, which would indicate $\left(\hat{Y}, G, Y\right)$ nearing separation. The parameter φ controls the trade-off between the emphasis on predictive accuracy versus fairness and can range from 0 to 1.

Romano et al. (2020)'s framework is based on a generative adversarial network architecture (Goodfellow et al., 2020) and involves training two sub-models, the predictive model $\hat{f}(\cdot)$ and a discriminator $\hat{d}(\cdot)$, both of which are FNNs. The discriminator is a binary classifier whose goal is to tell apart an observation as one of two class types, $\left(\hat{Y}, G, Y\right)$ or $\left(\hat{Y}, \tilde{G}, Y\right)$, which I will refer to as class observed (coded 1) and class dummy (coded 0), respectively. The performance of this discriminator serves as the basis of the discrepancy function $D(\cdot)$ in Equation 11. The predictive model is the focal machine learning model whose goal is to predict Y from predictors $\boldsymbol{X}$. It is trained to minimize prediction error while at the same time generate predictions $\hat{Y}$ that would fool the discriminator (i.e., impair the performance of the trained discriminator if it were applied to this $\hat{Y}$). In this way, the predictive model and the discriminator act as adversaries to one another.

To build the adversarial network structure, an initial predictive model is fit to the training data, minimizing prediction error only. Then, the following steps are repeated for *B* iterations to sequentially optimize each of the two sub-models within each iteration according to, or in response to, how the opposing sub-model was updated. First, as in the fair dummies test, a fair dummy is sampled from the distribution $P_{G|Y}$ for each training observation, or $\tilde{G}_i \sim P_{G|Y}(G_i \mid Y_i)$, $i \in I_{train}$. Using the current predictive model, the discriminator and its parameters $\boldsymbol{\theta_d}$ are optimized via gradient descent with the following cost function for *b* steps:

$$J_d(\boldsymbol{\theta_d}) = \frac{-1}{N_{train}}\left[\sum_i \log\left(\hat{d}_{\theta_d}\left(\hat{f}_{\theta_f}(\boldsymbol{X}_i), G_i, Y_i\right)\right) + \log\left(1 - \hat{d}_{\theta_d}\left(\hat{f}_{\theta_f}(\boldsymbol{X}_i), \tilde{G}_i, Y_i\right)\right)\right] \tag{12}$$

Equation 12 represents the binary cross entropy loss for the binary classification task of distinguishing between class observed and class dummy.[8] During this process, the parameters of the predictive model $\boldsymbol{\theta_f}$ remain static at their current values.

Given the updated discriminator, the parameters of the predictive model $\boldsymbol{\theta_f}$ are then optimized via gradient descent with the following cost function for *b* steps:

$$\begin{aligned} J_f(\boldsymbol{\theta_f}) = (1-\varphi)\frac{1}{N_{train}}\left[\sum_i \left(Y_i - \hat{f}_{\theta_f}(\boldsymbol{X}_i)\right)^2\right] + \varphi\left\|\mathrm{cov}(\hat{Y}, G) - \mathrm{cov}(\hat{Y}, \tilde{G})\right\|^2 \\ - \varphi\frac{1}{N_{train}}\left[\sum_i \log\left(\hat{d}_{\theta_d}\left(\hat{f}_{\theta_f}(\boldsymbol{X}_i), \tilde{G}_i, Y_i\right)\right)\right. \\ \left. + \log\left(1 - \hat{d}_{\theta_d}\left(\hat{f}_{\theta_f}(\boldsymbol{X}_i), G_i, Y_i\right)\right)\right] \end{aligned} \tag{13}$$

Equation 13 contains a term for the loss function (mean squared error; first term on the righthand side) and a discrepancy term (third term on the righthand side).[9] The second term on the righthand side stabilizes the learning process with an additional penalty to minimize the

[8] The discriminator is a two-layer FNN with a hidden layer of size 30 and ReLU activation.

[9] The predictive model is a two-layer FNN with a hidden layer of size 64 and ReLU activation.

difference between covariances of class observed and class dummy variables. During this process, the parameters of the discriminator $\boldsymbol{\theta_d}$ remain static at their current values. Note that in the discrepancy term of Equation 13, the placement of $\tilde{G}_i$ and G_i have flipped compared to how it appears in Equation 12. This is meant to penalize the predictive model for correct predictions made by the discriminator. This completes one iteration, and in the next iteration, the discriminator is optimized given this updated predictive model, and given the updated discriminator, the predictive model is optimized, and so on.

This model-fitting procedure can be implemented with the *fair_dummies* Python package (Romano et al., 2020). We used $B = 100$ iterations, $b = 50$ steps for both the predictive model and discriminator per iteration, and a learning rate of 0.01 in the simulation. While these hyperparameters could be tuned within the training set per model in each replication, this is computationally expensive and therefore, we used these fixed values throughout the simulation study. We used $\varphi = 0.9$ for the trade-off parameter to place a heavy emphasis on satisfying separation, which is the highest value considered in Romano et al. (2020).

Simulation outcomes and analysis

We studied the effects of measurement bias on MLMI by comparing predictions obtained from two different machine learning models per replication (Models 1 and 2; see Table 3 for summary of the different models). Using the model-fitting procedure of Romano et al. (2020), Model 1 was trained on an invariant outcome variable (Y_1), and Model 2 was trained on a non-invariant outcome variable (Y_4). Then, we obtained predictions on the test set from each model and assessed them each for MLMI using the fair dummies test, recording the resulting p-value from each replication. Within each condition, we compared the distribution of p-values for Model 1 versus Model 2, as well as the proportion of replications in which MLMI was violated

in Model 1 versus Model 2, according to an alpha level of 0.05. Distributions of p-values were compared graphically via boxplots per condition, and the difference in proportions was quantitatively assessed for significance using McNemar's test and Cohen's g effect sizes per condition, where pairings were made by replication. Cohen's g was calculated as $|Q - 0.5|$, where Q is the proportion of replications in a condition where Model 1 met MLMI but Model 2 did not, out of the replications where Models 1 and 2's MLMI outcome differed (Cohen, 1988).

In addition to MLMI results, we also examined the effects of measurement bias on predictive accuracy, given that predictive performance is another key consideration in machine learning. We recorded predictive performance using root mean squared error (RMSE; $\sqrt{N_{test}^{-1}\sum_{i=1}^{N_{test}}(Y_i - \hat{Y}_i)^2}$) and the correlation between η and each model's predictions on the test set ($r_{\eta,\hat{Y}}$). This allows for inspection of predictive performance from two different perspectives of (1) accurate reproduction of the observed outcome variable and (2) high correlation with the true underlying target latent variable. We made pairwise comparisons of each performance metric obtained in a replication between Model 1 versus Model 2. This was inspected for each metric graphically via boxplots and quantitatively via paired samples t-tests and Cohen's *d* effect sizes within each condition, where pairings were made by replication.

Note that Y_2 and Y_3 are not investigated as outcome variables in Models 1 nor 2, given that they were generated under identical properties as Y_1. However, generating these outcome variables was necessary for overidentification of the MG-CFA model, fit in part 2 of the simulation study.

Table 3

Summary of the four machine learning models built across simulation parts 1 and 2

	Outcome variable	Outcome variable exhibits measurement bias	Bias mitigation technique	Metrics to record from each replication
Simulation part 1				
Model 1	Y_1	No	Regularization for separation ($\varphi = 0.9$)	P-value from fair dummies test for MLMI, RMSE, $r_{\eta,\hat{Y}}$
Model 2	Y_4	Yes	Regularization for separation ($\varphi = 0.9$)	P-value from fair dummies test for MLMI, RMSE, $r_{\eta,\hat{Y}}$
Simulation part 2				
Model 3	Y_4	Yes	Regularization for MLMI using $\hat{\eta}$ (φ varies by condition)	P-value from fair dummies test for MLMI, RMSE, $r_{\eta,\hat{Y}}$
Model 4	Y_4	Yes	Regularization for separation (φ varies by condition)	P-value from fair dummies test for MLMI, RMSE, $r_{\eta,\hat{Y}}$

In addition to examining the effects of measurement bias within each condition separately, we studied the systematic influence of the simulation factors on these results using classification and regression trees (CART; Breiman et al., 1984). In these CART models, a single replication was the unit of analysis, the four simulation factors were the predictors, and a simulation outcome of interest was the outcome variable. Using CART to analyze Monte Carlo simulation results can be advantageous over more conventional approaches (e.g., analysis of variance) due to its ability to automatically detect higher-order interactions among predictors, flexibility in the type of outcome variable (continuous, binary, multinomial) it can handle, and its interpretable model output in the form of a decision tree diagram that intuitively describes relationships between predictors and outcomes (Gonzalez et al., 2018). We built a multi-class classification tree for the categorical simulation outcome of MLMI, with the following four

classes: (1) *Model 1 violated MLMI and Model 2 did not*; (2) *Model 2 violated MLMI and Model 1 did not*; (3) *both Models 1 and 2 violated MLMI*; and (4) *neither model violated MLMI*. We built two separate regression trees for the two continuous simulation outcomes related to predictive performance: (1) the paired difference in RMSE between Models 1 and 2; and (2) the paired difference in $r_{\eta,\hat{Y}}$ between Models 1 and 2. The number of splits in a tree was determined using cost-complexity pruning with cross-validation, as described in Gonzalez et al. (2018). Because a large, complex tree (i.e., many splits) challenges interpretation and generalizability to new data, this searches for an optimal tree size while maintaining the tree's ability to generate accurate out-of-sample predictions. Because splits on a predictor in CART are made sequentially according to the degree to which they homogenize the observations (replications) with respect to the outcome variable, the hierarchy of splits and the cut points of each split in the resulting tree diagrams were examined to interpret the importance and specific influence of the simulation factors on the simulation outcomes. We examined the R^2 values and misclassification rate from each of the regression and classification trees, respectively, to gauge how well the simulation factors predicted each simulation outcome. CART models were built using the R package *rpart* (Therneau & Atkinson, 2022).

Simulation hypotheses

We hypothesized that MLMI will generally be violated when a model is trained on a non-invariant outcome variable (Model 2) compared to when trained on an invariant outcome variable (Model 1). This effect of measurement bias on MLMI may be most present under larger magnitudes of non-invariance, larger factor loadings, and when G is included as a predictor. With regards to predictive performance, we hypothesized that in general, Model 2 will have comparable predictive performance as Model 1 in terms of RMSE when G is included as a

predictor, but worse in terms of $r_{\eta,\hat{Y}}$ across all conditions. This effect of measurement bias on predictive performance may be most pronounced under larger magnitudes of non-invariance and larger factor loadings.

It is worth noting that grouping variables often in practice have imbalanced groups rather than each group comprising approximately 50% of the sample, as configured in our simulation study. We therefore conducted a supplemental simulation study, in which groups 1 and 0 comprised 75% and 25% of the sample, respectively (i.e., generate the grouping variable as $G_i \sim \text{Bernoulli}(0.75)$ in Equation 4). Because the results remained largely the same as when using balanced groups, we only present the results from the main simulation study in Sections 3.2 and 4.2. Results of this supplemental simulation study can be found in the Appendix.

Section 3.2: Results

Machine learning fairness

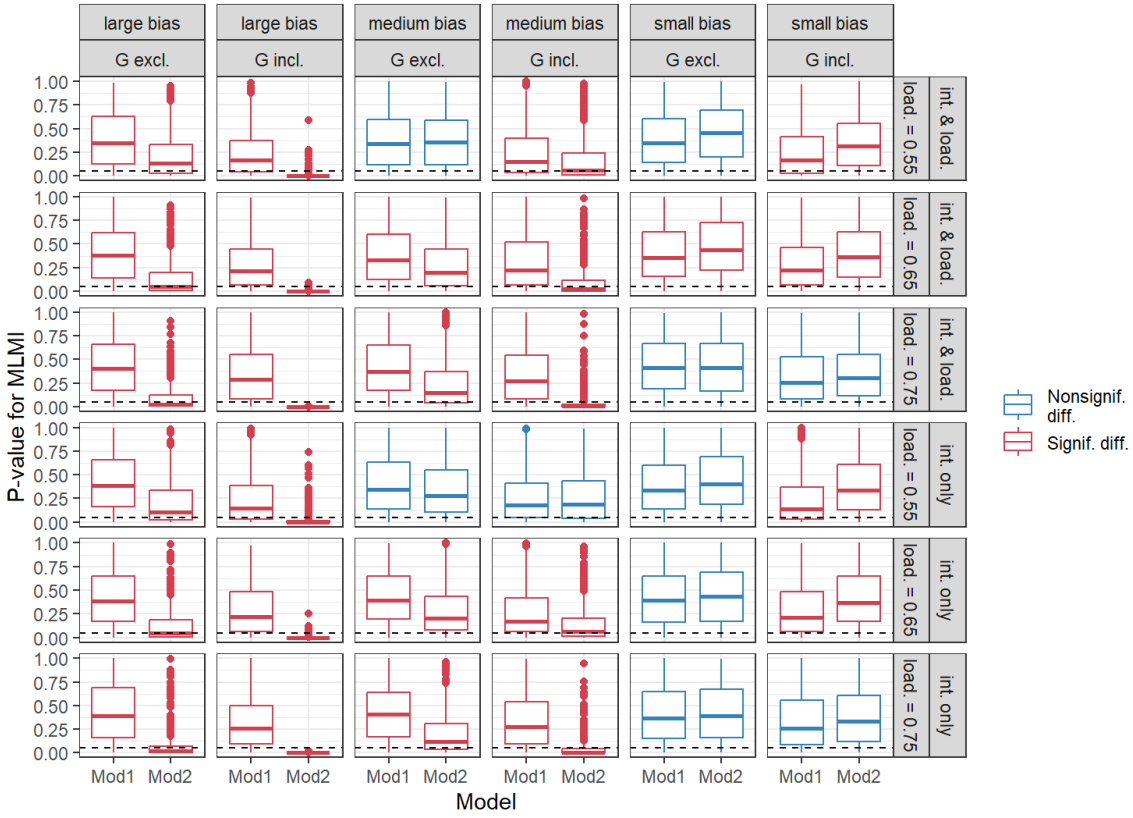
First, we investigated the effects of measurement bias on models' abilities to meet MLMI, by comparing the MLMI results between Model 1 (trained on invariant outcome variable) and Model 2 (trained on non-invariant outcome variable). The condition-wise distributions of p-values of the fair dummies test for MLMI from the two models are plotted in Figure 1. The results from Model 1 provide a baseline for what can be expected under an invariant outcome variable and show that when constrained for separation, MLMI is typically satisfied in all conditions (MLMI violations in 16.8% of replications across all 36 conditions). In contrast to this, we found Model 2 to violate MLMI substantially more in select conditions, indicating that with a non-invariant outcome variable, constraining for separation can be insufficient to satisfy MLMI. In particular, under medium and large magnitudes of non-invariance, Model 2 violated MLMI in 39.7% and 73.0% of replications, respectively, when

pooling across levels of all other simulation factors. MLMI violations were especially salient with higher factor loadings and the inclusion of G as a predictor. Under small magnitudes of non-invariance, Model 2 violated MLMI in just 9.52% of replications, indicating that with a small enough magnitude of bias, constraining for separation may be sufficient to mitigate the effects of measurement bias on MLMI.

The results of McNemar's tests and Cohen's g effect sizes calculated within each of the 36 conditions corroborated these observations. In all conditions with a large magnitude of non-invariance, regardless of the levels of all other simulation factors, there was a significant ($p < \frac{.05}{36}$; Bonferroni-corrected for multiple testing across conditions) and at least moderate ($g \geq 0.15$) difference in the proportion of MLMI violation between Models 1 and 2, with Model 2 consistently violating MLMI more than Model 1. All conditions with a medium magnitude of non-invariance, except when $\lambda = 0.55$, also showed a significant difference in the proportion of MLMI violation between Models 1 and 2, with Model 2 violating MLMI more than Model 1. Under small magnitudes of non-invariance, most conditions showed no significant difference in the proportion of MLMI violation between Models 1 and 2. However, in a couple of the small-magnitude conditions, namely when G was included as a predictor and with lower factor loadings ($\lambda = 0.55, 0.65$), we found significant differences, where Model 2 violated MLMI *less* than Model 1. This unexpected finding could be a result of the outcome variables being less reliable indicators of η (due to the low λ), which may have led to lower-quality predictions, combined with the magnitude of non-invariance in Y_4 not being large enough, such that patterns of MLMI violation did not surface as expected.

Figure 1

Condition-wise boxplots of the p-values of the fair dummies test for MLMI for Models 1 and 2



Note. A "significant" difference refers to McNemar's $p < \frac{.05}{36}$ and Cohen's $g \geq 0.15$.

Moving beyond condition-wise analyses, we next used CART models to analyze the Monte Carlo data from all conditions together to investigate the systematic influence of the simulation factors on the simulation outcome of MLMI in Models 1 and 2. Figure 2 presents the resulting classification tree diagram. While the MLMI simulation outcome modeled in this CART analysis was originally a four-level categorical variable, the optimal CART model found here using cross-validation, and its resulting decision rule, classified replications into one of two classes: (1) *neither model violated MLMI* (i.e., no consequences of measurement bias) or (2)

Model 2 violated MLMI and Model 1 did not (i.e., consequence of measurement bias present). The misclassification rate of this classification tree was 32.4%.

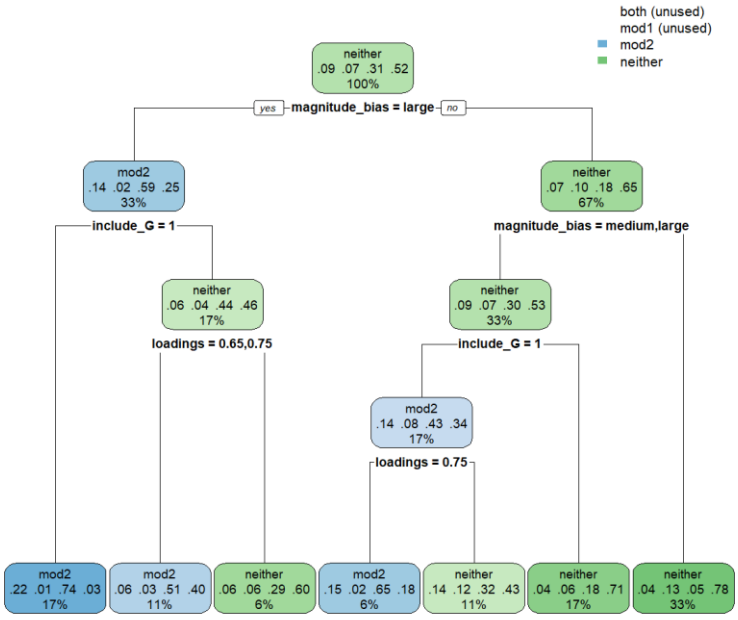
Consistent with the observations made from Figure 1, the hierarchy of splits in Figure 2 revealed that the most useful simulation factor in explaining the effects of measurement bias on MLMI was the magnitude of non-invariance, followed by the inclusion/exclusion of G and the value of λ. As hypothesized, the effects of measurement bias on MLMI were most prevalent with larger magnitudes of non-invariance, the inclusion of G, and higher values of λ. Figure 2 shows that under a large magnitude of non-invariance with G included as a predictor, or with G excluded but with higher factor loadings, replications were optimally classified as *Model 2 violated MLMI and Model 1 did not*. With a medium magnitude of non-invariance, G included as a predictor, and $\lambda = 0.75$, replications were also optimally classified as *Model 2 violated MLMI and Model 1 did not*. Under all other conditions, the optimal classification was *neither model violated MLMI*.

The effects of measurement bias on MLMI becoming more prevalent with larger magnitudes of non-invariance makes intuitive sense. The inclusion of G being associated with more prevalent effects of measurement bias is also expected, given that the machine learning model explicitly "sees" the grouping variable that characterizes the measurement bias in the outcome variable, such that this pattern of non-invariance may more easily be learned and manifested in the model predictions. On the other hand, with G excluded, the model may be more "oblivious" to the measurement bias, such that it is less likely to learn and generate predictions reflecting the pattern of non-invariance. A possible explanation for the effects of measurement bias on MLMI becoming more prevalent with higher factor loadings is that with a lower λ, the unreliability in the outcome variable leads to lower-quality model predictions that

may muddle any patterns of MLMI violation, that are otherwise generated under higher values of λ and therefore higher-quality predictions. Because the type of non-invariance is not used in any splits of the tree in Figure 2, we can interpret that the effect of measurement bias on MLMI did not meaningfully differ according to whether the bias was in intercepts only or in both loadings and intercepts.

Figure 2

Tree diagram of a classification tree modeling the MLMI outcome in Models 1 and 2, fit to simulation results across all conditions



Note. In each node of the tree diagram, the top value represents the majority class, the middle value represents the proportional breakdown of the classes of the observations belonging to that node, and the bottom value represents the percentage of total observations that belong to that node.

Predictive performance

Next, we studied the effects of measurement bias on models' predictive abilities by comparing the test-set performance between Models 1 and 2. Because predictive performance is a key consideration for the utility of machine learning models, it can be useful to contextualize

fairness results with performance considerations and to understand any trade-offs between fairness and predictive accuracy.

Condition-wise distributions of the performance metrics of Models 1 and 2 are presented in Figures 3a (RMSE) and 3b ($r_{\eta,\hat{Y}}$). In terms of accurate reproduction of the observed outcome variable (i.e., RMSE), as hypothesized, Model 2 performed worse than Model 1 in all conditions with medium and large magnitudes of non-invariance, suggesting a performance drop associated with measurement bias when the magnitude of bias is substantial. The performance gap grew with increasing factor loadings but narrowed with the inclusion of G as a predictor. This is likely because without "seeing" G (i.e., when G is excluded as a predictor), it becomes relatively more difficult for Model 2 to accurately predict Y_4 than it is for Model 1 to predict Y_1, given that Y_4 additionally (i.e., above and beyond the predictors in $\boldsymbol{X}$) and directly depends on G, as implied by the definition of measurement bias. On the other hand, with G included, Model 2 is better able to account for the additional variability in Y_4 that is directly due to G, such that the effect of measurement bias on predictive performance is subdued. Under small magnitudes of non-invariance, any performance gap between Models 1 and 2 appeared trivial. These observations were supported statistically with paired sample t-tests and Cohen's *d* effect sizes calculated within each of the 36 conditions, which revealed significant ($p < \frac{.05}{36}$) and at least moderate ($|d| \geq 0.5$) differences only in conditions with medium or large magnitudes of non-invariance.

Findings differed slightly when studying predictive performance in terms of the predictions' correlations with the target latent variable (i.e., $r_{\eta,\hat{Y}}$). We still found Model 2 to perform significantly ($p < \frac{.05}{36}$ and $|d| \geq 0.5$) worse than Model 1 in some conditions with medium and large magnitudes of non-invariance, with the performance gap growing with increasing factor loadings, as before. However, unlike the RMSE, the performance gap in $r_{\eta,\hat{Y}}$

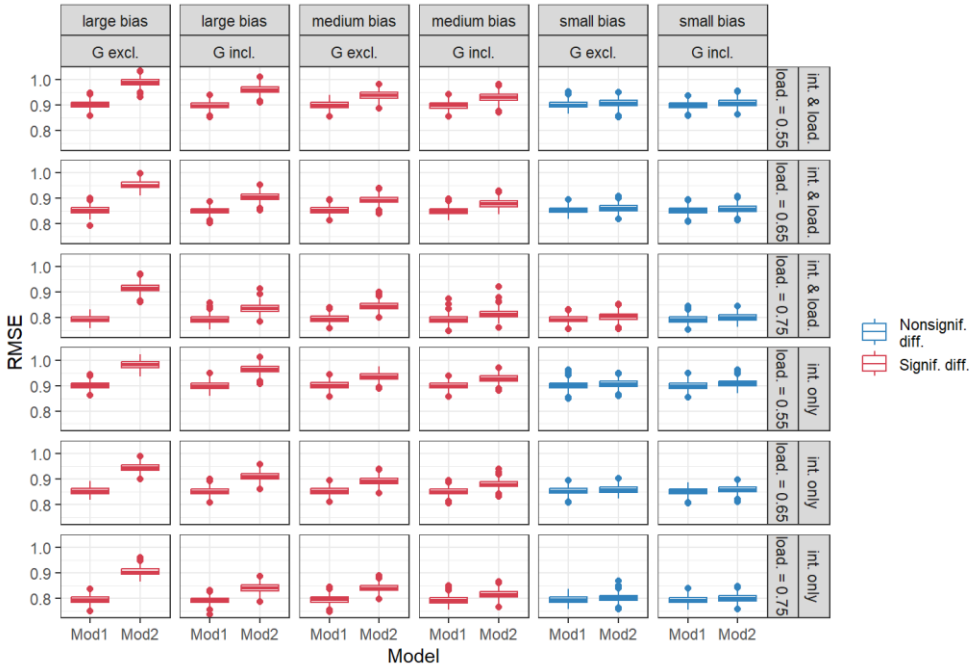
almost entirely closed with the inclusion of G as a predictor, such that Model 2 performed comparably to Model 1 in those conditions, meaning there was no effect of measurement bias on $r_{\eta,\hat{Y}}$ in those conditions. This pattern held for high magnitudes of non-invariance or when $\lambda = 0.75$. With medium magnitudes of non-invariance and lower factor loadings, we found no observable difference between Models 1 and 2, even when G was excluded as a predictor.

Furthermore, in three conditions with small magnitudes of non-invariance, namely when G was excluded as a predictor and with lower factor loadings ($\lambda = 0.55, 0.65$), we found Model 2 to outperform Model 1, although by a small margin. As discussed with the MLMI results above, these unexpected results associated with lower values of λ may be attributed to the unreliability of the outcome variable as indicators of η in those conditions, such that the "true" effect of the measurement bias on predictive performance was muddled due to the models struggling to generate accurate predictions in the first place. This may also explain the phenomenon of the performance gap narrowing with lower factor loadings in the medium- and high-magnitude conditions noted above.

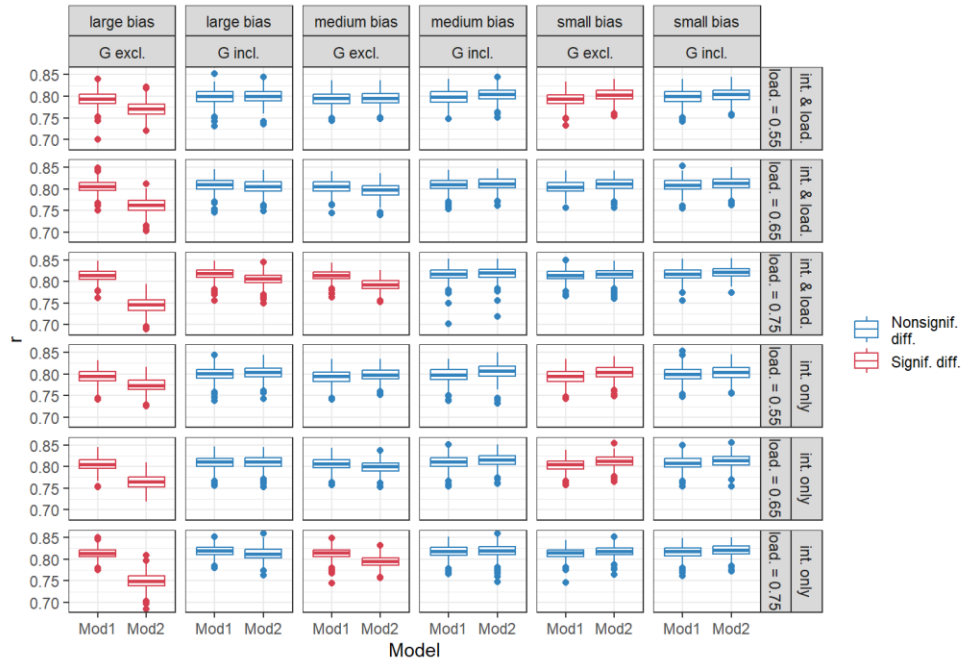
Figure 3

Condition-wise boxplots of the test performance of Models 1 and 2 in terms of RMSE (Panel a) and $r_{\eta,\hat{Y}}$ (Panel b)

Panel a



Panel b



Note. A "significant" difference refers to paired sample t-test $p < \frac{.05}{36}$ and Cohen's $|d| \geq 0.5$.

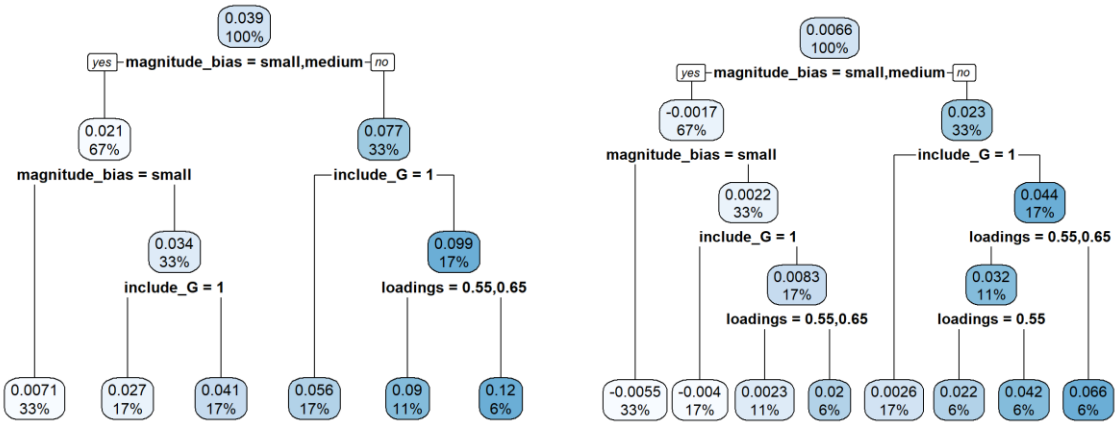
To study the systematic influence of the simulation factors on the performance gap between Models 1 and 2, we analyzed the simulation results from all conditions together using CART models. Figures 4a and 4b present the resulting regression tree diagrams modeling the pairwise (by replication) differences in RMSE and $r_{\eta,\hat{Y}}$, respectively, between Models 1 and 2. Note that in both regression trees, the differences in performance metrics were calculated such that a positive difference corresponds to Model 1 performing better than Model 2. The R^2 values of these regression trees were 0.70 and 0.53 for the RMSE and $r_{\eta,\hat{Y}}$ difference, respectively.

The tree diagrams in Figures 4a and 4b allow for similar interpretations and corroborate the observations made from Figures 3a and 3b: the performance gap between Models 1 and 2 (i.e., the effect of measurement bias on predictive performance) became more prevalent with a larger magnitude of non-invariance, exclusion of G as a predictor, and higher factor loadings. Again, because the type of non-invariance was not used as a split in either tree, we can interpret that the effect of measurement bias on predictive performance was consistent for both non-invariant intercepts only and non-invariant intercepts and loadings.

Figure 4

Tree diagram of a regression tree modeling the difference in RMSE (Panel a) and $r_{\eta,\hat{Y}}$ (Panel b) between Models 1 and 2, fit to simulation results across all conditions

Panel a **Panel b**



Note. In each node of the tree diagram, the top value represents the mean value of the simulation outcome among observations belonging to that node, and the bottom value represents the percentage of total observations that belong to that node.

Summary of simulation part 1

We proposed a framework that allows for tests of measurement invariance of outcome variables in machine learning. The importance and utility of such a framework is demonstrated in this simulation, which revealed several practical consequences of measurement bias on the fairness and predictive performance of machine learning models. First, as hypothesized, the results suggest that with medium and large magnitudes of non-invariance, we can expect the use of a non-invariant outcome variable to result in machine learning predictions that violate MLMI. These MLMI violations were observed even when a strong constraint ($\varphi = 0.9$) for separation was made during model training, meaning MLMI cannot be approximated with separation if the outcome variable exhibits measurement bias. In contrast, with an invariant outcome variable or

under a small magnitude of non-invariance, MLMI was often satisfied when constraining for separation. In general, we found measurement bias to harm predictive performance.

We also found that by excluding G as a predictor, MLMI was less frequently violated than when G was included. While this may appear to suggest *fairness through unawareness* as a possible treatment for mitigating the effects of measurement bias on MLMI, we do not unconditionally recommend this approach. First, from a predictive performance perspective, the simulation showed that the performance drop associated with measurement bias, both in terms of accurate reproduction of the observed outcome variable and a high correlation with the target latent variable, is most salient when G is excluded, which can disincentivize *fairness through unawareness*. Second, the simulation showed that the effect of measurement bias on MLMI can still be present even when G is excluded, if the magnitude of non-invariance is large enough. Third, while not demonstrated with this simulation, previous studies have discussed how *fairness through unawareness* may be ineffective if other variables that are highly correlated with G, that can be used to approximate G, are included as predictors in the machine learning model (Barocas et al., 2019; Kusner et al., 2017).

Finally, throughout this simulation, we encountered a couple of counterintuitive findings, where the effects of measurement bias on both MLMI and predictive performance were subdued or even reversed (e.g., no performance drop associated with measurement bias, MLMI being violated more with an invariant than a non-invariant outcome variable) when factor loadings were low. While this may appear to suggest that it is favorable to use outcome variables with low factor loadings in the presence of measurement bias, this is not necessarily the case. Low factor loadings themselves are unfavorable for the utility of a machine learning model in general, as they indicate low reliability of outcome variables as indicators of the target latent variable. In

fact, the simulation showed that with low factor loadings, MLMI can be violated even if the outcome variable does not exhibit measurement bias, and predictive performance was consistently lowered with decreasing factor loadings within each model. As such, having low loadings should not be viewed as a treatment to counter the effects of measurements bias—it should be more so gathered from this simulation that in those conditions with low factor loadings, the outcome variable was not reliable enough for its model predictions to generate patterns of non-invariance that were otherwise generated under higher loadings, such that measurement bias appeared to "matter less".

In sum, measurement bias in the outcome variable demonstrably presents concerns for both fairness and predictive performance in machine learning, especially with growing magnitudes of non-invariance. Furthermore, these effects are not entirely addressed with treatments such as *fairness through unawareness* or by constraining for separation during model training. Because η is unobserved and testing for MLMI is not possible in practice, testing for measurement invariance in the outcome variable is one way to check for a possible source of MLMI violation. This highlights the importance of assessing for measurement invariance using the proposed framework and that it brings additional value above and beyond existing fairness considerations (i.e., excluding G as a predictor, testing or constraining the model for separation).

CHAPTER 4: SIMULATION STUDY PART 2

Section 4.1: Methods

In simulation part 1, we demonstrated why assessing for measurement bias in machine learning using the proposed framework matters by studying its consequences for machine learning bias and predictive performance. Now, if evidence of non-invariance is found in an outcome variable using the proposed framework, a natural follow-up question may be what to do about it. One simple option is to abandon the non-invariant outcome variable and to seek a different one. However, there may be practical cases where changing the outcome variable is not well-received. For example, the machine learning model in question may already be deployed and integrated within a larger system (e.g., a medical expense prediction model used in a health system; see Chapter 1), such that there is hesitation to change such a fundamental aspect of the model. As the second part of the simulation study, we explored an alternative option—to address and mitigate the bias.

In a traditional psychometric setting, measurement bias may be accounted for at the level of the latent variable by modeling the non-invariance using a partially invariant model. In a partially invariant model, items found to be non-invariant have model parameters (i.e., loadings and intercepts) that are allowed to vary across groups in the MG-CFA model (Byrne et al., 1989). For example, a partial strong invariance model constrains only a portion of intercepts to be equal across groups. Non-invariant parameters are typically identified by examining residuals (difference between observed and model-implied moments) and modification indices (Millsap &

Olivera-Aguilar, 2012), which help to determine which across-group equality constraint should be lifted to improve model fit and for partial invariance to hold. We combined this with the bias mitigation method of Romano et al. (2020) to explore how machine learning models approximately satisfying MLMI may be trained, even when the outcome variable is non-invariant.

Proposed bias mitigation technique

To do so, we propose to simply modify the discrepancy term in the cost function of the predictive model in Romano et al. (2020)'s fair dummies model-fitting procedure (Equation 11) to minimize $D\left(\left(\hat{Y}, G, \eta\right), \left(\hat{Y}, \tilde{G}, \eta\right)\right)$ instead of $D\left(\left(\hat{Y}, G, Y\right), \left(\hat{Y}, \tilde{G}, Y\right)\right)$. However, because η is unobserved in practice, we approximate it with factor scores estimates obtained from a partially invariant MG-CFA model (fit following the proposed framework), in which the measurement bias has been "corrected" at the level of the latent variable. Therefore, a predictive model $\hat{f}(\cdot)$ is fit, such that

$$\hat{f}(\boldsymbol{X}) = \underset{f}{\operatorname{argmin}} \left\{ (1-\varphi) \frac{1}{N_{train}} \left[\sum\nolimits_i \ell(Y_i, f(\boldsymbol{X}_i)) \right] + \varphi D\left(\left(\hat{Y}, G, \hat{\eta}\right), \left(\hat{Y}, \tilde{G}, \hat{\eta}\right)\right) \right\} \tag{14}$$

where $\hat{\eta}$ are factor score estimates from a partially invariant MG-CFA model, Y is a non-invariant outcome variable, and the fair dummy $\tilde{G}$ is now sampled from $P_{G|\hat{\eta}}$. Factor score estimates are used to represent each observation's placement on the latent variable according to a factor model (DiStefano et al., 2009). However, factor score estimates are indeterminate, meaning that there can be infinitely many sets of factor scores that are consistent with the factor model at hand (i.e., there is no unique solution; Grice, 2001).

Simulation factors and model building

Thus, the goal of simulation part 2 was to study the effectiveness of using factor score estimates in generating machine learning predictions fulfilling MLMI in the presence of measurement bias. Following simulation part 1, we recorded the simulation conditions under which Model 2 (trained on non-invariant outcome variable) predictions failed to meet MLMI for over 25% of the replications. Then, for those conditions only, we repeated the data generation steps from simulation part 1 and trained a new machine learning model (Model 3) using the proposed bias mitigation technique, as in Equation 14, with the non-invariant Y_4 as the outcome variable.

For simulation part 2, we considered two additional simulation factors: (1) the value of the trade-off parameter φ (three levels; $\varphi = 0.5, 0.7, 0.9$) and (2) the use of covariate-informed factor score estimates (two levels; covariates are used or not used in calculating factor scores). A possible way to improve factor score estimates is to model the structural relationships between η and the predictors in $\boldsymbol{X}$ (i.e., predictors of the latent factor, or "covariates") in the factor model (Curran et al., 2016). Therefore, in simulation conditions using covariate-informed factor scores, we included the additive main effects of each predictor on the latent factor in the MG-CFA model from which the factor scores were calculated. However, it should be noted that these structural relationships are misspecified, given that the predictor matrix $\boldsymbol{X}$ contains some noise predictors, $\boldsymbol{X}$ excludes some signal predictors, and the data generating function for η is not an additive function of all predictors in $\boldsymbol{X}$. We believe this misspecification is expected in practice, given that in a machine learning setting, the analyst typically lacks strong *a priori* knowledge of structural relationships among variables and aims to empirically identify those relationships in a data-driven manner (see Section 1.1). As such, with this simulation factor, our aim was to

investigate whether the use of covariates, despite the misspecification, improves the calculation of factor scores and, as a result, the effectiveness of the proposed bias mitigation technique.

We ran 500 replications in each condition, with a total of $6 \times c$ conditions, where c was the number of conditions flagged from simulation part 1 as described above. To compute factor scores, a MG-CFA model was fit to $\boldsymbol{Y} = [Y_1, Y_2, Y_3, Y_4]$, as in the proposed framework, using the R package *lavaan* (Rosseel, 2012). We used the regression method (Thurstone, 1935) to calculate factor scores, which are computed per observation i as

$$\hat{\eta}_i = (\boldsymbol{Y}_i - \boldsymbol{\mu_Y})\boldsymbol{\Sigma_Y}^{-1}\boldsymbol{\Lambda}_{G_i}\boldsymbol{\Psi} + \boldsymbol{\kappa} \tag{15}$$

where $\boldsymbol{\mu_Y}$ and $\boldsymbol{\Sigma_Y}$ are the mean vector and (co)variance matrix, respectively, of the observed outcome variables, $\boldsymbol{\Lambda}_{G_i}$ is the factor loading matrix for group for group G_i, and $\boldsymbol{\kappa}$ and $\boldsymbol{\Psi}$ refer to the factor mean vector and (co)variance matrix, respectively. The MG-CFA model was identified by standardizing the factor in group 1 and constraining all intercepts and loadings except for Y_4 to be equal across groups. For Y_4, intercepts and/or loadings, depending on the simulation condition, were freely estimated with no across-group equality constraints to model partial invariance. Model 3 was fit using the *fair_dummies* package (Romano et al., 2020), manually customized (in the source code) to train models with the modified cost function, as in Equation 14. We used the same hyperparameters as described in simulation part 1. To provide a benchmark to which the proposed bias mitigation technique (Model 3) can be compared, we trained an additional model (Model 4) in each replication using the original model-fitting procedure for separation (Romano et al., 2020), as in Equation 11.

Simulation outcomes and analysis

Using the predictions obtained from Models 3 and 4, we recorded the p-value for the fair dummies test for MLMI (testing for $\hat{Y} \perp G \mid \eta$, not $\hat{Y} \perp G \mid \hat{\eta}$) from each model in each

replication. Per condition, we examined the proportion of replications in which Model 3 predictions violated MLMI according to an alpha level of 0.05, to determine how successful the proposed technique is in building a "fair" model in the presence of measurement bias. We additionally compared the proportions of MLMI violation between Models 3 and 4 to investigate the effectiveness of the proposed strategy relative to the original model-fitting procedure for separation. Similar to simulation part 1, the distributions of p-values were compared graphically via boxplots per condition, and the difference in proportions of MLMI violation were assessed for significance using McNemar's test and Cohen's *g* effect sizes per condition, where pairings were made by replication.

To evaluate the proposed technique from the perspective of predictive performance, we additionally examined how predictive accuracy may be affected due to the modified cost function, compared to the original cost function for separation. Per condition, we made pairwise comparisons in the RMSE and $r_{\eta,\hat{Y}}$ obtained in each replication from Model 3 versus Model 4. We made across-model comparisons for each performance metric graphically via boxplots and quantitatively via paired samples t-tests and Cohen's *d* effect sizes within each condition, where pairings were made by replication.

Similar to simulation part 1, we additionally studied the systematic influence of the simulation factors on these simulation outcomes using CART models. We built a multi-class classification tree for the categorical simulation outcome of MLMI, with the following four classes: (1) *Model 3 violated MLMI and Model 4 did not*; (2) *Model 4 violated MLMI and Model 3 did not*; (3) *both Models 3 and 4 violated MLMI*; and (4) *neither model violated MLMI*. We built two regression trees for the two continuous simulation outcomes for predictive performance: (1) the paired difference in RMSE between Models 3 and 4 and (2) the paired

difference in $r_{\eta,\hat{Y}}$ between Models 3 and 4. All other procedures to analyze the simulation results remained the same as part 1.

Simulation hypotheses

We hypothesized that Model 3 (proposed technique) will satisfy MLMI with higher values of φ and with larger factor loadings. Model 3 may also show a higher rate of MLMI satisfaction under covariate-informed factor scores given that the quality of factor score estimates are expected to improve, such that they become better approximations of the target latent variable. However, differences across factor score types may not manifest when φ is high and with large factor loadings, as MLMI will likely be satisfied under both types of factor scores in such conditions. In terms of predictive performance, we hypothesized that Model 3 will have a lower RMSE but a higher $r_{\eta,\hat{Y}}$ compared to Model 4, with this difference becoming more pronounced with larger magnitudes of non-invariance.

Section 4.2: Results

Of the 36 total conditions tested in simulation part 1, there were $c = 20$ conditions where over 25% of its replications resulted in MLMI violation by a model trained on a non-invariant outcome variable, even when constrained for separation. This included all 12 conditions with large magnitudes of non-invariance, plus 8 conditions with medium magnitudes of non-invariance when G was included as a predictor (6 conditions) or when G was excluded and $\lambda = 0.75$ (2 conditions).

In simulation part 2, we investigated whether the proposed bias mitigation technique can act as an alternative treatment to satisfy MLMI in the presence of measurement bias, in those 20 conditions where the original bias mitigation technique for separation was unsuccessful in satisfying MLMI. We crossed those 20 conditions with two new simulation factors (value of φ

and the use of covariate-informed factor scores), which resulted in 120 total conditions tested in simulation part 2. We did not encounter any issues of model convergence for all MG-CFA models fit throughout this simulation.

Machine learning fairness

First, we investigated the effectiveness of the proposed bias mitigation technique to train models satisfying MLMI in the presence of measurement bias. Figure 5 presents the condition-wise distributions of p-values from the fair dummies test for MLMI for Model 3 (proposed technique regularizing for $\hat{Y} \perp G \mid \hat{\eta}$) and Model 4 (original technique regularizing for separation, i.e., $\hat{Y} \perp G \mid Y$). Focusing first on the MLMI results of Model 3 only, we found that if G is excluded as a predictor, the proposed technique was largely successful in training models satisfying MLMI, regardless of the levels of all other simulation factors (11.7% MLMI violation among all replications where G is excluded). If G is to be included as a predictor, Figure 3 shows that a strong emphasis on the regularization term ($\varphi = 0.9$) was needed for Model 3 to largely meet MLMI (23.1% MLMI violation). With $\varphi = 0.7$ or 0.5, Model 3 largely violated MLMI (63.4% and 95.4% MLMI violation, respectively), meaning the proposed technique was ineffective in satisfying MLMI with any weaker emphasis on the regularization term. Because φ signifies the emphasis on fairness (i.e., satisfying $\hat{Y} \perp G \mid \hat{\eta}$) during model training, it follows that the higher the value of φ, the more successful the proposed technique was in meeting MLMI. In general, we found conditions with higher values of λ to have higher success in meeting MLMI. This is likely due to the quality of factor score estimates improving with increasing factor loadings, such that $\hat{Y} \perp G \mid \hat{\eta}$ becomes a better approximation to $\hat{Y} \perp G \mid \eta$ (i.e., MLMI). In fact, the factor score estimates on average correlated with η at 0.83, 0.88, and 0.92 for $\lambda = 0.55, 0.65, 0.75$, respectively, when pooling across levels of all other simulation factors.

Contrary to our hypothesis, Model 3's ability to meet MLMI did not appear to meaningfully differ according to the use of covariate-informed factor scores. This may be due to the misspecification of the structural relationships between the predictors and η in the factor model, as discussed in Section 4.1. Accordingly, examining the correlation between the factor score estimates and η revealed only a minimal improvement with the inclusion of covariates, with the average correlation (when pooling across levels of all other simulation factors) being 0.90 for covariate-informed factor scores and 0.87 for factor scores computed without covariates.

Next, we compared the MLMI results of Model 3 to those of Model 4 in Figure 5. This comparison helps to demonstrate the improvement that regularizing for $\hat{Y} \perp G \mid \hat{\eta}$ with the proposed technique provides over the original model-fitting procedure for separation. Figure 5 shows that differences between Models 3 and 4 with respect to MLMI violation were most salient in conditions where G was included as a predictor and $\varphi = 0.9$, where Model 3 violated MLMI much less frequently than Model 4 (23.1% MLMI violation for Model 3 versus 76.3% for Model 4). In simulation part 1, we found MLMI violations were most prevalent when G was included as a predictor, so it follows that these are the conditions in which the proposed technique was most useful and differentiated from Model 4's results. This is corroborated with McNemar's tests and Cohen's g effect sizes, which revealed significant ($p < \frac{.05}{120}$; Bonferroni-corrected for multiple testing across conditions) and at least moderate-sized ($g \geq 0.15$) differences in the proportion of MLMI violation between Models 3 and 4 in nearly all such conditions. With G included as a predictor and $\varphi = 0.7$, we found Model 3 to frequently violate MLMI, although still significantly ($p < \frac{.05}{120}$ and $g \geq 0.15$) less than Model 4. With G included as a predictor and $\varphi = 0.5$, Model 3 violated MLMI just as much as Model 4, resulting in some nonsignificant ($p \geq \frac{.05}{120}$ and $g < 0.15$) differences between the two models. This again shows

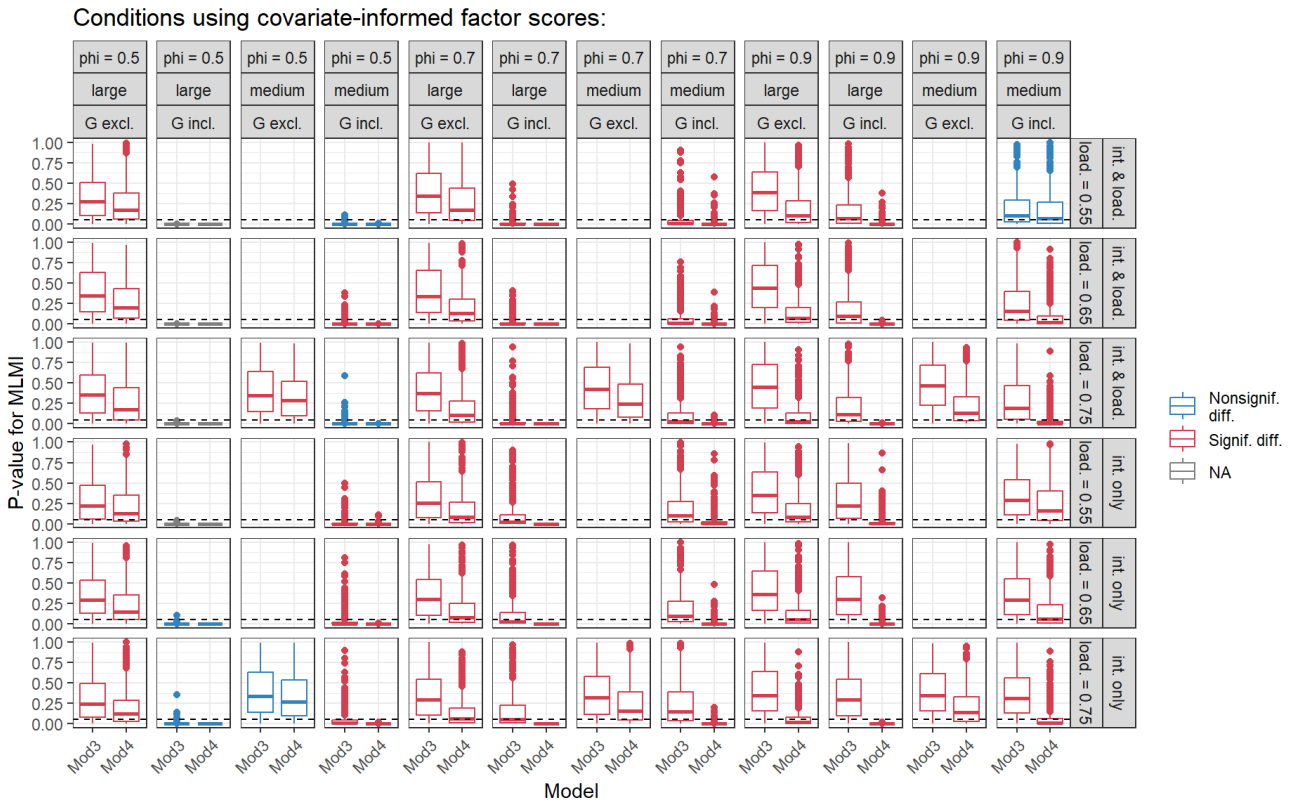
that the proposed technique becomes increasingly more effective (over constraining for separation) with increasing values of φ. When G was excluded as a predictor, Model 3 still consistently had fewer MLMI violations than Model 4, although the differences did not appear as notable. This is likely because when G was excluded, the original model-fitting procedure for separation (Model 4) already did a decent job of satisfying MLMI, such that the proposed technique (Model 3) did not make as large of a difference from separation as when G was included.

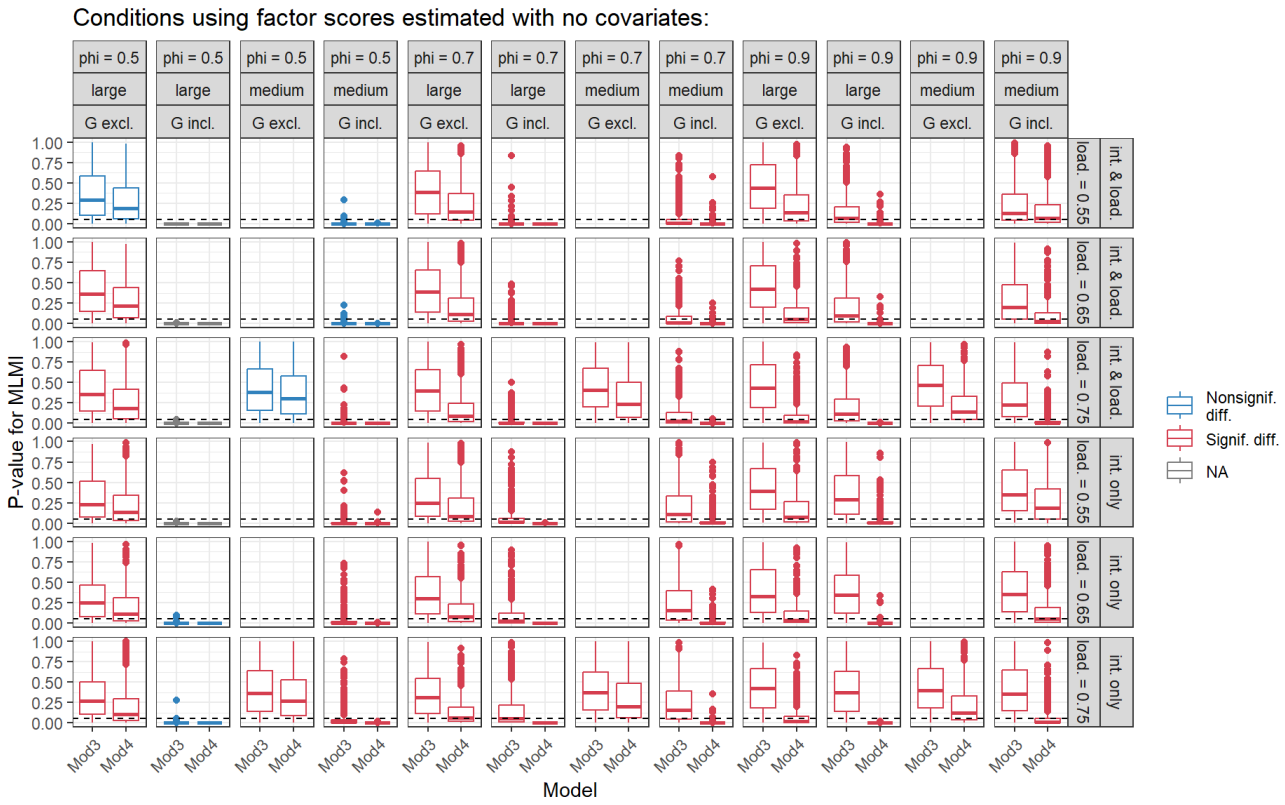
These patterns were observed for both types of non-invariance and for both kinds of factor score estimates. In general, differences in MLMI violation between Models 3 and 4 appeared most notable under higher factor loadings. This is likely because Model 3's ability to meet MLMI improved with increasing factor loadings, and on top of that, Model 4's MLMI violations became more crystallized with increasing reliability of the outcome variable as an indicator of η, as discussed in Section 3.2. Similarly, differences were also more apparent with a large magnitude of non-invariance, given that MLMI violations by Model 4 were also most prevalent under large magnitudes of non-invariance.

Note that in eight of the 120 conditions, McNemar's test and Cohen's g effect size could not be calculated, because there was only one pattern of MLMI results across Models 3 and 4 among all replications (i.e., both Models 3 and 4 violated MLMI in all 500 replications of the condition).

Figure 5

Condition-wise boxplots of the p-values of the fair dummies test for MLMI for Models 3 and 4





Note. A "significant" difference refers to McNemar's $p < \frac{.05}{120}$ and Cohen's $g \geq 0.15$. "NA" refers to conditions with no variability in the MLMI results among all 500 replications.

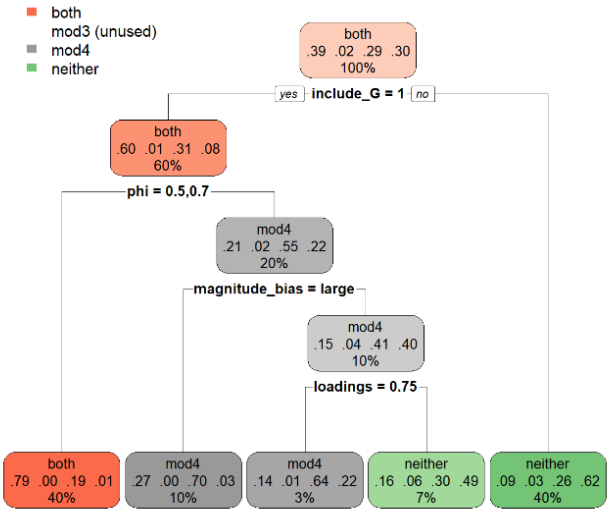
The above observations from Figure 5 can be collectively summarized with results from the CART analysis, with the classification tree diagram presented in Figure 6. The MLMI simulation outcome modeled in this classification tree was originally a four-level categorical variable, but no replications were classified as *Model 3 violated MLMI and Model 4 did not* in the optimal decision rule found. The misclassification rate of this classification tree was 31.0%.

The hierarchy of the splits in Figure 6 suggests that the most useful simulation factor for explaining the improvement in MLMI associated with the proposed technique over separation was the inclusion/exclusion of G, followed by the value of φ. When G was excluded as a predictor, the optimal classification of a replication was *neither model violated MLMI*, meaning both approaches were effective in meeting MLMI in the presence of measurement bias (i.e., no improvement). When G was included, the optimal classification depended on the value of φ.

When $\varphi = 0.5$ or 0.7, a replication was optimally classified as *both models violated MLMI*, meaning neither approach was effective (i.e., no improvement). When $\varphi = 0.9$, the optimal classification was *Model 4 violated MLMI and Model 3 did not*, meaning only the proposed bias mitigation technique was effective (i.e., improvement present). The tree diagram further split this node (G is included and $\varphi = 0.9$), such that only when the magnitude of non-invariance is medium and factor loadings are low ($\lambda = 0.55, 0.65$), the optimal classification was *neither model violated MLMI*. All other conditions from this node were classified as *Model 4 violated MLMI and Model 3 did not.* Because the tree did not split on the type of factor scores or the type of non-invariance, we can interpret that the MLMI results did not differ meaningfully according to whether covariates are used in the calculation of factor scores or whether there is bias in intercepts only or both intercepts and loadings.

Figure 6

Tree diagram of a classification tree modeling the MLMI outcome in Models 3 and 4, fit to simulation results across all conditions



Note. In each node of the tree diagram, the top value represents the majority class, the middle value represents the proportional breakdown of the classes of the observations belonging to that node, and the bottom value represents the percentage of total observations that belong to that node.

Predictive performance

Next, we examined for any predictive performance differences between Models 3 and 4 to understand how performance may be altered with the proposed bias mitigation technique in comparison to the original model-fitting procedure for separation, if at all. Figures 7a and 7b present the condition-wise distributions of the RMSE and $r_{\eta,\hat{Y}}$, respectively, from Models 3 and 4. In terms of the RMSE, we found patterns to differ according to the inclusion/exclusion of G. When G was included, Model 3 performed significantly worse than Model 4 ($p < \frac{.05}{120}$ and $|d| \geq 0.5$), indicating that predictive performance was sacrificed when implementing the proposed strategy. This performance gap was largest under a large magnitude of non-invariance and with higher loadings. This is expected, given that these are the conditions in which we observed the biggest improvement in MLMI satisfaction by implementing the proposed technique over constraining for separation. Therefore, we may anticipate observing the biggest sacrifices in predictive performance where there are the biggest gains in MLMI satisfaction. When G was excluded, we did not observe notable differences between Models 3 and 4, meaning predictive performance was preserved in those conditions.

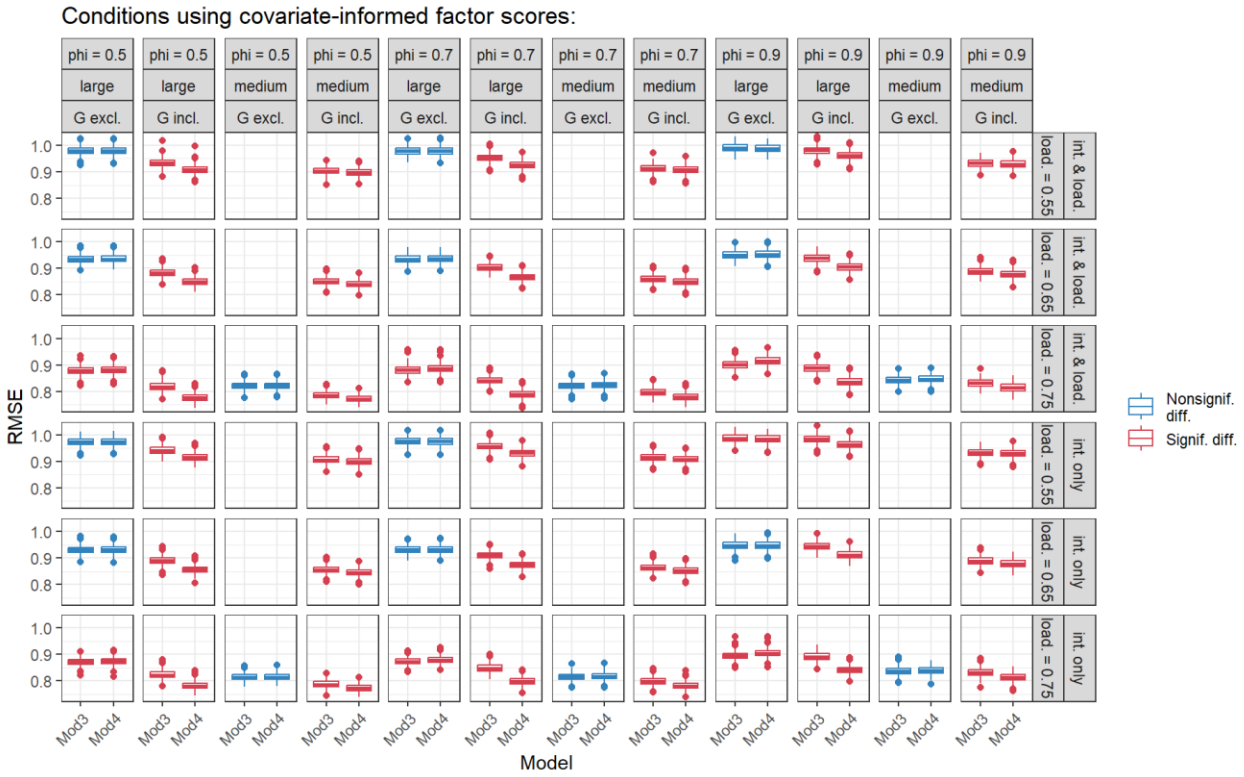
In terms of $r_{\eta,\hat{Y}}$, Model 3 performed significantly *better* than Model 4 ($p < \frac{.05}{120}$ and $|d| \geq 0.5$) in a majority of the conditions, meaning the proposed technique improved performance over the original model-fitting procedure for separation. The exception was in conditions where G was included and $\varphi = 0.9$, where the two models performed comparably to one another. As hypothesized, this is opposite from the pattern observed when examining performance in terms of the RMSE, where the proposed technique led to worse performance than constraining for separation in most conditions. This difference in patterns across the two metrics highlights the distinct natures of the RMSE and $r_{\eta,\hat{Y}}$ as performance indicators. Because the RMSE involves

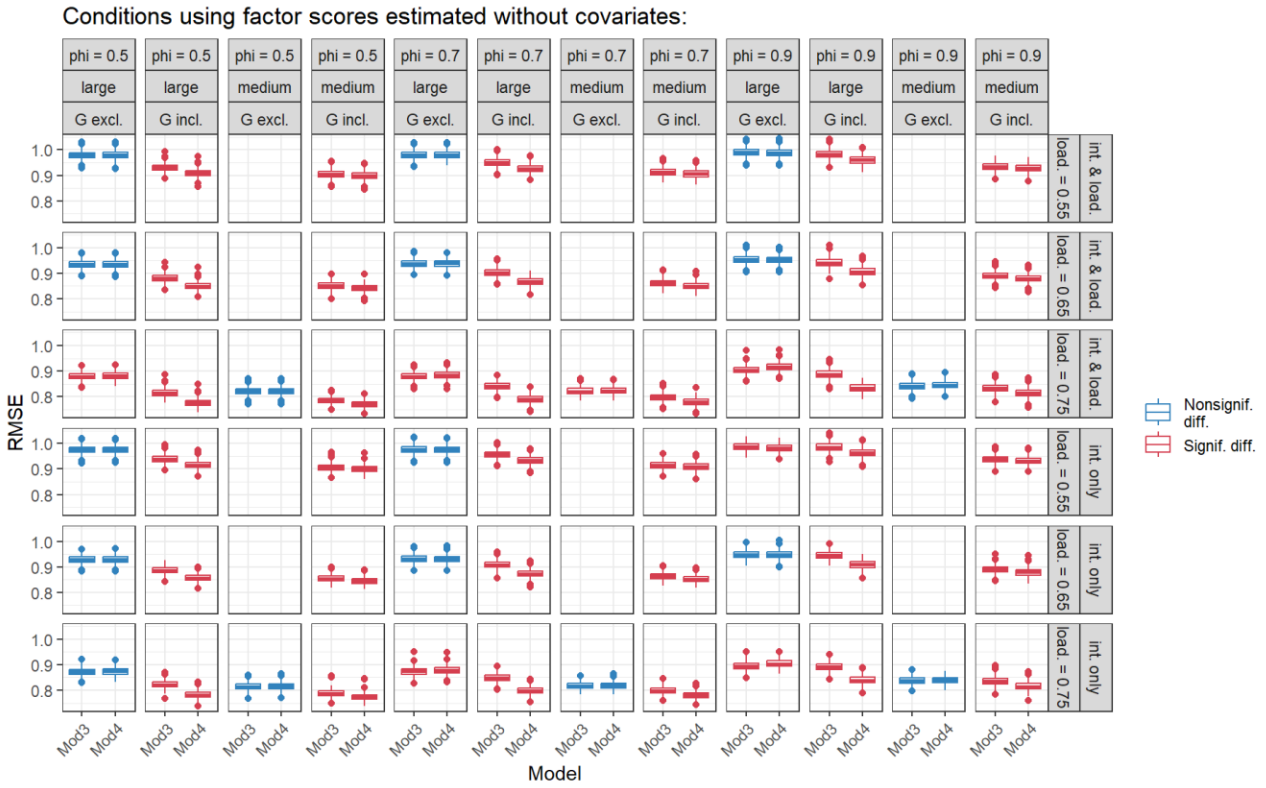
observed variables only in its calculation, it may generally favor a model that regularizes for separation (Model 4), which also involves observed variables only, over Model 3. On the other hand, $r_{\eta,\hat{Y}}$ involves the target latent variable, or the factor, in its calculation, so it may favor a model that regularizes for $\hat{Y} \perp G \mid \hat{\eta}$ (Model 3), which involves the factor score estimates, over Model 4. This means that in attempting to satisfy MLMI by constraining for $\hat{Y} \perp G \mid \hat{\eta}$, the proposed bias mitigation technique generates predictions that better correlate with the target latent variable, but worse at reproducing the observed outcome variable, compared to the original model-fitting procedure for separation.

Figure 7

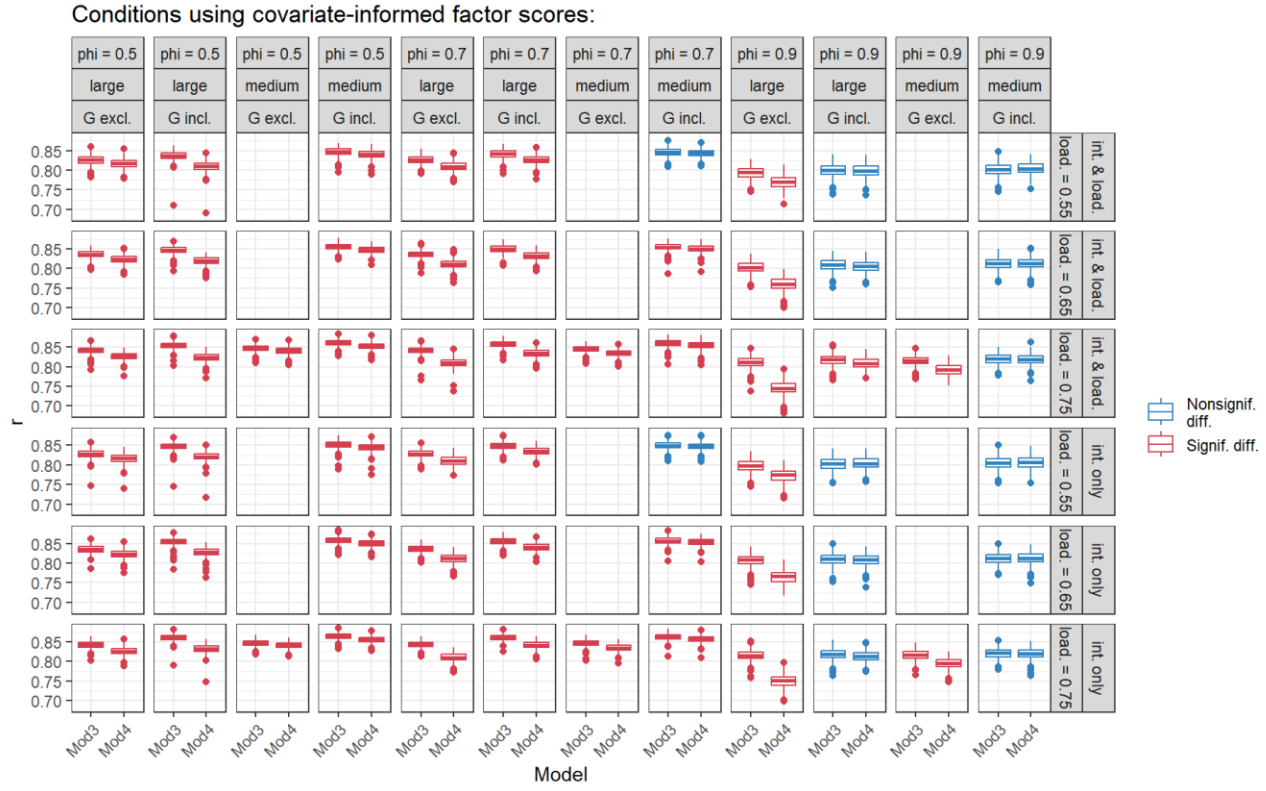
Condition-wise boxplots of the test performance of Models 3 and 4 in terms of RMSE (Panel a) and $r_{\eta,\hat{Y}}$ (Panel b)

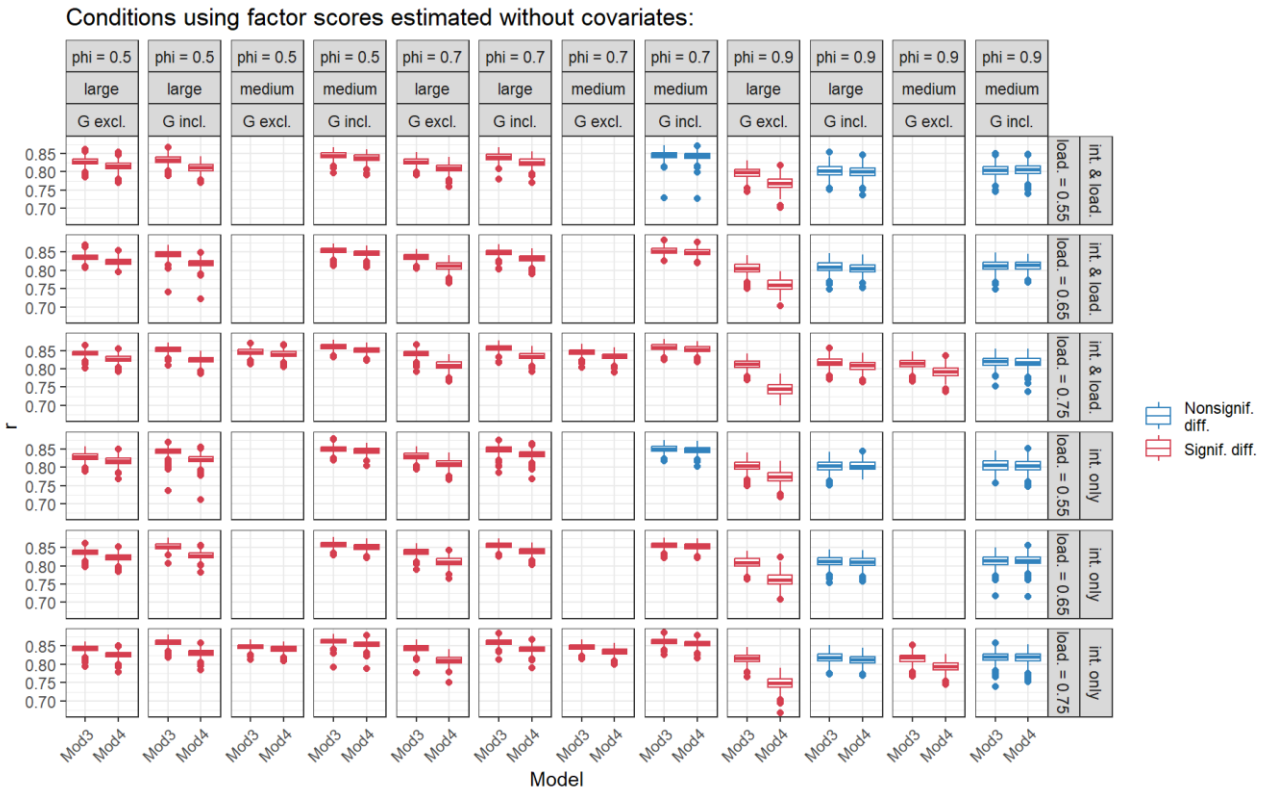
Panel a




Conditions using factor scores estimated without covariates:
RMSE
Model
Nonsignif. diff.
Signif. diff.

Panel b


Conditions using covariate-informed factor scores:
r
Model
Nonsignif. diff.
Signif. diff.



Note. A "significant" difference refers to paired sample t-test $p < \frac{.05}{120}$ and Cohen's $|d| \geq 0.5$.

To understand the systematic influence of the simulation factors on the performance differences between Models 3 and 4, we again used CART models to analyze results across all simulation conditions together. Figures 8a and 8b present the resulting regression tree diagrams for the pairwise (by replication) differences in RMSE and $r_{\eta,\hat{Y}}$, respectively, between Models 3 and 4. Note that in both regression trees, the differences in performance metrics were calculated such that a positive difference corresponds to Model 3 performing better than Model 4, and a negative difference corresponds to Model 3 performing worse than Model 4. The R^2 values of these regression trees were 0.84 and 0.55 for the RMSE and $r_{\eta,\hat{Y}}$ difference, respectively.

From Figure 8a, we can gather that the most important determinant of the RMSE difference between Models 3 and 4 was the inclusion/exclusion of G, followed by the magnitude of non-invariance and value of λ. As observed from Figure 7, the exclusion of G often resulted in

comparable performance between Models 3 and 4. When G was included, the performance drop associated with the proposed technique grew in size with a larger magnitude of non-invariance and with higher factor loadings. Again, these are the same conditions in which we observed the biggest gains in MLMI satisfaction by implementing the proposed technique as opposed to constraining for separation, such that we may expect the biggest sacrifices in RMSE associated with the proposed technique in those conditions.

From Figure 8b, we can gather that the most influential determinant of the difference in $r_{\eta,\hat{Y}}$ between Models 3 and 4 was the magnitude of non-invariance, followed by the inclusion/exclusion of G and the value of φ. First, a larger magnitude of non-invariance led to a larger performance boost in $r_{\eta,\hat{Y}}$ associated with Model 3. Figure 8b also shows that the effect of φ depended on whether G is included as a predictor. When G was excluded, a higher φ was associated with a larger performance gap between Models 3 and 4. This is intuitive, given that a larger φ indicates a greater emphasis on the regularization term (i.e., constraining for $\hat{Y} \perp G \mid \hat{\eta}$ for Model 3 versus separation for Model 4) that distinguishes the two models. On the other hand, when G was included, we found an unintuitive pattern, where a larger φ led to a smaller performance gap between Models 3 and 4. This was observed earlier in Figure 7b, where Model 3 performed significantly better than Model 4 in all conditions except when G was included and $\varphi = 0.9$. A possible explanation for this is that when G is included as a predictor, changing φ has a differential effect on $r_{\eta,\hat{Y}}$ between Models 3 and 4 compared to when G is excluded. When G is included, increasing φ for Model 3 steadily nears it closer to MLMI satisfaction, whereas increasing φ for Model 4 appears to have little impact on nearing it to MLMI satisfaction, because Model 4 violates MLMI even with a strong constraint for separation, such that increasing φ does not affect performance as much for Model 4. Therefore, as φ increases, any

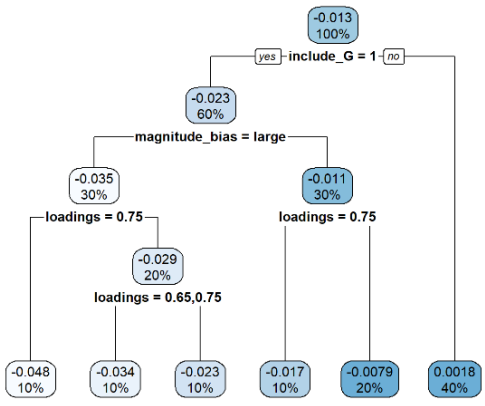
changes in MLMI satisfaction within each model and as a result, differences in $r_{\eta,\hat{Y}}$ between Models 3 and 4 becomes narrower, given that Model 4's performance remains relatively unharmed compared to that of Model 3.

Because neither tree in Figure 8a nor 8b split on the type of factor score estimates or type of non-invariance, we can interpret that performance differences between the two cost functions (in terms of both RMSE and $r_{\eta,\hat{Y}}$) did not differ meaningfully according to whether covariates are used in the calculation of factor scores or whether there is bias in intercepts only or both intercepts and loadings.

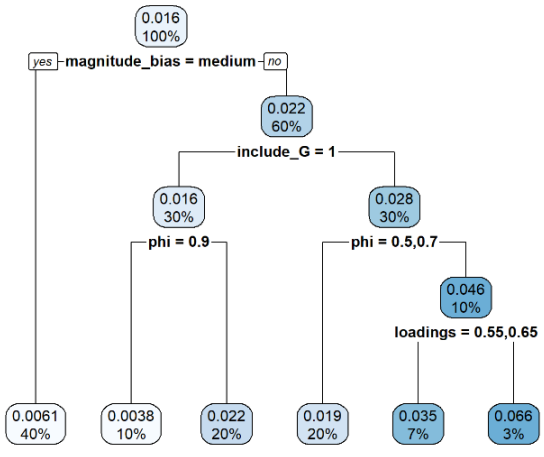
Figure 8

Tree diagram of a regression tree modeling the difference in root mean squared error (Panel a) and $r_{\eta,\hat{Y}}$ (Panel b) between Models 3 and 4, fit to simulation results across all conditions

Panel a



Panel b



Note. In each node of the tree diagram, the top value represents the mean value of the simulation outcome among observations belonging to that node, and the bottom value represents the percentage of total observations that belong to that node.

Summary of simulation part 2

In simulation part 1, we found several conditions with substantial (over 25% of replications) MLMI violations associated with the use of a non-invariant outcome variable, even when the machine learning model was constrained for separation during training. We explored the effectiveness of a possible bias mitigation technique to address this effect of measurement bias on MLMI in simulation part 2, which showed that under certain conditions, MLMI can be met even in the presence of measurement bias using this technique.

First, if G is to be excluded as a predictor, MLMI was often successfully met using the proposed bias mitigation technique, regardless of the type or magnitude of non-invariance, value of λ or φ, and the type of factor score estimates used. However, the added utility of the proposed technique over the original model-fitting procedure for separation may be maximized when G is to be *included* as a predictor, which is where MLMI violations associated with a non-invariant outcome variable were most salient (see simulation part 1; Section 3.2). When G is included and a strong emphasis is placed on regularizing for $\hat{Y} \perp G \mid \hat{\eta}$ ($\varphi = 0.9$) during training, the proposed bias mitigation technique was largely effective in satisfying MLMI, where constraining for separation was ineffective. In such conditions, the proposed technique was effective for both medium and large magnitudes of non-invariance, for both types of non-invariance, and for both types of factor score estimates. However, the caveat to meeting MLMI using the proposed technique is that predictive performance in terms of RMSE was negatively impacted, although performance in terms of $r_{\eta,\hat{Y}}$ was preserved, if not improved. Therefore, there is a trade-off to be made in generating predictions that satisfy MLMI versus those that accurately reproduce the observed outcome variable. With a weaker emphasis on regularizing for $\hat{Y} \perp G \mid \hat{\eta}$ ($\varphi = 0.5$ or 0.7), MLMI violations were still frequently prevalent, although to a lesser extent than when

constraining for separation. Therefore, while constraining for $\hat{Y} \perp G \mid \hat{\eta}$ provided a relative improvement over separation across all values of φ tested, a strong emphasis ($\varphi = 0.9$) was needed for the proposed technique to be effective in the absolute sense, if G is to be included as a predictor.

In sum, the proposed technique can be an effective treatment for mitigating the effects of measurement bias on the fairness of machine learning models under certain conditions where constraining for separation is unsuccessful in meeting MLMI. Further, the predictions obtained from the proposed technique often showed a higher correlation with the target latent variable, compared to predictions obtained from the original model-fitting procedure for separation. However, these improvements in MLMI satisfaction and $r_{\eta,\hat{Y}}$ were qualified by an increase in RMSE. Given that the observed outcome variable exhibits measurement bias and should therefore not be regarded as the "ground truth", this sacrifice in accurately reproducing the observed outcome variable may not be so detrimental.

CHAPTER 5: APPLIED EXAMPLE

To demonstrate the application of the proposed methods on empirical data, we analyzed the public use files of the Medical Expenditure Panel Survey (MEPS). The household component of this nationally representative, large-scale survey, given by the Agency for Healthcare Research and Quality (AHRQ), collects data from sampled U.S. households and their members about their use of health services, health conditions and status, medical expenditures, sources of payment, health insurance coverage, access to care, employment, income, and demographic characteristics (AHRQ, 2019). MEPS has been previously used in the machine learning fairness literature to emulate a use case where the task is to develop a "fair" machine learning model for healthcare utilization or expense prediction that scores patients to aid in care management prioritization and enrollment decisions (e.g., Bellamy et al., 2019; Fabris et al., 2022; Romano et al., 2020; Singh & Ramamurthy, 2019). Future healthcare utilization and expense have been used as outcome variables in algorithms in healthcare systems as proxies or indicators of health to prospectively identify those with high projected health needs or risk (Fleishman & Cohen, 2010; Morid et al., 2017; Obermeyer et al., 2019; Rakovski et al., 2002; Singh & Ramamurthy, 2019; Wherry et al., 2014). It should be noted that while MEPS is not the data source that was used in the motivating *Science* article of Obermeyer et al. (2019) from Chapter 1, it provides a similar example.

In MEPS, a new panel or cohort is initiated each year and is surveyed over five rounds of interviews spanning two calendar years. For example, Panel 22 was surveyed in 2017 for rounds 1, 2, and 3 and in 2018 for rounds 3, 4, and 5. We followed the methodology of Fleishman and

Cohen (2010) and Singh and Ramamurthy (2019) in using a panel's data collected from the first year to predict their medical expenses in the second year, given that in a realistic use case, the prediction of future, rather than current, medical expenses is often of interest. Specifically, we used the MEPS full-year consolidated data files of 2017 and 2018 (AHRQ, 2019; AHRQ, 2020) for Panel 22 to build a machine learning model, using health characteristics and demographic variables collected in 2017 to predict expenses in 2018.

As such, using the MEPS data, the emulated task was to build a "fair" medical expense prediction model with respect to race/ethnicity, where we took a measurement approach to machine learning fairness, using the framework and bias mitigation technique proposed in this thesis. The grouping variable used was a binary indicator of race/ethnicity, with the two levels being non-Hispanic white (coded 1) and non-Hispanic black (coded 0). All other race/ethnicity categories were excluded from the analysis. To infer a target latent variable of health needs, we collected additional indicators of health needs, besides medical expenses, that were captured in the survey to be considered as proxy outcome variables. Then, we used the proposed framework and bias mitigation technique respectively to test for measurement invariance and build a medical expense prediction model that takes fairness into consideration from a measurement perspective.

Section 5.1: Measurement Invariance Testing using the Proposed Framework

It is possible that the medical expense measure exhibits measurement bias with respect to race. In other words, medical expenses may indicate a differential level of underlying health needs across black and white individuals. Applying the proposed framework to test for measurement invariance, we gathered multiple other indicators of health needs available in the MEPS data to fit a MG-CFA model. These additional indicators included three patient-reported

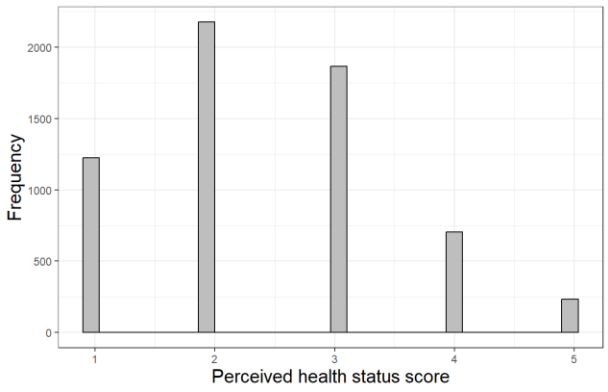
outcome measures: a measure of self-assessed perceived health, the health in general score of the Veterans RAND 12-Item Health Survey (VR-12), and another item from the VR-12 on the extent to which one's physical health limit the kind of work or daily activities they are able to do ("daily limitations" hereafter). Medical expense was calculated as the sum of payments made (including out-of-pocket payments and payments made by insurance) for various health services, including office-based care, hospital-based care (e.g., inpatient stays, outpatient visits, emergency room visits), home health care, dental and vision care, and prescribed medications. Perceived health status and general health were rated on a scale of 1 (excellent) to 5 (poor). Daily limitations were rated on a scale of 1 (none of the time) to 5 (all of the time). Because the health in general score and daily limitations score came from the self-administered questionnaire (SAQ) portion of the MEPS, which was only administered to respondents 18 and older, the analysis was limited to those who were eligible for the SAQ. For simplicity, we additionally limited the analysis to those with a non-zero medical expenditure. This resulted in an unweighted analytic sample size of $N = 6{,}232$ ($N = 5{,}021$ for non-Hispanic whites; $N = 1{,}211$ for non-Hispanic blacks). All four measures were treated as continuous in the analysis.

First, we examined the distributions of each measure, plotted in Figure 9. The medical expense measure was log-transformed to bring the distribution closer to normality and the scale closer to the rest of the measures, which are on a scale of 1 to 5 (references to medical expenses hereafter assume log-transformation). Due to the skewness seen in the distributions, particularly the daily limitations measure, we used maximum likelihood estimation with robust standard errors, test statistics, and fit indices. Missing data (0.42%, 3.23%, 3.67%, 0% missingness for perceived health status, health in general, daily limitations, and medical expenses, respectively) were handled with full-information maximum likelihood.

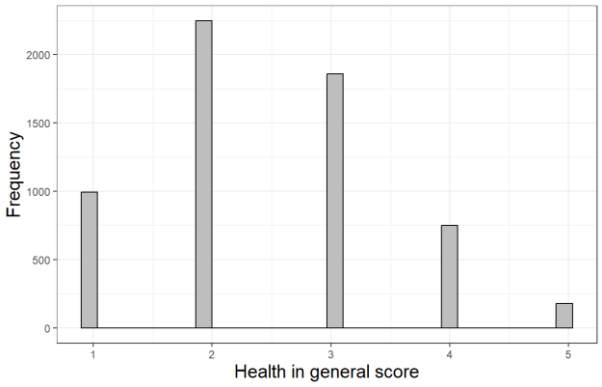
Figure 9

Histograms of the four measures, including perceived health status (Panel a), health in general (Panel b), daily limitations (Panel c), and medical expenses (Panel d)

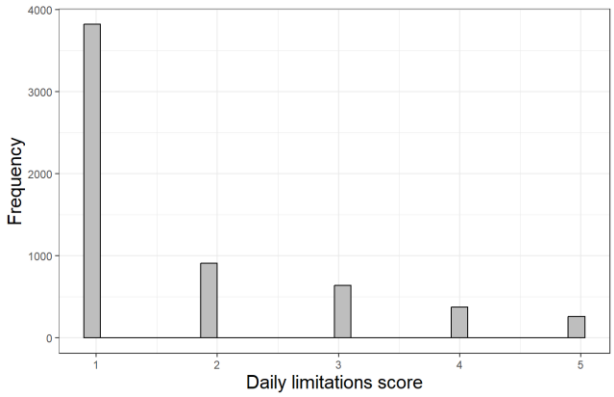
Panel a



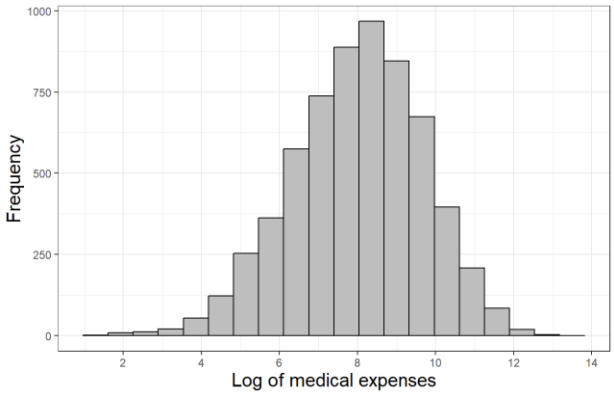
Panel b



Panel c



Panel d



We began the analysis by fitting a single-group CFA model separately in each of the two groups. For the non-Hispanic white group, a one-factor CFA model had suboptimal fit to the data ($\chi^2(2) = 165.11, p < .01; CFI = 0.97;\ TLI = 0.92;\ RMSEA = 0.13, 95\%\ CI: [0.11, 0.15]; SRMR = 0.03$). Examining the top modification indices led us to estimate a residual covariance between the perceived health status and health in general measures. Given that these measures are similar in content, we found this modification to be theoretically defensible. Therefore, we next estimated a one-factor CFA model with a residual

covariance and found this model to fit well to the data ($\chi^2(1) = 11.73, p < .01; CFI = 1.00; TLI = 0.99; RMSEA = 0.05, 95\%\ CI: [0.03, 0.08]; SRMR = 0.01$). A scaled chi-square likelihood ratio test also indicated a significant improvement in model fit following this modification ($\chi^2(1) = 160.88, p < .01$). Repeating the analysis on the non-Hispanic black group also led to a one-factor CFA model with a residual covariance between the perceived health status and health in general measures to be optimally fitting ($\chi^2(1) = 6.47, p = .01; CFI = 1.00; TLI = 0.98; RMSEA = 0.07, 95\%\ CI: [0.03, 0.12]; SRMR = 0.01$).

Next, we estimated this one-factor model in both groups simultaneously using a MG-CFA, starting with a configural model with an equal general factor structure across groups. The model was identified by setting the mean and variance of the factor to 0 and 1, respectively, for the non-Hispanic white group and freely estimating them for the non-Hispanic black group, and freely estimating but setting across-group equality constraints on the intercept and factor loading of the perceived health status measure. This configural model showed adequate fit to the data, so we moved on to test for weak invariance by constraining all factor loadings to be equal. This model also showed adequate fit to the data, and a non-significant chi-square likelihood ratio test indicated no decrement in fit following these constraints ($\chi^2(3) = 4.53, p = .21$). This indicates that weak invariance is met. Next, we tested for strong invariance by placing across-group equality constraints on all factor loadings and intercepts. This led to a significant decrement in model fit according to a chi-square likelihood ratio test ($\chi^2(3) = 167.41, p < .01$). After examining the top modification indices, we lifted the across-group equality constraint on the intercept of the medical expense measure. This partial strong invariance model fit significantly better than the strong invariance model ($\chi^2(1) = 129.42, p < .01$), but it still fit significantly worse than the weak invariance model ($\chi^2(2) = 56.89, p < .01$). Therefore, we examined the

top modification indices of the partial strong invariance model and additionally lifted the across-group equality constraint on the intercept of the daily limitations measure. This led to a significant improvement in model fit over the previous partial strong invariance model ($\chi^2(1) = 55.78, p < .01$), and it further showed no decrement in fit over the weak invariance model ($\chi^2(1) = 0.04, p = .85$). Therefore, we conclude that partial strong invariance is met, with two measures' intercept parameters being freely estimated with no across-equality constraints. The parameter estimates of this final partial strong invariance model are given in Table 4.

Table 4

Parameter estimates of partial strong invariance MG-CFA model

	Group 0 (non-Hispanic black)				**Group 1 (non-Hispanic white)**			
	Estimate	Std. error	P-value	Std. estimate	Estimate	Std. error	P-value	Std. estimate
Factor loadings								
Perceived health status	0.66	0.02	<.01	0.63	0.66	0.02	<.01	0.64
Daily limitations	0.93	0.02	<.01	0.79	0.93	0.02	<.01	0.82
Health in general	0.72	0.02	<.01	0.70	0.72	0.02	<.01	0.73
Medical expenses	0.82	0.02	<.01	0.45	0.82	0.02	<.01	0.50
Item intercepts								
Perceived health status	2.40	0.01	<.01	2.28	2.40	0.01	<.01	2.30
Daily limitations	1.43	0.04	<.01	1.21	1.71	0.02	<.01	1.51
Health in general	2.43	0.01	<.01	2.36	2.43	0.01	<.01	2.45
Medical expenses	7.39	0.05	<.01	4.09	8.06	0.02	<.01	4.92
Item residual (co)variances								
Perceived health status	0.67	0.04	<.01	0.60	0.64	0.02	<.01	0.59
Daily limitations	0.53	0.05	<.01	0.38	0.41	0.03	<.01	0.32
Health in general	0.53	0.03	<.01	0.50	0.46	0.02	<.01	0.47
Medical expenses	2.60	0.13	<.01	0.80	2.02	0.05	<.01	0.75
Perceived health status ~~ Health in general	0.23	0.03	<.01	0.38	0.25	0.02	<.01	0.46
Factor mean	0.38	0.03	<.01	0.38	0.00			0.00
Factor variance	1.00	0.06	<.01	1.00	1.00			1.00

Note. ~~ indicates a covariance between two item residuals.

While Table 4 shows that the standardized factor loading of the medical expense measure is rather low ($\lambda = 0.50$) and it therefore may lack in utility as a reliable indicator of health needs,

we proceeded to use this as the outcome variable of the machine learning task for demonstration purposes.

Finally, all fit measures from the series of MG-CFA models are summarized in Table 5, as well as results from all chi-square likelihood ratio tests comparing the relative fit of each subsequent model.

Table 5

Results of measurement invariance testing, including fit statistics and chi-square likelihood ratio tests of the sequence of MG-CFA models

MG-CFA model	Absolute fit						Relative fit			
	χ^2	df	CFI	TLI	RMSEA	SRMR	Rel. to model	$\Delta\chi^2$	Δdf	p-value
1. Configural	18.38	2	1.00	0.99	0.05 [0.03, 0.08]	0.01				
2. Weak	22.80	5	1.00	0.99	0.04 [0.02, 0.05]	0.01	1	4.53	3	.21
3. Strong	177.70	8	0.98	0.97	0.08 [0.07, 0.10]	0.03	2	167.41	3	< .01
4. Partial strong (expense)	78.23	7	0.99	0.98	0.06 [0.05, 0.07]	0.02	3	129.42	1	< .01
							2	56.89	2	< .01
5. Partial strong (expense, daily limitations)	23.07	6	1.00	1.00	0.03 [0.02, 0.05]	0.01	4	55.78	1	< .01
							2	0.04	1	.85

Note. df = degrees of freedom; CFI = comparative fit index; TLI = Tucker-Lewis Index; RMSEA = root mean squared error of approximation; SRMR = standardized root mean squared residual.

Section 5.2: Meeting MLMI using the Proposed Bias Mitigation Technique

We found evidence of measurement bias in our chosen observed outcome variable, the medical expense measure, where its intercept parameter differed across non-Hispanic whites and non-Hispanic blacks. This means that the same level of underlying health needs is manifested as differential levels of expenses according to race. As we learned in simulation part 1, using this non-invariant measure as the outcome variable of the machine learning model, without accounting for this bias, can lead to an "unfair" model that violates MLMI (i.e., machine learning

model gives individuals with the same underlying health needs differing predictions according to race). Therefore, we used factor score estimates calculated from the partial strong invariance model presented in Table 4 to implement the proposed bias mitigation technique, in an attempt to build a "fair" model satisfying MLMI in the presence of measurement bias in the outcome variable. We used factor scores computed without covariates, given that we did not observe notable improvements in the effectiveness of the proposed bias mitigation technique associated with the use of covariate-informed factor scores in simulation part 2 (see Section 4.2), and we do not have strong *a priori* knowledge of the structural relationships between the predictors of the machine learning model and health needs.

We split the MEPS data into a training (60%) and test (40%) set. Predictors of the machine learning model included age, marital status, military active-duty status, diagnoses of priority conditions (e.g., high blood pressure, diabetes, high cholesterol, asthma), physical limitations (e.g., hearing, vision), health behaviors (e.g., smoking), and insurance coverage. Missing values in the predictors (0.4% missing) were imputed using the R package and function missForest (Stekhoven, 2013). There were 29 predictors in total, including race.

Table 6 summarizes the test-set performance of the various machine learning models we built, which differed in their values of φ ($\varphi = 0.9, 0.7, 0.5$), whether race was included as a predictor in the machine learning model, and the fairness constraint used (model-fitting procedure for separation of Romano et al., 2020 or proposed bias mitigation technique). We used the same set of hyperparameter values as used in the simulation study. Table 6 shows that performance differed only negligibly by changing these factors. Therefore, it may be sound to select a model that is most favorable with respect to fairness (i.e., most likely to satisfy MLMI),

which would be the model trained using the proposed bias mitigation technique, with a strong emphasis on the regularization term ($\varphi = 0.9$), and excluding race as a predictor.

Table 6

Predictive performance of the medical expense prediction model

Fairness constraint	Predictive performance (RMSE)					
	$\varphi = 0.9$		$\varphi = 0.7$		$\varphi = 0.5$	
	Race included	Race excluded	Race included	Race excluded	Race included	Race excluded
Proposed technique	1.46	1.48	1.46	1.46	1.46	1.46
Separation	1.45	1.45	1.45	1.45	1.46	1.46

CHAPTER 6: DISCUSSION

As machine learning continues to govern sensitive and high-stakes decisions in everyday life, studying machine learning fairness remains critical in ensuring and advancing the equity of machine learning-based decisions. In this thesis, we presented how the psychometric literature on test fairness contributes a useful perspective to machine learning fairness, focusing on how *bias in measurement* channels into *bias in machine learning*.

First, in taking a measurement perspective to the problem of machine learning bias, we emphasized the idea of a target latent variable, or the true, underlying construct to be predicted in the machine learning task—an entity to be distinguished from the observed outcome variable, which is often only an imperfect proxy of it. To ensure that the observed outcome variable is an unbiased measure of the target latent variable, we proposed a simple framework to conduct measurement invariance testing among multiple, candidate observed outcome variables in machine learning. We also introduced the concept of MLMI, or the conditional independence of machine learning predictions and group membership given the target latent variable—a definition of machine learning fairness that is analogous to the psychometric concept of measurement invariance.

We demonstrated the importance of assessing for measurement invariance using this proposed framework in simulation part 1, which showed that training a machine learning model on a non-invariant outcome variable often leads to "unfair" predictions that violate MLMI. In other words, the model gives unequal predictions to observations according to group membership, despite them having equal levels of the target latent variable. Furthermore, with

growing magnitudes of non-invariance, violations of MLMI were not entirely mitigated by constraining the model for separation (the observed-variable counterpart to MLMI) or by removing the grouping variable as a predictor from the model. Because these effects of measurement bias on machine learning bias are not entirely addressed with these techniques, we emphasize the added utility and importance of assessing for measurement invariance in machine learning beyond these existing fairness considerations.

To counter the effects of measurement bias on machine learning bias, we proposed a bias mitigation technique, a natural byproduct of the proposed framework, to train "fair" models in the presence of measurement bias. We demonstrated the effectiveness of this proposed technique in simulation part 2, which showed that MLMI can be satisfied even with a non-invariant outcome variable and that MLMI was better met using the proposed technique than an existing technique constraining for separation. While these improvements in MLMI satisfaction were accompanied by inflations in RMSE (i.e., worse reproduction of the observed outcome variable), we view this trade-off to be warranted, given that the observed outcome variable in this case is non-invariant and therefore, not necessarily the "truth". This may lead one to wonder why we would not simply use a different, invariant outcome variable or alternatively, predict the factor score estimates directly as the outcome variable. Regarding the former point, we view the proposed bias mitigation technique to be useful in cases where there is hesitation to change the outcome variable for practical reasons (see Section 4.1). Regarding the latter point, predicting the observed outcome variable provides enhanced interpretability of model predictions compared to predicting factor score estimates, as the units of factor scores are uninterpretable. Given that model predictions are the main output of interest that machine learning users take away from the

model (i.e., users are interested in making decisions based on model predictions), we view this advantage to be an essential aspect of the proposed technique.

There are several limitations associated with this thesis that lend themselves to useful future directions. First, the simulation study applies only Romano et al. (2020)'s bias mitigation method to train models satisfying separation, and the tests of MLMI as well as the proposed bias mitigation technique are both modified versions of Romano et al. (2020)'s methods leveraging the fair dummy. While there exist many other bias mitigation methods in the machine learning fairness literature, we used Romano et al. (2020)'s technique throughout this thesis because these methods provide a unified framework to both train and test models for separation, and they are flexibly adaptable to the modified applications used in this thesis (e.g., testing for MLMI instead of separation; modifying the cost function to use factor score estimates in the regularization term). However, future studies may benefit from a broader consideration of other existing methods in the machine learning fairness literature and how they may be adapted to be used to study fairness from a measurement perspective.

Second, the simulation set-up and conditions considered in the study are rather limited and may be overly simplistic. For example, the data were simulated such that groups differ only in their mean values of the predictor variables (with up to a medium effect size) and not in the (co)variances, and the predictor variables define the mean of the target latent variable, but not the variance. Furthermore, we have only considered the case in which the grouping variable is a single binary variable, and there is measurement bias only with respect to that grouping variable and no other predictors. Given these narrow specifications, an important extension of this thesis from a simulation perspective would be to consider more complex data generating mechanisms and measurement bias configurations. From a methodological development perspective, a useful

extension of the proposed methods is to expand the proposed bias mitigation technique to be applicable for cases where the grouping variable is continuous or has more than two groups. In the psychometric literature, there exist methods that provide enhanced flexibility over the MG-CFA model in the evaluation of measurement invariance, such as assessing invariance with respect to a grouping variable with many groups, allowing for both continuous and categorical grouping variables, and simultaneously evaluating more than one grouping variable (Asparouhov & Muthén, 2014; Bauer, 2017). Similarly, there exist bias mitigation methods in the machine learning fairness literature which provide similar flexibility, such as methods to train "fair" machine learning models with respect to multiple grouping variables or allow for both continuous and categorical grouping variables (Chakraborty et al., 2021; Zhang et al., 2018). A creative integration of such methods from the psychometric and machine learning fairness literatures would be an interesting future direction.

Third, the findings from the simulation study may be limited to the quality of the predictions generated by the machine learning models in this simulation. With increasing predictive ability of the models such that the $\hat{Y}$'s become closer to the Y's, we may find results to differ. As such, we warn against overinterpretation of the results obtained particularly from simulation conditions with low factor loadings, where unexpected patterns emerged, as further investigation into the replicability of such unexpected patterns may be needed. We nevertheless chose to test these values of factor loadings ($\lambda = 0.55, 0.65, 0.75$) in the simulation, given that we anticipate encountering indicators (observed outcome variables) with factor loadings in this range in practice. To heighten predictive ability in other ways, future simulations may benefit from building more complex machine learning models (e.g., a deeper FNN) or conducting a

thorough hyperparameter (e.g., number of epochs, learning rate) tuning process during training, although it will be more computationally expensive.

From a practical standpoint, another limitation of the proposed methods is the conceivability of having multiple indicators of the target latent variable, such that one could overidentify the MG-CFA model—the core component of the proposed methods. We recognize that in many practical settings, the observed outcome variable used in the machine learning model may be the only indicator of the target latent variable readily available to the analyst. Nevertheless, this limitation should underscore that a single proxy or indicator is not enough to infer an unobserved, underlying construct and to take a measurement perspective to the problem of machine learning bias. We therefore hope that the proposed methods can bring awareness to this risk and at the least facilitate the consideration of collecting more data to allow access to multiple indicators of the target latent variable, rather than relying on a single proxy outcome variable.

With these limitations in mind, we emphasize that while the methods proposed in this thesis may be presented and studied under rather narrow specifications in the simulation, the larger (and arguably more meaningful) contribution lies in the novelty of the integration of concrete, psychometric techniques into the study of machine learning fairness. Building upon the connection between predictive invariance testing and machine learning fairness drawn in Hutchinson and Mitchell (2019), we presented a psychometric addition to the toolbox with which researchers can identify and address sources of machine learning bias. We believe measurement is a key perspective to machine learning fairness that the field of psychometrics is uniquely equipped to contribute, given its closely paralleled work in test fairness and foundation in measurement theory. We therefore envision future work to continue to emphasize the importance

and relevance of *fairness in measurement* for *fairness in machine learning*, build upon the ties between test bias and machine learning bias, and elucidate how these two domains may further intersect or learn from one another.

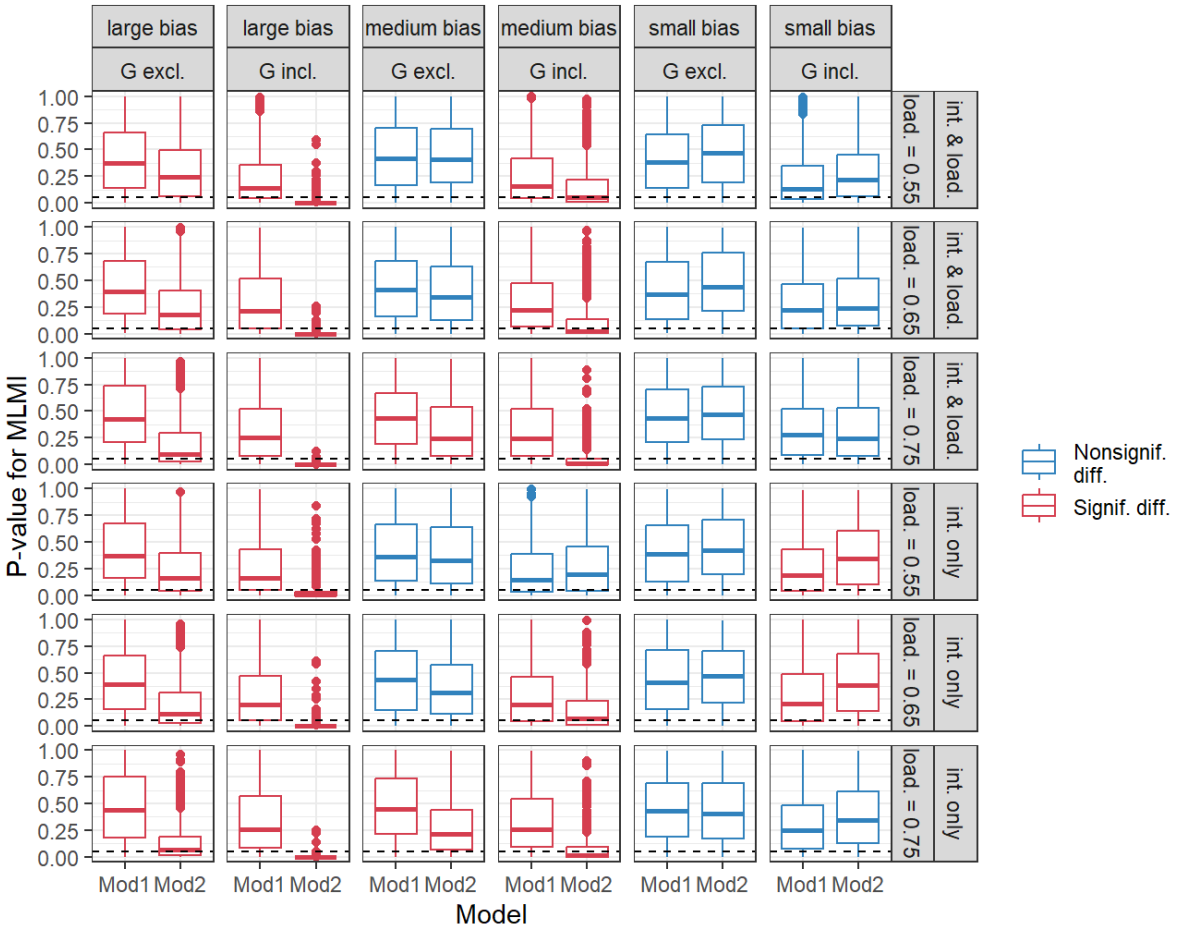
APPENDIX

Below, we present results from a supplemental simulation study, in which the group proportions are imbalanced. This was done by simulating the binary grouping variable as $G_i \sim \text{Bernoulli}(0.75)$ instead of $G_i \sim \text{Bernoulli}(0.5)$, as in the main study's simulation. All other aspects of this supplemental simulation study remained the same as the main study's simulation. See Table 3 for reference to the details of Models 1-4.

Supplemental Simulation Part 1

Figure A1

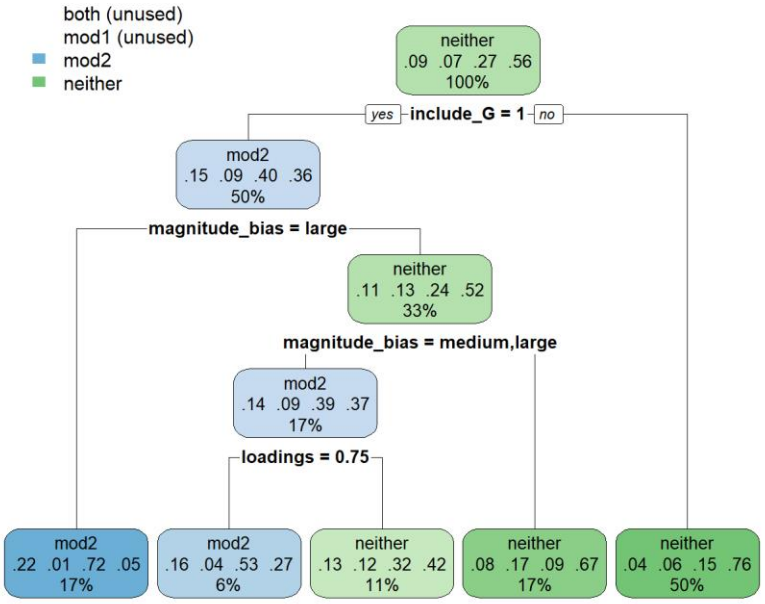
Condition-wise boxplots of the p-values of the fair dummies test for MLMI for Models 1 and 2



Note. A "significant" difference refers to McNemar's $p < \frac{.05}{36}$ and Cohen's $g \geq 0.15$.

Figure A2

Tree diagram of a classification tree modeling the MLMI outcome in Models 1 and 2, fit to simulation results across all conditions

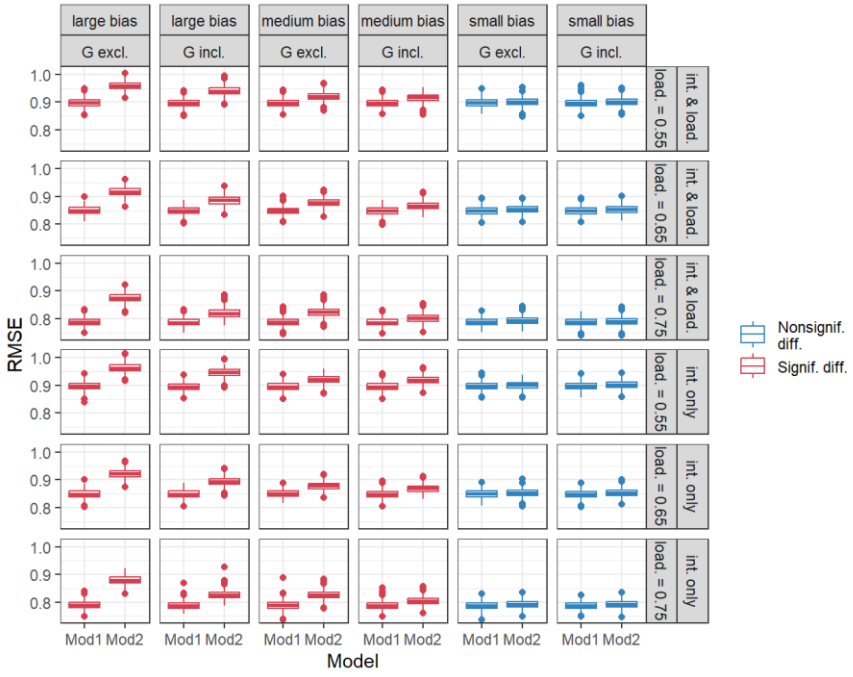


Note. In each node of the tree diagram, the top value represents the majority class, the middle value represents the proportional breakdown of the classes of the observations belonging to that node, and the bottom value represents the percentage of total observations that belong to that node.

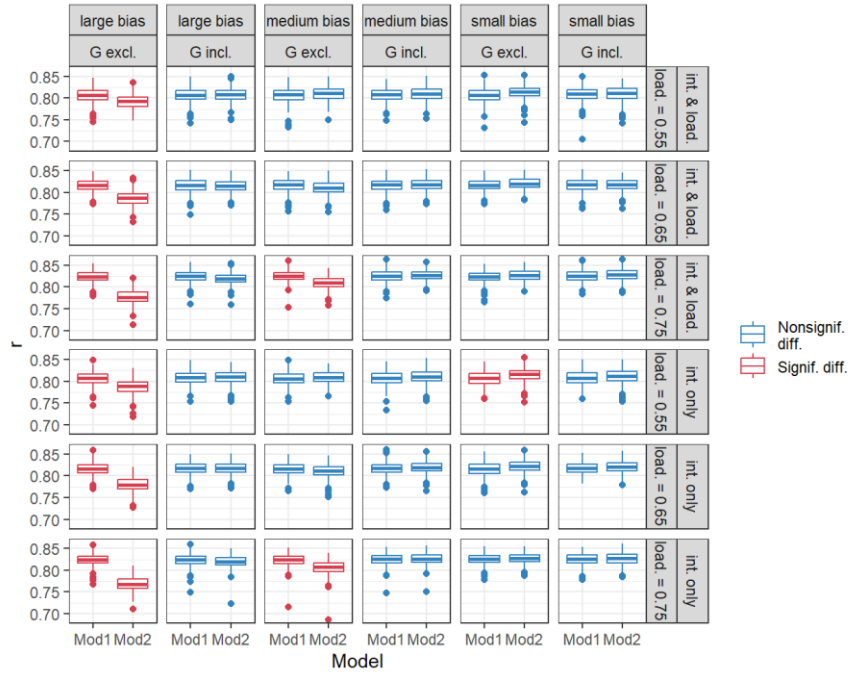
Figure A3

Condition-wise boxplots of the test performance of Models 1 and 2 in terms of RMSE (Panel a) and $r_{\eta,\hat{Y}}$ (Panel b)

Panel a



Panel b

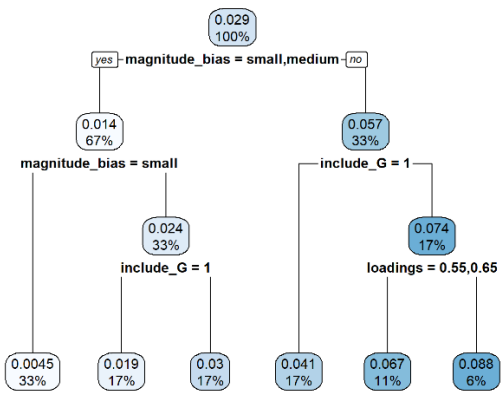


Note. A "significant" difference refers to paired sample t-test $p < \frac{.05}{36}$ and Cohen's $|d| \geq 0.5$.

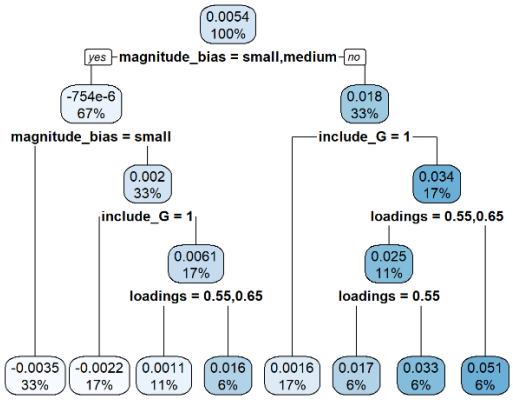
Figure A4

Tree diagram of a regression tree modeling the difference in RMSE (Panel a) and $r_{\eta,\hat{Y}}$ (Panel b) between Models 1 and 2, fit to simulation results across all conditions

Panel a



Panel b



Supplemental Simulation Part 2

Of the 36 total conditions tested in supplemental simulation part 1, there were $c = 17$ conditions where over 25% of its replications resulted in MLMI violation by Model 2 (non-invariant outcome variable), even when constrained for separation. This included all conditions with large magnitudes of bias, except for when G was excluded, $\lambda = 0.55$, and the type of bias is both intercept and loading (11 conditions), plus all conditions with medium magnitudes of bias with G included as a predictor (6 conditions). As done in the main study's simulation, we crossed these 17 conditions with two new simulation factors (value of φ and the use of covariate-informed factor scores), which resulted in 102 total conditions tested in supplemental simulation part two.

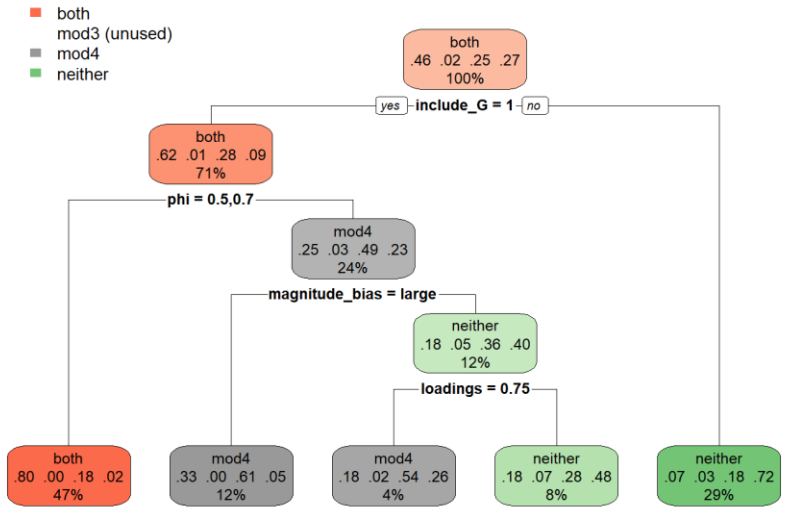
Figure A5

Condition-wise boxplots of the p-values of the fair dummies test for MLMI for Models 3 and 4

Note. A "significant" difference refers to McNemar's $p < \frac{.05}{102}$ and Cohen's $g \geq 0.15$. "NA" refers to conditions with no variability in the MLMI results among all 500 replications.

Figure A6

Tree diagram of a classification tree modeling the MLMI outcome in Models 3 and 4, fit to simulation results across all conditions

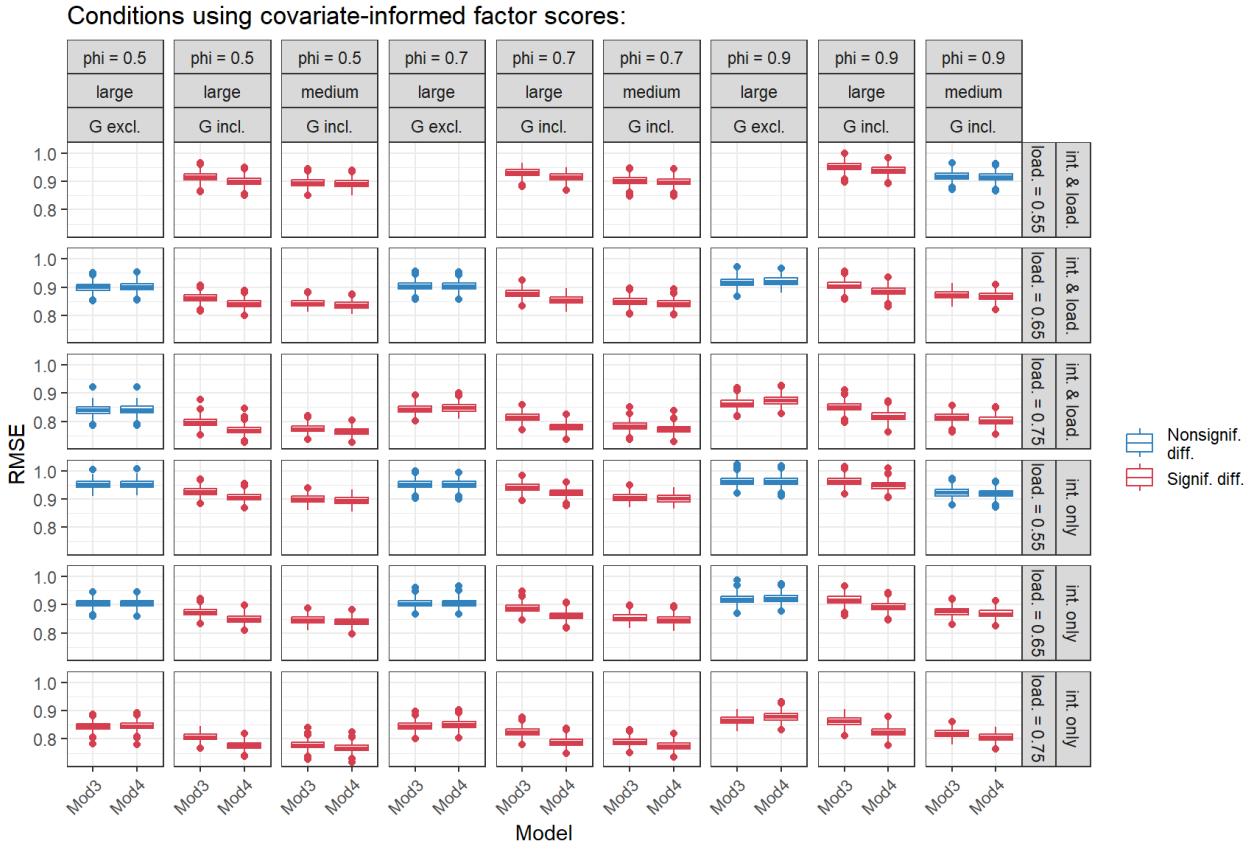


Note. In each node of the tree diagram, the top value represents the majority class, the middle value represents the proportional breakdown of the classes of the observations belonging to that node, and the bottom value represents the percentage of total observations that belong to that node.

Figure A7

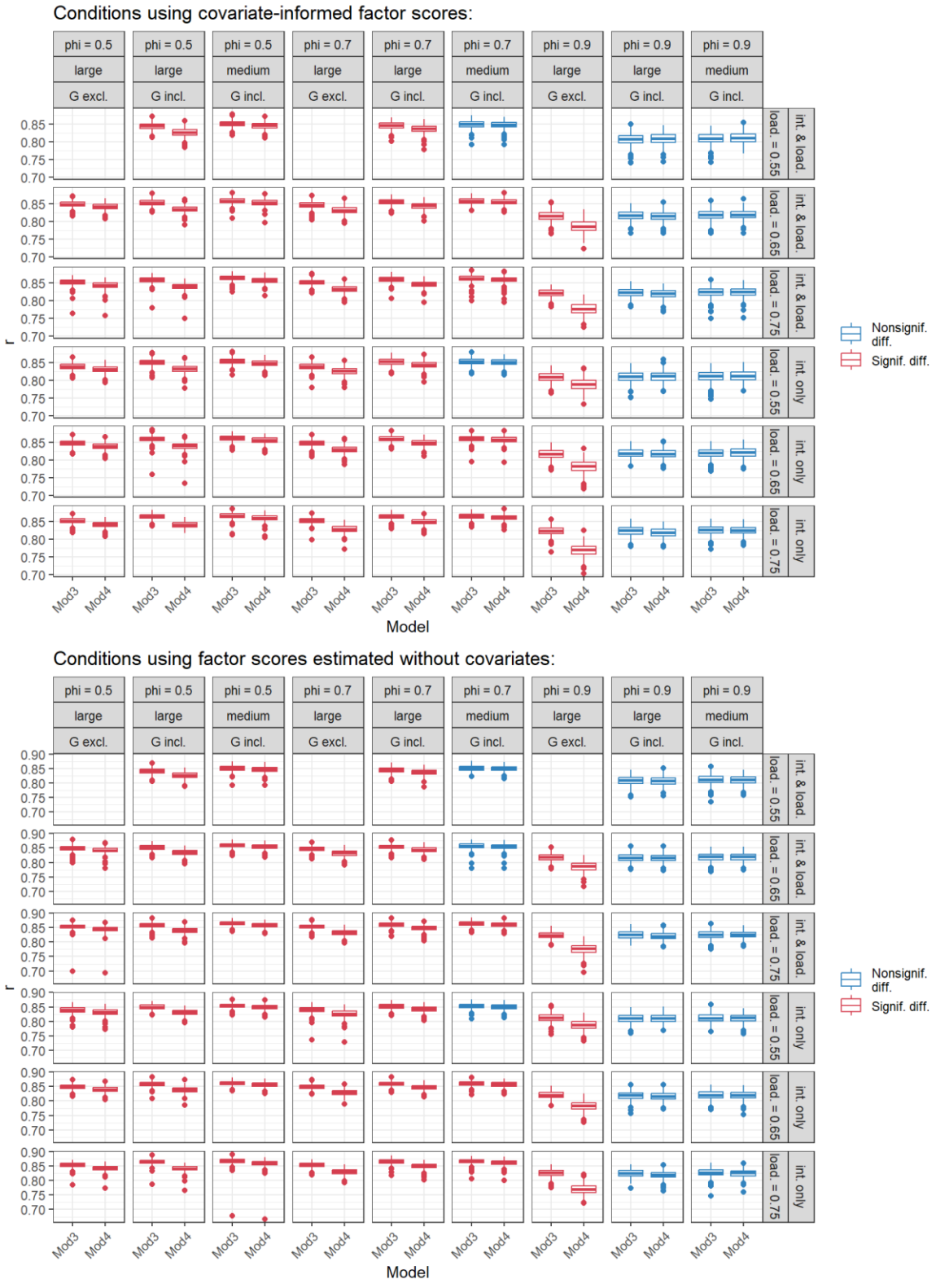
Condition-wise boxplots of the test performance of Models 3 and 4 in terms of RMSE (Panel a) and $r_{\eta,\hat{Y}}$ (Panel b)

Panel a




Conditions using factor scores estimated without covariates:
phi = 0.5
large
G excl.
phi = 0.5
large
G incl.
phi = 0.5
medium
G incl.
phi = 0.7
large
G excl.
phi = 0.7
large
G incl.
phi = 0.7
medium
G incl.
phi = 0.9
large
G excl.
phi = 0.9
large
G incl.
phi = 0.9
medium
G incl.
int. & load.
load. = 0.55
load. = 0.65
load. = 0.75
int. only
RMSE
1.0
0.9
0.8
Mod3
Mod4
Model
Nonsignif. diff.
Signif. diff.

Panel b

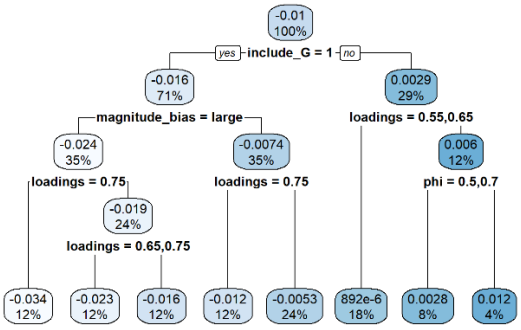


Note. A "significant" difference refers to paired sample t-test $p < \frac{.05}{102}$ and Cohen's $|d| \geq 0.5$.

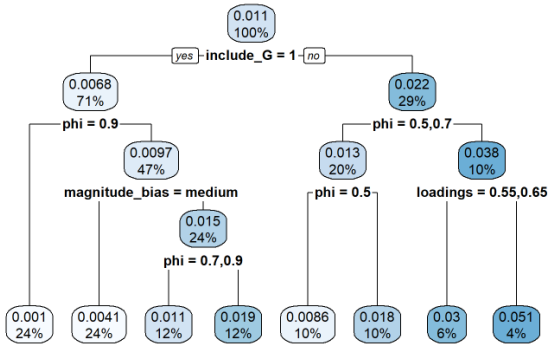
Figure A8

Tree diagram of a regression tree modeling the difference in root mean squared error (Panel a) and $r_{\eta,\hat{Y}}$ (Panel b) between Models 3 and 4, fit to simulation results across all conditions

Panel a



Panel b



Note. In each node of the tree diagram, the top value represents the mean value of the simulation outcome among observations belonging to that node, and the bottom value represents the percentage of total observations that belong to that node.

REFERENCES

Agency for Healthcare Research and Quality (2019, April 22), Survey Background. *Medical Expenditure Panel Survey*. Retrieved from https://www.meps.ahrq.gov/mepsweb/about_meps/survey_back.jsp

Agency for Healthcare Research and Quality (2019, August), MEPS HC-201: 2017 Full Year Consolidated Data File. Retrieved from https://www.meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-201

Agency for Healthcare Research and Quality (2020, August), MEPS HC-209: 2018 Full Year Consolidated Data File. Retrieved from https://www.meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-209

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Adjerid, I., & Kelley, K. (2018). Big data in psychology: A framework for research advancement. *American Psychologist*, 73(7), 899-917. https://psycnet.apa.org/doi/10.1037/amp0000190

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(4), 495-508. https://doi.org/10.1080/10705511.2014.919210

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and Machine Learning: Limitations and Opportunities*. http://www.fairmlbook.org

Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological methods*, *22*(3), 507-526. https://psycnet.apa.org/doi/10.10 37/met0000077

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Natesan Ramamurthy, K., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5), 1-15. https://doi.org/10.1147/JRD.2019.2942287

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

Berk, R., Heidari, H., Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., Neel, S., & Roth, A. (2017). A convex framework for fair regression. https://doi.org/10.48550/arXiv.1706.02409

Bleidorn, W., & Hopwood, C. J. (2019). Using machine learning to advance personality assessment and theory. *Personality and Social Psychology Review*, 23(2), 190-203. https://doi.org/10.1177/1088868318772990

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

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification And Regression Trees (1st ed.). Routledge. https://doi.org/10.1201/9781315139470

Brown, T. A., & Moore, M. T. (2012). Confirmatory factor analysis. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 361-379). New York, NY: Guilford Press.

Burckhardt, C. S., Anderson, K. L., Archenholtz, B., & Hägg, O. (2003). The Flanagan quality of life scale: Evidence of construct validity. *Health and quality of life outcomes*, *1*(59), 1-7. https://doi.org/10.1186/1477-7525-1-59

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*(3), 456-466. https://doi.org/10.1037/0033-2909.105.3.456

Chakraborty, J., Majumder, S., & Menzies, T. (2021, August). Bias in machine learning software: Why? how? what to do?. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 429-440. https://doi.org/10.1145/3468264.3468537

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, *5*(2), 153-163. https://doi.org/10.1089/big.2016.0047

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and white students in integrated colleges. *Journal of Educational Measurement, 5*(2), 115-124. https://www.jstor.org/stable/1434406

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences (2nd ed.). Routledge. https://doi.org/10.4324/9780203771587

Cole, N. S. (1973). Bias in selection. *Journal of educational measurement*, *10*(4), 237-255. https://doi.org/10.1111/j.1745-3984.1973.tb00802.x

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

Curran, P. J., Cole, V., Bauer, D. J., Hussong, A. M., & Gottfredson, N. (2016). Improving factor score estimation through the use of observed background characteristics. *Structural equation modeling: a multidisciplinary journal*, *23*(6), 827-844. https://doi.org/10.1080/10705511.2016.1220839

Darlington, R. B. (1971). Another Look at "Cultural Fairness." *Journal of Educational Measurement*, 8(2), 71–82. https://doi.org/10.1111/j.1745-3984.1971.tb00908.x

DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical assessment, research, and evaluation*, *14*(1), 20. https://doi.org/10.7275/da8t-4g52

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10), 78-87. https://doi.org/10.1145/2347736.2347755

Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. *Annual review of clinical psychology*, 14, 91-118. https://doi.org/10.1146/annurev-clinpsy-032816-045037

Einhorn, H. J., & Bass, A. R. (1971). Methodological considerations relevant to discrimination in employment testing. *Psychological Bulletin*, *75*(4), 261-269. https://doi.org/10.1037/h0030871

Executive Office of the President, (2016, May). *Big data: a report on algorithmic systems, opportunity, and civil rights*. Washington, D.C. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf

Fabris, A., Messina, S., Silvello, G., & Susto, G. A. (2022). Algorithmic fairness datasets: the story so far. *Data Mining and Knowledge Discovery*, 36, 2074-2152. https://doi.org/10.1007/s10618-022-00854-z

Fleishman, J. A., & Cohen, J. W. (2010). Using information on clinical conditions to predict high-cost patients. *Health services research*, *45*(2), 532-552. https://doi.org/10.1111/j.1475-6773.2009.01080.x

Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2016). On the (im) possibility of fairness. arXiv preprint arXiv:1609.07236.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. http://www.deeplearningbook.org

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, 63(11), 139–144. https://doi.org/10.1145/3422622

Gonzalez, O. (2021). Psychometric and machine learning approaches to reduce the length of scales. *Multivariate Behavioral Research*, 56(6), 903-919. https://doi.org/10.1080/00273171.2020.1781585

Gonzalez, O., O'Rourke, H. P., Wurpts, I. C., & Grimm, K. J. (2018). Analyzing Monte Carlo simulation studies with classification and regression trees. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(3), 403-413. https://doi.org/10.1080/10705511.2017.1369353

Goretzko, D., & Israel, L. S. F. (2021). Pitfalls of Machine Learning-Based Personnel Selection. *Journal of Personnel Psychology*, 21(1), 37-47. https://doi.org/10.1027/1866-5888/a000287

Graham, S., Depp, C., Lee, E. E., Nebeker, C., Tu, X., Kim, H. C., & Jeste, D. V. (2019). Artificial intelligence for mental health and mental illnesses: an overview. *Current psychiatry reports*, *21*(116), 1-18. https://doi.org/10.1007/s11920-019-1094-0

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods*, *6*(4), 430–450. https://doi.org/10.1037/1082-989X.6.4.430

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Hunter, J. E., & Schmidt, F. L. (1976). Critical analysis of the statistical and ethical implications of various definitions of test bias. *Psychological Bulletin*, 83(6), 1053-1071. https://doi.org/10.1037/0033-2909.83.6.1053

Hutchinson, B., & Mitchell, M. (2019, January). 50 years of test (un) fairness: Lessons for machine learning. *In Proceedings of the conference on fairness, accountability, and transparency* (pp. 49-58). https://doi.org/10.1145/3287560.3287600

Jacobs, A. Z., & Wallach, H. (2021, March). 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

Jacobucci, R., & Grimm, K. J. (2020). Machine learning and psychological research: The unexplored effect of measurement. *Perspectives on Psychological Science*, 15(3), 809-816. https://doi.org/10.1177/1745691620902467

Johnson, M. S., Liu, X., & McCaffrey, D. F. (2022). Psychometric methods to evaluate measurement and algorithmic bias in automated scoring. *Journal of Educational Measurement,* 59(3), 338-361. https://doi.org/10.1111/jedm.12335

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245), 255-260. https://doi.org/10.1126/science.aaa8415

Koedinger, K. R., D'Mello, S., McLaughlin, E. A., Pardos, Z. A., & Rosé, C. P. (2015). Data mining and education. *Wiley Interdisciplinary Reviews: Cognitive Science*, *6*(4), 333-353. https://doi.org/10.1002/wcs.1350

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. *Advances in neural information processing systems*, 30.

Landers, R. N., & Behrend, T. S. (2022). Auditing the AI auditors: A framework for evaluating fairness and bias in high stakes AI predictive models. *American Psychologist*. https://doi.org/10.1037/amp0000972

Liem, C., Langer, M., Demetriou, A., Hiemstra, A. M., Sukma Wicaksana, A., Born, M. P., & König, C. J. (2018). Psychology meets machine learning: Interdisciplinary perspectives on algorithmic job candidate screening. In *Explainable and interpretable models in computer vision and machine learning* (pp. 197-253). Springer, Cham. https://doi.org/10.1007/978-3-319-98131-4_9

Linn, R. L. (1973). Fair test use in selection. *Review of Educational Research*, 43(2), 139-161. https://psycnet.apa.org/doi/10.2307/1169933

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6), 1-35. https://doi.org/10.1145/3457607

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

Millsap, R. E. (2007). Invariance in Measurement and Prediction Revisited. *Psychometrika*, 72, 461-473. https://doi.org/10.1007/s11336-007-9039-7

Millsap, R. E. (2011). Statistical Approaches to Measurement Invariance (1st ed.). *Bias in Measurement and Prediction* (pp. 281-303). New York: Routledge. https://doi.org/10.4324/9780203821961

Millsap, R. E., & Olivera-Aguilar, M. (2012). Investigating measurement invariance using confirmatory factor analysis. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 380-392). New York, NY: Guilford Press.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8, 141-163. https://doi.org/10.1146/annurev-statistics-042720-125902

Morid, M. A., Kawamoto, K., Ault, T., Dorius, J., & Abdelrahman, S. (2018). Supervised learning methods for predicting healthcare costs: systematic literature review and empirical evaluation. In *AMIA Annual Symposium Proceedings 2017*, American Medical Informatics Association, 1312-1321. https://pubmed.ncbi.nlm.nih.gov/29854200

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447-453. https://doi.org/10.1126/science.aax2342

Olivera-Aguilar, M., Rikoon, S. H., Gonzalez, O., Kisbu-Sakarya, Y., & MacKinnon, D. P. (2017). Bias, type I error rates, and statistical power of a latent mediation model in the presence of violations of invariance. *Educational and Psychological Measurement*, *78*(3), 460-481. https://doi.org/10.1177/0013164416684169

O'Neil, C. (2017). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

Orrù, G., Monaro, M., Conversano, C., Gemignani, A., & Sartori, G. (2020). Machine learning in psychometrics and psychological research. *Frontiers in Psychology*, 10, 2970. https://doi.org/10.3389/fpsyg.2019.02970

Petersen, N. S., & Novick, M. R. (1976). An evaluation of some models for culture-fair selection. *Journal of Educational Measurement*, 3-29. http://www.jstor.org/stable/1434489

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental review*, *41*, 71-90. https://doi.org/10.1016/j.dr.2016.06.004

Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020, January). Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 469-481). https://doi.org/10.1145/3351095.3372828

Rakovski, C. C., Rosen, A. K., Wang, F., & Berlowitz, D. R. (2002). Predicting elderly at risk of increased future healthcare use: How much does diagnostic information add to prior utilization? *Health Services & Outcomes Research Methodology*, *3*, 267-277. https://doi.org/10.1023/A:1025866331616

R Core Team (2023). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. https://www.R-project.org/

Romano, Y., Bates, S., & Candes, E. (2020). Achieving equalized odds by resampling sensitive attributes. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., & Lin, H. (Eds.), *Advances in Neural Information Processing Systems, 33*, 361-371. Curran Associates, Inc. Available at https://proceedings.neurips.cc/paper_files/paper/2020/file/03593ce517feac573fdaafa6dcedef61-Paper.pdf

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, 48(2), 1-36. https://doi.org/10.18637/jss.v048.i02

Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., Rodolfa, K.T., & Ghani, R. (2018). Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577*.

Sawyer, R. L., Cole, N. S., & Cole, J. W. (1976). Utilities and the issue of fairness in a decision theoretic model for selection. *Journal of Educational Measurement*, 59-76. https://www.jstor.org/stable/1434493

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual review of psychology*, *69*, 487-510. https://doi.org/10.1146/annurev-psych-122216-011845

Singh M. & Ramamurthy, K. N. (2019). Understanding racial bias in health using the medical expenditure panel survey data. *NeurIPS 2019 workshop: "Fair ML for health".* arXiv:1911.01509

Spitzer, R. L., Kroenke, K., Williams, J. B., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: the GAD-7. *Archives of internal medicine*, *166*(10), 1092-1097. https://doi.org/10.1001/archinte.166.10.1092

Stekhoven, D. J. (2013). missForest: Nonparametric Missing Value Imputation Using Random Forest. R package version 1.4. Available at: https://cran.r-project.org/web/packages/missForest/missForest.pdf

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

Suresh, H., & Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and access in algorithms, mechanisms, and optimization* (pp. 1-9). https://doi.org/10.1145/3465416.3483305

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

Tay, L., Woo, S. E., Hickman, L., & Saef, R. M. (2020). Psychometric and validity issues in machine learning approaches to personality assessment: A focus on social media text mining. *European Journal of Personality*, 34(5), 826-844. https://doi.org/10.1002/per.2290

Therneau T. & Atkinson, B. (2022). *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1.19, https://CRAN.R-project.org/package=rpart.

Thorndike, R. L. (1971). Concepts of culture-fairness. *Journal of Educational Measurement*, 8(2), 63-70. https://doi.org/10.1111/j.1745-3984.1971.tb00907.x

Thurstone, L. L. (1935). The appraisal of abilities. In L. L. Thurstone, *The vectors of mind: Multiple-factor analysis for the isolation of primary traits* (pp. 226-231). University of Chicago Press. https://doi.org/10.1037/10018-010

Urban, C. J., & Gates, K. M. (2021). Deep learning: A primer for psychologists. *Psychological Methods*, *26*(6), 743-773. https://psycnet.apa.org/doi/10.1037/met0000374

Wherry, L. R., Burns, M. E., & Leininger, L. J. (2014). Using Self-Reported Health Measures to Predict High-Need Cases among Medicaid-Eligible Adults. *Health services research*, *49*(S2), 2147-2172. https://doi.org/10.1111/1475-6773.12222

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12(6), 1100-1122. https://doi.org/10.1177/1745691617693393

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018, December). Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 335-340. https://doi.org/10.1145/3278721.3278779

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