IMPROVING METHODS FOR PROPENSITY SCORE ANALYSIS WITH MIS-MEASURED VARIABLES BY INCORPORATING BACKGROUND VARIABLES WITH MODERATED NONLINEAR FACTOR ANALYSIS

Noah Greifer

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 & Neuroscience (Quantitative) in the College of Arts & Sciences.

Chapel Hill
2018

Approved by:

Patrick Curran

Kirsten Kainz

David Thissen

Daniel Bauer

ABSTRACT

Noah Greifer: Improving Methods for Propensity Score Analysis with Mis-Measured Variables by Incorporating Background Variables with Moderated Nonlinear Factor Analysis
(Under the direction of Patrick Curran)

There has been some research in the use of propensity scores in the context of measurement error in the confounding variables; one recommended method is to generate estimates of the mis-measured covariate using a latent variable model, and to use those estimates (i.e., factor scores) in place of the covariate. I describe a simulation study designed to examine the performance of this method in the context of differential measurement error and propose a method based on moderated nonlinear factor analysis (MNLFA) to try to address known problems with standard methods. Although MNLFA improves effect estimation somewhat in the presence of differential measurement error relative to standard factor analysis methods, the greatest gains come from the nonstandard practice of including the treatment variable as an indicator in the scoring models. More research is required on the effects of model misspecification on the performance of these methods for causal inference applications.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ATE	average treatment effect
CBCL	Child Behavior Checklist
CE	conditional exchangeability
DIF	differential item function
IRT	item response theory
LV	latent variable
MNLFA	moderated nonlinear factor analysis
PBR	percent bias remaining
RMS	root mean squared
RMSBD	root mean squared balance discrepancy
SMD	standardized mean difference
SUTVA	stable unit treatment value assumption
wABC	weighted area between curves
WLS	weighted least squares

INTRODUCTION

Randomized control trials have long been considered the "gold standard" design method for making causal inferences in the social and health sciences (Jones & Podolsky, 2015). When randomization is successful, any observed difference in outcomes after receipt of the treatment can be due only to treatment status and not to some other variable that might otherwise explain the relationship between treatment status and outcome (Shadish, Cook, & Campbell, 2002). Although random assignment is frequently used in psychology to answer causal questions, often random assignment is unethical or impossible. For example, researchers cannot randomly assign whether people experience childhood trauma, use illicit substances in adolescence, or are held back a year in school, but clearly the causal effects of these events are of great concern to researchers and policymakers. There may also be scenarios in which random assignment is not desirable, because it involves forcing a subset of individuals into a treatment condition they might not otherwise want be placed in; the conclusions of randomized studies can therefore lack external validity (Rothwell, 2005).

Instead, psychologists often employ observational studies, studies in which the event of interest is not randomly assigned by the researcher, but rather is chosen by the participant or some other actor. It is no longer so simple to draw a causal effect estimate from a comparison of the outcomes of those in various conditions, because the observed differences may be due to some quality of the subjects that also influenced their treatment assignment. For example, Odgers et al. (2008) sought to estimate the causal effect of early adolescent substance use on adult criminal convictions, but substance users were more likely to have a parent who was

convicted, which itself might explain why substance users have more convictions in adulthood. Instead of simple comparisons between conditions, researchers will have to rely on statistical methods and sets of assumptions to identify causal effects and make valid causal inferences for their data. These methods are further complicated by measurement error, ubiquitous in psychology, in which the constructs of interest are often not directly observable.

The structure of this paper is as follows. First, I introduce the assumptions required and methods available for making causal inferences with social science data, with a focus on propensity score methods. Next, I describe the problems with propensity score methods caused by measurement error and recent attempts to solve these problems. Next, I present a new potential solution that builds off prior methods involving latent variable analysis by including a recent methodological innovation, namely moderated nonlinear factor analysis (Bauer, 2017; Bauer & Hussong, 2009). Finally, I present the results of a simulation study designed to examine the plausibility and effectiveness of this proposed method to improve the performance of propensity score methods in conditions of measurement error; this is the focus of my work here.

The Potential Outcomes Framework

The potential outcomes framework is a valuable conceptual and mathematical tool to understand the problems associated with making causal inferences. In this framework, each individual has two potential outcomes, $Y^{z=0}$ and $Y^{z=1}$, where Y is a continuous outcome variable and z is an instance of random variable Z denoting treatment status taking the values 0 (control) and 1 (treated). These correspond to the potential outcomes if the unit were to receive control and if the unit were to receive treatment. We can define the individual treatment effect for a unit *i* to be

$$\tau_i = Y_i^{z=1} - Y_i^{z=0} \quad (1)$$

and the average treatment effect in the population (ATE) to be

$$ATE = E[Y_i^{z=1} - Y_i^{z=0}] = E[Y_i^{z=1}] - E[Y_i^{z=0}] \quad (2)$$

In fact, though, for each individual, only one of the potential outcomes is observed. The other is the counterfactual outcome (i.e., counter to fact). This is considered the fundamental problem of causal inference: neither $E[Y_i^{z=1}]$ nor $E[Y_i^{z=0}]$ can be directly computed to attain a causal effect estimate. Instead, we can compute functions only of the observed outcomes $E[Y_i^{z=1}|Z=1]$ and $E[Y_i^{z=0}|Z=0]$, the potential outcomes realized corresponding to the actual treatment assignment of the individuals.

However, the expectations of the potential outcomes are identifiable under several assumptions. These are the stable unit treatment value assumption (SUTVA), positivity, consistency, and exchangeability (Foster, 2010). SUTVA requires that a unit's potential outcomes do not depend on the treatment status of other individuals (and thus are "stable"). Positivity requires that it is theoretically possible for all units to be in either treatment condition (i.e., no combinations of qualities systematically exclude units from either condition). Consistency requires that there are no unspecified versions of treatment. Although these assumptions are worthy of study, the final assumption of exchangeability, described below, is the focus of this paper, and the aforementioned assumptions will be taken for granted here.

Exchangeability and Confounding

Exchangeability is a core assumption and requires that there is no association between treatment status and potential outcomes:

$$Y^z \perp Z \; for \; all \; z \quad (3)$$

Insofar as potential outcomes are a function of some set of covariates and the actual treatment to be received, this means that this set of covariates is not also associated with treatment

assignment. The assumption of exchangeability will be the focus of this work and is the focus of most work in causal inference research. Conceptually, exchangeability can be thought of as the assumption that there are no alternative explanations for the observed association between actual treatment assignment and actual outcome value other than the causal effect of treatment on the outcome. In other words, there is no confounding.

Asymptotically, randomization guarantees exchangeability because when units are randomized to conditions the joint distribution of factors that influence potential outcomes will be the same across conditions (even if the factors are not measured). In this way, treatment assignment is independent of potential outcomes (because it is independent of all covariates that relate to potential outcomes), and exchangeability is met. In this case,

$$E[Y^{z=1}|Z = 1] \; = \; E[Y^{z=1}] \tag{4}$$

and

$$E[Y^{z=0}|Z = 0] \; = \; E[Y^{z=0}] \tag{5}$$

Because the first expression in each of the equations (4) and (5) is identifiable from the observed data, the desired estimand—the unconditional expectation of each potential outcome—can be computed using equation (2). A simple comparison of group means will yield a valid estimate of the treatment effect.

Exchangeability can be extended to the case of conditionally randomized experiments, wherein within each level of some factor (i.e., a random variable C), randomization occurs and assignment is independent of potential outcomes. In this case, conditional on C, there is exchangeability; this is known as conditional exchangeability (CE; Hernán & Robins, 2018, Ch. 2), or the strongly ignorable treatment assumption (Rosenbaum & Rubin, 1983). Expanding on equation (3), CE is formalized as

$$Y^z \perp Z \,|C = c \; for \; all \; z, c \tag{6}$$

For example, to examine the effect of a promising school policy intervention aimed at improving the academic performance of financially disadvantaged children, researchers might randomly assign schools in low income neighborhoods to treatment with 0.75 probability and might randomly assign schools in middle income neighborhoods to treatment with 0.5 probability. It is clear that neighborhood income class will affect the potential outcomes of the schools, regardless of treatment status, but conditional on (i.e., within) neighborhood class, treatment is independent of the potential outcomes, and so a marginal causal effect can be identified (Hernán & Robins, 2018). This insight is critical to understanding exchangeability in observational studies, which are more common than conditionally randomized experiments in psychology.

When randomization does not occur and, for example, individuals are allowed to choose their own treatment condition, equations (4) and (5) do not hold and the causal effect cannot be immediately identified. Instead, a simple comparison of group means will include information both about the treatment effect and the other associations between assignment and potential outcomes. For example, if those with parents who have histories of conviction are more likely to engage in illicit substance use as adolescents, the difference in adult convictions between adolescent substance users and abstainers will include not only the causal effect of drug use on the individual's convictions but also the association between parental conviction history and the individual's convictions through pathways other than the increased propensity to use substances. This failure to identify the correct causal effect is called "bias" and is a direct result of confounding (or a violation of any of the aforementioned assumptions). Figure 1 is a schematic depiction of a confounding scenario, where **C** is the set of common causes of Treatment and Outcome. In the example above, we can consider adolescent illicit substance use the treatment,

adult convictions the outcome, and having parents with a history of convictions a common cause of treatment and outcome that creates confounding.

In observational studies, it may be possible to condition on a set of confounding variables and arrive at CE. This will be true if the observational study can be thought of as a conditionally randomized experiment: conditional on a set of pre-treatment variables, individuals are randomly assigned (with some nonzero probability) to condition (Hernán & Robins, 2018, Ch. 3). Therefore, to identify causal effects in observational studies, one can condition on a set of pre-treatment variables associated with treatment and potential outcomes and arrive at a valid estimate of the causal effect without true randomization. The identification of the set of sufficient variables to eliminate confounding is its own area of research (e.g., Brookhart et al., 2006), but the focus for the rest of this study will be on the act of conditioning on those variables, given that they are already known.

Adjusting for Confounding

Three methods of conditioning on variables include matching, stratification, and regression. In matching, individuals in one treatment condition are matched to individuals in the other based on similarity of covariate values, and the causal effect estimate is the average of the pairwise outcome differences. In stratification, units are stratified into quantiles of covariate values, and treatment effects are estimated within each stratum. A problem with these two methods is that with many confounders, which are often necessary to eliminate confounding in observational studies, these methods fail: it will be impossible to find units with the same or even similar values of the entire set of covariates, yielding units without matches and strata with too few units (Rosenbaum & Rubin, 1984). This problem is often known as the "curse of dimensionality."

Regression is regarded as a solution to this problem because regression models can include many variables while yielding efficient estimates of the treatment effect. Regression is

commonly used as a statistical method to condition on sets of variables in conditionally randomized and observational studies. A weakness of regression is that the functional form of the relationship between all of the included covariates and the outcome must be correctly specified in the regression model, or else residual confounding can occur (Schafer & Kang, 2008). For example, if the true outcome model includes interactions and nonlinear terms but a simple linear regression missing those forms is specified, the effect estimate will be biased and inconsistent (Hernán & Robins, 2018, Ch. 15; Schafer & Kang, 2008). There are other reasons why researchers may want to avoid regression in favor of some other method to adjust for confounding; these will be discussed later to contrast regression with the methods presented next.

Balancing Scores and the Propensity Score

Consider a set of variables **C** for which conditioning on **C** is sufficient to eliminate confounding and arrive at CE. A balancing score b(**C**) is a value or set of values that, when conditioned upon, yields conditional independence between the covariates and the treatment, thereby satisfying the requirements for CE. The full set of confounding variables itself is (trivially) a balancing score. However, as described above, conditioning on the full set through matching or stratification can fail due to the curse of dimensionality, and conditioning on the full set through regression requires model assumptions that are unlikely to be met. In their landmark paper, Rosenbaum and Rubin (1983) discovered a unidimensional balancing score, known as the propensity score, which could allow for matching and stratification to achieve exchangeability. Formally, they discovered a b(**C**) such that

$$Y^z \perp Z \,|\boldsymbol{C} \Rightarrow Y^z \perp Z \,|b(\boldsymbol{C}), \tag{7}$$

where

$$b(\boldsymbol{C}) \equiv P(Z = 1|\boldsymbol{C}). \tag{8}$$

That is, if there is a set of covariates **C** for which conditioning on **C** is sufficient to eliminate

confounding and arrive at CE, then conditioning on the conditional probability of receiving treatment given **C**—the propensity score—is also sufficient to arrive at CE. In a conditionally randomized experiment, the probability of receiving treatment is set by the researcher, and non-parametric techniques like matching and stratification on this known probability will yield unbiased estimates of a treatment effect.

A variety of methods have been developed for matching on the propensity score, including the traditional and simple nearest neighbor matching, as well as more sophisticated alternatives such as full matching (Stuart & Green, 2008) and genetic matching (Diamond & Sekhon, 2013). With stratification, researchers typically form quantiles of the propensity score and estimate treatment effects within each quantile; this technique is used less frequently than matching (Thoemmes & Kim, 2011). Propensity scores can be used in inverse probability weighting for marginal structural models (Robins, Hernan & Brumback, 2000), where a function of the propensity score is used as a sampling weight in weighted estimation. Occasionally, the propensity score itself is used as a covariate in a regression model, but this method has fallen out of favor due to its poor empirical properties and additional required assumptions (Austin, 2011; Thoemmes & Kim, 2011). The primary focus here will be on propensity score weighting.

Estimating Propensity Scores

In practice, the true propensity score is not known, so it must be estimated from the sample data at hand. A variety of methods can be used to generate propensity scores from sample data, which generally involve specifying a parametric or non-parametric model predicting treatment assignment from the available covariates. The fitted values from this model then form the estimated propensity scores. A common method is to use logistic regression and to use the model-predicted probabilities of treatment assignment as the propensity scores. The propensity score model can include all or a subset of the relevant covariates as well as polynomials,

interactions, and other non-linear terms; machine learning techniques such as generalized boosted modeling (McCaffrey, Ridgeway, & Morral, 2004) can simplify this process by requiring fewer modeling decisions from the user in cases of uncertainty (Lee, Lessler, & Stuart, 2010).

The resulting estimated propensity scores must then be evaluated for their ability to achieve CE by empirically examining whether, after conditioning on the propensity score, the joint distributions of covariates are similar across treatment groups (Stuart, 2010). This distributional similarity, an approximation of CE in the sample, is often referred to as *balance*. Balance assessment is an ongoing area of research, but typical methods involve comparing the means and higher moments of covariate distributions and interactions across treatment groups and using visual diagnostics such as kernel density and Q-Q plots (Kainz et al., 2017). A common and recommended measure of balance is the standardized mean difference (SMD), defined by the following expression (Austin, 2011):

$$SMD = \frac{M_1 - M_0}{\sqrt{\frac{s_1^2 + s_0^2}{2}}} \tag{9}$$

where M_1 and M_0 are the (weighted) group means of the covariate under study, and s_1^2 and s_0^2 are the group variances. In propensity score weighting, the SMD is computed with the group means weighted by the estimated propensity score weights, but the group variances remain unweighted (Stuart, 2010). Typically, SMD values below 0.10 in absolute value are considered adequate (Stuart, 2010). It is important to note that propensity score models are not to be evaluated on traditional model evaluation criteria such as goodness of fit or parsimony; attaining balance on the covariates of interest is paramount (Stuart, 2010).

If the estimated propensity score and conditioning specifications do not yield satisfactory

balance, the propensity score model can be respecified, such as by adding squared terms or interactions, and reevaluated until balance is achieved (Rosenbaum and Rubin, 1984; Stuart & Rubin, 2008; Stuart, 2010). Because this process does not involve the outcome variable, there is no risk of capitalizing on chance to arrive at a specific treatment effect estimate; the propensity score stage of a full analysis is akin to the design stage of a study, in that applying the propensity score is essentially adjusting the selection parameters for the sample (Rubin, 2001). Unlike typical statistical analyses, overfitting to the data is not a problem because the goal of propensity score analysis is to arrive at sample covariate balance irrespective of the interpretability, plausibility, reproducibility, or parsimony of the propensity score model (Augurzky & Schmidt, 2001; Stuart, 2010).

Estimating a Treatment Effect

Once propensity scores have been estimated and balance has been achieved, an analyst can then estimate the treatment effect in their propensity score-conditioned data. For propensity score matched samples, a matched pairs t-test on the matched samples is recommended (Austin, 2011). For propensity score weighted data, an outcome regression model can be specified as follows:

$$E[Y_j] = \beta_0 + \beta_1 Z_j, \quad (10)$$

where Y_j is the outcome, Z_j is treatment assignment, and β_0 and β_1 are the intercept and treatment effect, respectively, which are to be estimated (Robins et al., 2000). This model is then fit with weighted least squares (WLS) regression (i.e., by minimizing the weighted sum of squares) or by generalized estimating equations to arrive at an estimate for β_1, which corresponds to the causal treatment effect estimate (Hernán & Robins, 2018, Ch. 12). After effect estimation, there are a variety of methods to assess the potential impact of unobserved confounding on estimates of the treatment effect (Liu, Kuramoto, & Stuart, 2013).

Propensity Scores vs. Regression Models

In psychology, the practice of statistically modeling relationships among variables using regression or structural equation modeling is popular (Foster, 2010). Researchers can specify a parametric model for the outcome conditional on a linear combination of predictors, including treatment and variables required to eliminate confounding. The estimated treatment effect can then often be identified by examining the coefficient estimate for the treatment variable in the model (Schafer & Kang, 2008). On the other hand, propensity score methods are explicitly a non-modeling approach, in that the functional form of the relationship between the covariates and the outcome variable does not have to be modeled (Ho, Imai, King, & Stuart, 2007). Though the propensity score itself must be modeled, the process and reasons for modeling the propensity score vastly differ from those of confirmatory parametric models, opening up the possibility of employing machine learning and other optimization-based techniques that would normally be reserved for exploratory data analysis.

Several authors have discussed the differences between regression and propensity score approaches for conditioning on confounding variables. The core of these discussions is that the commonly employed structural assumptions of the form of the relationship between covariates in regression are often untenable, and failing to specify the correct functional form for important covariates can yield additional bias (Foster, 2010; Ho et al., 2007; Schafer & Kang, 2008). Because regression is a confirmatory approach that involves having access to the outcome variable and estimating a treatment effect with each run, continually respecifying an outcome regression model to improve the plausibility of exchangeability can run the risk of overfitting or a temptation to select an outcome model that yields a treatment effect in accordance with a researcher's hypothesis (King & Nielsen, 2016; Rubin, 2001). Because the respecification of propensity score models depends on criteria distinct from inference (i.e., covariate balance) and

is done without considering the outcome data or a treatment effect estimate, propensity score methods do not face these issues.

Despite these potential pitfalls, regression can be a valuable tool for causal inference. It has been shown that when researchers apply regression and propensity score methods separately to analyze the same data, the substantive conclusions are almost identical (Shah, Laupacis, Hux, & Austin, 2005). Several authors recommend regression on datasets preprocessed through propensity score or other matching methods (e.g., Ho, Imai, King, & Stuart, 2007; Rubin, 2001). Freedman and Berk (2008) found in their simulations that regression alone actually performed better than using propensity score weights alone or propensity score weighted regression. Indeed, there is still debate about the value of propensity score methods when linear regression often arrives at the same conclusion and does so with greater apparent precision (Shadish, Clark, & Steiner, 2008).

In the end, the debate boils down to a bias-variance tradeoff: with propensity score methods, there is a major emphasis on reducing bias, often at the expense of statistical efficiency (i.e., by removing cases after matching or down-weighting cases with weighting); whereas with regression, the efficiency properties of maximum likelihood estimation yield high precision of the estimates, but often at the expense of potential unbiasedness when the model is misspecified (Golinelli, Ridgeway, Rhoades, Tucker, & Wenzel, 2012; Schafer & Kang, 2008). For the purposes of this study, I focus on the application of propensity scores for effect estimation on observational studies, though regression-based techniques should not be discounted in estimating causal effects.

Propensity Scores and Measurement Error

In many of the social sciences, psychological variables can produce confounding because they are often related both to treatment selection (such as when individuals get to choose their own treatment) and to outcomes (such as when the outcome depends on motivation or is caused by baseline psychological characteristics). A major issue with psychological variables is that they are almost always measured with error (Bollen, 2002). Until recently, the problems associated with conditioning on an indicator of a mis-measured variable rather than on the true variable itself had been largely ignored in the causal inference literature. Steiner, Cook, and Shadish (2011) were the first to systematically investigate the effects of measurement error in covariates of the propensity score model on bias in the estimated treatment effect. They found that decreasing the reliability of measures of latent covariates associated with the treatment and outcome led to increased bias in treatment effect estimates. Other researchers found similar results in simulations that attempted to provide solutions to the problem of measurement error (e.g., Jakubowski, 2015, McCaffrey, Lockwood, & Setoji, 2013). Rodríguez De Gil et al. (2015) ran a large, comprehensive simulation study examining the effects of covariate unreliability on treatment effects estimated with propensity scores. They found that even small unreliability in covariates (e.g., reliabilities of .8) can lead to marked increases in bias, increased Type I error rates, and decreased confidence interval coverage. They did not examine a solution to these problems.

Several approaches have been developed to deal with covariate measurement error and related issues such as covariate missingness and unmeasured confounding, all of which can be considered issues in the same vein (e.g., Cole, Chu, & Greenland, 2006). These approaches include multiple imputation with calibration (Webb-Vargas, Rudolph, Lenis, Murakami, & Stuart, 2015), corrected propensity score weighting (McCaffrey, Lockwood, & Setoji, 2013), and

subclassification on latent classes (Masyn & Walderman, 2016). Although promising, most current approaches have limitations that hamper their widespread use: they are challenging to implement for substantively oriented researchers, some require external validation samples or otherwise untestable but major distributional assumptions, and they are still in their infancy, with little empirical validation.

Another solution that has gained some attention is to use latent variable (LV) models to generate estimates of the mis-measured covariate and then use those estimates in place of the true variable in standard propensity score analysis. This method involves generating factor scores from the observed indicators of the LV using standard factor analysis or principal components analysis. Raykov (2012) proposed this solution and supported it with analytical derivations and a simulated demonstration of its efficacy. Jakubowski (2015) explored this method as well, using simulations to more systematically examine its effectiveness under a variety of circumstances, including treatment model misspecification. Both authors found modest reductions in bias relative to using the observed covariates in propensity score analysis.

In a manuscript under review for publication, Nguyen, Hong, Ebnesajjad, and Stuart (under review) expanded on this method by incorporating other covariates and the treatment variable into a structural equation model and using factor scores generated from this model in a standard propensity score analysis. In this way, treatment was modeled as an indicator of the LV. This method yielded dramatic improvement in bias reduction compared to the method of using factors scores generated from the measurement model only. Because of its effectiveness and the ease of performing structural equation modeling and generating factor scores, this method holds promise for widespread use. Though it may seem unusual to include the treatment variable as an indicator of the LV when estimating factor scores as done in Nguyen et al. (under review), there is some

precedent for including the outcome in estimating the predictor of that outcome.

First, the Mantel-Haenszel technique, used to determine whether a test item functions differentially for two groups of units that are otherwise identical on their level of the measured construct, involves matching units with similar levels of the construct to be measured and comparing their responses to items (Michaelides, 2008). Holland and Thayer (1988) and Zwick (1990) found that matching on a proxy for the construct that included the item under study (i.e., the item for which the differential functioning between groups was in question) yielded superior performance for accurately assessing whether the item functioned differentially. In this way, the relationship between group membership and item response is examined conditional on a measure that includes the item response, similar to how Nguyen et al. (under review) proposed that the outcome of the propensity score model (i.e., the treatment) should be used to construct the score that is used as a predictor of that same outcome.

Second, when missing data techniques (e.g., multiple imputation or full-information maximum likelihood) are used to "fill in" values for the missing variable, a distribution for the missing variable must be specified conditional on a set of related covariates (Collins, Schafer, & Kam, 2001). When the variable with missingness is used to predict an outcome (and therefore its relationship with the outcome is in question), it is recommended to include the outcome in the estimation of the conditional distribution of the variable (Leyrat et al., 2017; Meng, 1994; Moons, Donders, Stijnen, & Herrell, 2006). In the sense that LVs and missing data are related concepts (Blackwell, Honaker, & King, 2017), it would seem to be prudent to apply the lessons from the missing data literature to problems in measurement error (Mislevy, Johnson, & Muraki, 1992). Indeed, in the context of categorical latent variables, this line of reasoning led to the development of a similar approach for latent class analysis with distal outcomes, in which the

distal outcome is included as a covariate in the latent class model used for class assignment before the outcome is regressed on class membership (Bray, Laza, & Tan, 2015).

Instead, Burt (1976) described the procedure of using outcomes as indicators of LV models as leading to "interpretational confounding," in that the meaning of the LV (and therefore its relationship with its indicators) changes when including an outcome in its measurement model. To the degree the meaning of the LV affects its relationship with an outcome, some authors argue that measures should be taken to separate the measurement step and outcome modeling step entirely (Burt, 1976). The approach of Nguyen et al. (under review) explicitly violates this recommendation by including the outcome (i.e., the treatment) in the measurement model and using the estimated scores from that model in predicting the same outcome.

There is both theoretical precedent and controversy in the use of the treatment as an indicator to estimate factor scores for propensity score analysis. Below I discuss factor scores and measurement models in more detail, primarily considering standard measurement models (in the sense that only indicators of a LV, as opposed to outcomes, are included).

Factor Analysis and Factor Scores

Factor scores, also known as scale scores in item response theory (IRT), are point estimates of units' values on a LV given some indicators and a model linking the LV to the indicators (i.e., the measurement model). There are variety of ways to compute factor scores (Grice, 2001), though they are often highly correlated with each other (Fava & Velicer, 1992). The first step is often to specify and estimate a (generalized) factor score model, which parametrically links the LV to its indicators, using an equation such as

$$\mu_{ij} = g_i^{-1}(\nu_i + \lambda_i \eta_j) \quad (11)$$

where μ_{ij} is the expected value of indicator i for person j, η_j is the value of the LV for individual j

assumed be normally distributed as

$$\eta_j \sim N(\alpha, \psi) \tag{12}$$

where α and ψ are the mean and variance, respectively, of the LV, and v_i and λ_i are parameters describing the linear relationship between η_j and μ_{ij} (Bauer, 2017; Bauer & Hussong, 2009). In relation to a generalized linear model, $g_i^{-1}(.)$ is an inverse link function, which may differ for each item depending on its type; a response function is also specified to relate the expected value of the indicator μ_{ij} to actual indicator values u_{ij}. For example, a linear factor model with continuous indicators might include an identity link (and inverse link) and a normally distributed response function, while a 2-PL IRT model with binary indicators would include a logit link and a binomial response function (Bauer & Hussong, 2009). For this 2-PL model, the relationship between μ_{ij} and the observed response u_{ij} for item i would be defined as $P(u_{ij} = 1| \eta_j) = \mu_{ij}$.

Once parameters for the factor model have been estimated, they can be used to generate factor scores based on individuals' patterns of responses and assumptions about the distribution of the LV. For each individual, a "posterior" distribution f for the LV can be formed by taking the product of the conditional item response functions (i.e., measurement equations) for each item and the conditional LV distribution (usually specified as Gaussian), as in the following equation:

$$f(\eta_j|\boldsymbol{u}_j) = \prod_{i=1}^{nitems} T_i(u_{ij}|\eta_j)\phi(\eta) \tag{13}$$

where $\mathbf{u}_j$ is the vector of all item response for unit j, T_i is the estimated response curve for item i, and ϕ is the probability density function of η (i.e., mentioned in equation 12). Factor scores can be computed as the expectation or mode of this posterior distribution. Though there are a variety of choices to be made when considering how to compute factor scores, especially with

continuous items (Grice, 2001), Lu and Thomas (2008), building off core results of Skrondal and Laake (2001), recommend using the expected *a posteriori* (EAP) score estimates, which involve computing the expectation of this posterior distribution, when the scores are to be used as predictors in regression models.

A problem with estimating factor scores, and with factor analysis in general, is factor indeterminacy; perfect estimates of the LVs are impossible because the LV does not have a posterior point distribution after observing and modeling the indicators (Maraun, 1996). That is, $f(\eta_j|\boldsymbol{u}_j)$ has nonzero variance. Modeling more information about the LV can reduce indeterminacy by reducing the posterior variance, yielding more precise estimates of the LV in the form of factors scores (Fava & Velicer, 1992). This information can come in the form of a larger sample size, more indicators, and higher multiple correlation of other covariates with the LV (Bollen, 2002). Given a constant sample size, including more indicators (i.e., consequences) of the LV and modelling covariances with other observed variables will thereby reduce factor indeterminacy, leading to superior factor score estimates (Mislevy et al., 1992). These properties may help explain why the approach to propensity score estimation of Nguyen et al. (under review) led to improved performance of the propensity scores: including the treatment variable as a consequent of the LV increased the number of its indicators, and including the covariances between other covariates and the LV increased the multiple correlation for the LV, both of which reduce its indeterminacy by reducing the variance of it posterior distribution.

Moderated Nonlinear Factor Analysis

Factor score estimates can be improved with available background variables in other ways. Curran et al. (2016) found that using background variables in the measurement model improved factor score estimates. The reason is that when holding the often-untenable assumption that the

same scoring algorithm exists for all units, estimates of the LV are biased and imprecise, resulting from incorrect modeling of the relationship between the LV and its indicators. Two ways background variables might influence measurement are known as *impact* and *differential item function* (DIF). Impact occurs when background variables affect the distribution of the LV (i.e., by creating different values of α and ψ in equation (12) for each individual). The LV mean and variance for individual *j* are instead specified as

$$\alpha_j = f_\alpha(\boldsymbol{X}_j) \tag{14}$$

and

$$\psi_j = f_\psi(\boldsymbol{X}_j), \tag{15}$$

where α_j and ψ_j are the mean and variance, respectively, of the LV for individual *j* with background variables $\mathbf{X}_j$, and f_α and f_ψ relate $\mathbf{X}_j$ to α_j and ψ_j. DIF occurs when background variables affect the parameters relating the LV to its indicators (i.e., by creating different values of ν_i and λ_i in equation (11) for each individual). In this case, the item intercept and loading for individual *j* are specified as

$$\nu_{ij} = f_{\nu_i}(\boldsymbol{X}_j) \tag{16}$$

and

$$\lambda_{ij} = f_{\lambda_i}(\boldsymbol{X}_j), \tag{17}$$

where ν_{ij} and λ_{ij} are the factor intercept and loading for item *i* for individual *j* with background variables $\mathbf{X}_j$, and $f_{\nu i}$ and $f_{\lambda i}$ relate $\mathbf{X}_j$ to ν_{ij} and λ_{ij}.

Unmodeled impact and DIF affect factor score estimation because the parameters of the LV distribution and measurement equations will be assumed to be constant across individuals (i.e., not involving on $\mathbf{X}_j$), thereby yielding an incorrect posterior distribution from which to compute the factor scores. It is possible to incorporate background variables into LV models using

moderated nonlinear factor analysis (MNLFA), an expansion of generalized factor analysis that allows for the simultaneous modeling of the relationships between covariates and both the LV and its indicators (Bauer & Hussong, 2009; Curran, et al., 2014). Doing so involves specifying a model for the LV mean, a model for the LV variance, and models for the parameters in the measurement models of the indicators, all of which can contain background covariates of any variable type, subject to identifiability constraints. Thus, each of the parameters estimated in the models in equations (11) and (12) can vary across individuals, and can be modeled in terms of functions containing background covariates as in equations (14) through (17).

The models are simultaneously estimated, yielding parameter estimates that can be used to generate factor score estimates using an updated version of equation (13) (Bauer & Hussong, 2009):

$$f\left(\hat{\eta}_j \middle| \boldsymbol{u}_j, \boldsymbol{X}_j\right) = \prod_{i=1}^{nitems} T_i(u_{ij}|\eta_j, \boldsymbol{X}_j)\phi(\eta_j|\boldsymbol{X}_j) \quad (18)$$

which now conditions on $\mathbf{X}_j$. T_i and ϕ now explicitly involve *j* and involve estimating the parameters in f_α, f_ψ, f_{vi}, and $f_{\lambda i}$.

Curran, et al. (2016) used a simulation to examine the performance of this technique and others for estimating factor scores in various conditions of covariate involvement in the measurement model. Simulation factors included percent of items with DIF, ratio of mean impact to variance impact, number of items, and the magnitude of DIF. Examined scoring models included proportion scores (i.e., the mean of the indicators), unconditional MNLFA (i.e., a 2-PL IRT model), MNLFA accounting for impact, and MNLFA accounting for both impact and DIF. The authors found fairly substantial improvements in factor score recovery (as measured by the correlation between the estimated scores and the true scores and the RMSE of

the estimated scores) under all scenarios when using MNLFA that accounted for both impact and DIF, even when the covariate involvement was small, and especially when it was large. The authors conclude by recommending the inclusion of background variables when available through MNLFA in estimating factor scores.

Propensity Scores and MNLFA

Given these findings, it would appear that involving background variables in factor score estimation for propensity score modeling is a plausible solution to the problem of improving propensity score-based methods with mis-measured covariates. In settings in which propensity scores are commonly used, background variables are often plentiful, as they are included in the models used to estimate the propensity scores. These background variables may be able to provide more information than solely their prediction of the probability of treatment assignment; including some of them in a MNLFA model for the estimation of the factors scores of the latent confounder may also help to decrease bias and improve the precision of the effect estimate. As the use of factor scores in propensity score estimation is still in its infancy, these possibilities have not yet been examined, and they may provide a benefit to research assessing causal effects in the ubiquitous circumstance of confounding and measurement error.

With the present study, I aimed to address a fundamental problem in propensity score analysis with human data: that measurement error is often inherent in the covariates one must account for to eliminate confounding. By incorporating the findings by Jakubowski (2015) and Nguyen et al. (under review) that factor scores can improve the results of propensity score analyses and the finding by Curran et al. (2016) that background variables can be used to improve factor score estimation, I hoped to provide a new set of procedures for applied researchers in the context of covariate measurement error. The application of propensity score methods and LV models in tandem may help bring advances in psychometric techniques to bear

on these causal methods often used in other disciplines with little development in accounting for measurement error.

Drawing from theory and prior results on measurement error, factor scoring, and propensity score analysis described previously, I proposed the following hypotheses to be addressed in this study. First, I hypothesized that failing to model impact and DIF when they are present will yield degraded factor score estimates, thereby biasing causal effect estimates in addition to increasing their variability. Second, I hypothesized that using MNLFA and including background covariates will improve causal effect estimation by yielding factor score estimates that better emulate the true confounder addressed using propensity scores, thereby reducing the bias and variability of the effect estimate. Third, I hypothesized that in cases of uncertainty about the population impact and DIF effects, the detriments of over-modeling when impact and DIF are not present will be outweighed by the benefits of correct modeling when they are present, especially when more information about the LV is available. Finally, I hypothesized that the factor scores generated from models that most closely matched the data-generating model would yield the most accurate information about balance on the true latent confounding variable.

I used a simulation study to systematically test these hypotheses by varying the factors related to factor score recovery and examining their effects on effect estimates and balance in the confounders. In particular, I varied the presence of impact and DIF, the number of items, and the method of generating the factor score estimates in the context of a propensity score analysis of the treatment effect of a binary treatment on a continuous outcome, a situation that is common in causal studies in psychology.

CHAPTER 1: METHODS

Data Scenario

To provide concreteness to the simulation, it is helpful to consider a hypothetical observational study comparing the effect of a standard vs. an experimental after-school program for elementary school students on emotional wellbeing at the program's end. We might imagine this study spread over two sites implementing the same programs. Students are allowed to choose the program in which they take part, and we can imagine their choice depends on factors including site-specific characteristics, their age, and their level of an LV (e.g., depression) as measured by an error-prone psychological scale with binary items and known DIF such as the Child Behavior Checklist (CBCL; Achenbach & Edelbrock, 1981). In this example, our "treatment" is the experimental program, while the "control" is the standard program. The outcome is measured emotional well-being at the program's end, which itself is known to depend on a variety of factors, including those that influence selection into the chosen program[1]. We will imagine that age, site, and the LV constitute a sufficient set of variables to identify the causal effect of the experimental program relative to the standard one. We can assume one is interested in the ATE, the hypothetical average effect of moving all students from the standard program to the experimental program. We will imagine that levels and the measurement of the LV may be affected by each student's site and age, which is plausible given prior research (e.g., Curran et al., 2014).

[1] For the purposes of this simulation, it will be assumed that the outcome construct of interest is measured perfectly; however, psychological variables like this are often subject to measurement error.

I simulated data consistent with the path model in Figure 2. The data-generating model includes two components: a binary treatment (program) with a causal effect on an observed continuous outcome (well-being) confounded by an observed binary variable (site), an observed continuous variable (age), and a latent continuous variable (depression); and a measurement model for the LV that includes binary indicators, mean and variance impact from the observed confounders, and slope and intercept DIF from the observed confounders. The causal portion of the model is a setup common in propensity score simulations (e.g., Jakubowski, 2015; Nguyen et al., under review), while the measurement portion is similar to that used in Curran et al. (2016).

Latent and Observed Confounders

The LV is measured with a set of *k* binary items. I specified the corresponding measurement model as follows[2]: for $j = 1, \ldots, n$ individuals assessed on $i = 1, \ldots, k$ binary indicators, each indicator w_{ij} will follow a Bernoulli distribution with probability p_{ij} defined by the factor model as

$$p_{ij} = (1 + exp(-(\nu_{ij} + \lambda_{ij}\eta_j)))^{-1} \quad \text{(M1)}$$

where ν_{ij} is the intercept of item *i* for person *j*, λ_{ij} is the loading of item *i* for person *j*, and η_j is value of the LV for person *j*, distributed as $\eta_j \sim N(\alpha_j, \psi_j)$. This corresponds to a generalized linear model with a logit link and a Bernoulli response function, equivalent to a 2-PL IRT model when considering η_j as latent.

The observed covariates *site* and *age* were generated, respectively, as being drawn from a Bernoulli distribution with probability 0.6 and from a uniform distribution with bounds (-3, 3) (i.e., as if age were centered at its population mean for the group of interest). Although typically many more confounders and at least several moderators of the factor model would be present in

[2] Formulas in this section are numbered with a preceding letter "M" to distinguish them from statistical formulas used elsewhere.

an analysis with real data, here I limited my simulation to just two variables for simplicity of illustration, with the hope that results would generalize to situations with more variables and variables of types other than binary and uniform continuous. Substantive theory would typically drive the selection of these variables.

Measurement Moderation

I defined measurement moderation using the following models: I defined mean and variance impact, respectively, as

$$\alpha_j = \gamma_0 + \gamma_1 age_j + \gamma_2 site_j + \gamma_3 age_j \times site_j \quad \text{(M2)}$$

and

$$\psi_j = \beta_0 exp(\beta_1 age_j + \beta_2 site_j). \quad \text{(M3)}$$

These equations are instantiations of equations (14) and (15). By this specification, the mean (α_j) and variance (ψ_j) of the LV depend on (i.e., are impacted by) the levels of age and site; there is also an interactive effect of age and site on the mean (i.e., the effect of age on the LV mean is moderated by site). Equation (M3) for the LV variance ψ_j corresponds to a log-linear model so that the variance is bounded at 0 (Bauer, 2017; Bauer & Hussong, 2009). The parameters were chosen so that the marginal LV mean and variance were 0 and 1, respectively, in the population.

I defined intercept and loading (i.e., slope) DIF, respectively, as

$$\nu_{ij} = \kappa_{0i} + \kappa_{1i} age_j + \kappa_{2i} site_j \quad \text{(M4)}$$

and

$$\lambda_{ij} = \omega_{0i} + \omega_{1i} age_j + \omega_{2i} site_j \quad \text{(M5)}$$

The equations correspond to equations (16) and (17). As above, the intercept (ν_{ij}) and loading (λ_{ij}) each depend on the levels of site and age.

Treatment and Outcome Models

I specified the treatment effect model as follows, consistent with many propensity score simulations, notably Nguyen et al. (under review). I defined the treatment selection model as

$$\log\left(\frac{e_j}{1-e_j}\right) = a_0 + a_1 age_j + a_2 site_j + a_3 \eta_j \quad \text{(M6)}$$

where e_j is the probability of selecting the experimental program; that is, e_j is the "true" propensity score (i.e., the model-generated probability of treatment assignment for each individual). This model is consistent with a logistic regression model where selection probability is determined only by site, age, and the LV. Program selection (Z_j) is drawn from a Bernoulli distribution with probability e_j for each unit *j*. The coefficients are listed in Table 1 and were chosen so that age, site, and the LV uniquely explain 5%, 5%, and 10%, respectively, of the variance in the logit of the true propensity scores, and the marginal probability of treatment is .35. These reflect small to medium effect sizes for the observed covariates and a medium effect size for the LV (Cohen, 1988, p. 413). SMDs (defined in equation 9) were 0.52 (age), 0.39 (site), and 0.80 (LV); these indicate significant covariate imbalance, especially for the LV, given that absolute SMDs of 0.10 or lower are considered tolerable (Stuart, 2010).

I defined the outcome model as

$$Y_j = \tau Z_j + b_1 age_j + b_2 site_j + b_3 \eta_j + b_4 age_j \times \eta_j + b_5 site_j \times \eta_j + \varepsilon_j \quad \text{(M7)}$$

where $\varepsilon_j \sim N(0, \sigma^2)$. The purpose of including interactions in this model is to simulate a level of model complexity that applied researchers may not know to specify beforehand in their outcome model. A benefit of propensity score methods is that they yield unbiased results even when then outcome model is complex, so long as an appropriate propensity score model has been specified. Here there is no effect moderation on treatment and the potential outcome models for each unit

differ only on Z_j. Thus, τ is the causal effect of the experimental program on the outcome, and its estimate is the causal effect estimate that is the focus of the analysis of the simulations. The coefficients are listed in Table 1 and were chosen so that age and site each uniquely explained 6% of the variance in the outcome, the LV uniquely explained 12% of the variance in the outcome, and the interactions each uniquely explained 2% of the variance in the outcome, so that all variables jointly explained 34% of the variance in the outcome. Together with the treatment selection, these parameters created significant confounding by the covariates, yielding an unadjusted treatment effect estimate with a Cohen's *d* of approximately 0.6, indicating a moderate to large effect of treatment (Cohen, 1988, p. 40), when the true treatment effect was 0. The residual variance (σ^2) was chosen to scale and identify the effects of the covariates.

Design Factors

The design factors include the presence of impact (two levels), the presence of DIF (two levels), and the number of items (two levels), for a total of eight cells. Each cell was replicated 1000 times, for a total of 2 x 2 x 2 x 1000 = 8000 simulated data sets. Each data set contained 1000 individuals, which is reasonable for a two-site observational study and appropriate for both MNLFA and propensity score applications (e.g., Curran et al., 2017; Nguyen, Ebnesajjad, Stuart, Kennedy, & Johnson, 2018). Sample size will not be included as a varying design factor, given that its effects are predictable and Curran et al. (2016) found meager effects and no interactions with other design factors.

Impact. As exogenous predictors, it is expected that the observed and latent variables would naturally covary. I defined the presence of impact as covariate effects on the LV mean and variance above and beyond this covariance between the covariates and the LV. Regardless of the presence of impact in the simulation design, the covariates jointly explained 15% of the variance in the LV mean model (M2), yielding correlations between the LV and age of .33 and between

the LV and site of .21, which correspond to medium correlations commonly observed in practice (coefficient values are in Table 1). In the "impact absent" condition, the LV mean model contained only the linear terms, and the LV variance was constant across individuals. In the "impact present" condition, an interaction between site and age was included in the LV mean model, and covariate effects were also included in the LV variance model (M3) so that the variance of the LV differed for each individual based on their covariate values. The coefficients of the variance model are listed in Table 1 and were chosen so that across individuals, the inter-quartile range of the model-generated LV variances was 0.5, corresponding to a medium amount of variability in the variances due to age and site, which was enough to perturb estimates using traditional method methods in Curran et al. (2016) and differentiated the results between the "impact present" and "impact absent" conditions in pilot testing. To ensure equality across conditions, parameters were chosen so that the marginal mean and variance of the LV were approximately 0 and 1, respectively, in the population.

DIF. I defined the presence of DIF as covariate effects on the intercept and loading parameters of the indicator measurement models. In the "DIF absent" condition, the parameters of the measurement models were the same across individuals for each item. In the "DIF present" condition, these parameters differed across individuals based on age and site on half the items. The parameters are listed in Table 2 and were chosen to yield weighted area between curves (wABC)—a standardized measure of effect size for DIF—of approximately 0.3 on average, which has been considered a cutoff for delineating problematic DIF from unproblematic DIF (Edelen, Stucky, & Chandra, 2015). The items that had DIF in the "DIF present" condition were identical to those that did not have DIF so that any differences caused by DIF were attributable to the presence of DIF broadly and not the presence of DIF for a certain type of item. Item

endorsements ranged between .5 and .75. Item communalities, computed as the squared correlation between the LV and the continuous latent propensity underlying the item endorsement probabilities (Long, 1997), ranged from .23 to .7. The parameters and proportion of items with DIF are comparable to those found in empirical integrative data analysis applications, including Curran et al. (2014), and differentiated the results between the "DIF present" and "DIF absent" conditions in pilot testing. Specific values are in Table 2.

Number of Items. The number of items was either six or 12. As in many simulation studies, the effect of the number of items is straightforward: the more items, the better the precision of the estimates because of increased factor determinacy (Bollen, 2002). This was found in both Curran et al. (2016) and Jakubowksi (2015). The reason for including it here was to examine whether the potential loss in precision due to over-modeling impact and DIF when there are none is comparable to the gain in efficiency when increasing the number of items. Curran et al. (2016) also found in metamodels that the number of items interacted with other design factors, including the magnitude of impact and DIF, which are studied here.

Data Generation and Analysis

Data was generated in R (R Core Team, 2018) in the following sequential steps for each replication. First, values of site and age were generated from the distributions previously described. Next, values of the LV were generated using equations (M2) and (M3) to define the mean and variance of the LV. Next, the indicators of the LV were generated using equations (M1), (M4), and (M5). Next, the treatment variable was generated using equation (M6), and finally, the outcome variable was generated using equation (M7).

Effect Estimation Models

Within each simulation, several estimation models were employed. In all but the naïve model (defined below), the following steps occurred. First, a score was estimated to represent the LV. Second, a logistic regression model was fit with treatment status as the response and the observed covariates and the estimated LV score as predictors; this served as the propensity score model from which predicted probabilities were estimated as propensity scores. Third, the propensity scores were transformed into weights with the following formula:

$$w_j = \frac{Z_j}{\hat{e}_j} + \frac{1 - Z_j}{1 - \hat{e}_j}, \tag{M8}$$

where Z_j is the treatment status for individual j and $\hat{e}_j$ is their estimated propensity score. These weights are appropriate for estimating the ATE (Austin, 2011). Finally, an outcome regression model as specified in equation (10) was fit using the estimated weights for WLS estimation to acquire a treatment effect estimate for each data set. Additionally, using each set of weights, I computed the weighted SMD of the estimated LV score between the treated and control groups, and the same for the true value of the corresponding LV; these served as balance summaries for the estimated and true LVs.

The following are the methods that were employed in this simulation, which involve the basic, current best practice, and proposed improved methods.

Method 1: Naïve Model. In this method, neither scores for the LV nor propensity scores were estimated. A regression of the outcome on only the treatment variable (i.e., a t-test) was fit to attain a naïve treatment effect estimate with no adjustment for confounding. The goal of the subsequent methods was to arrive at improved treatment effect estimates relative to this unadjusted estimate; therefore, this model served as a baseline.

Method 2: Individual Items. In this method, no LV scores were estimated; the propensity score model was fit using the individual items and the observed covariates to predict the log odds of treatment, as follows:

$$\log\left(\frac{P(Z_j = 1)}{1 - P(Z_j = 1)}\right) = b_0 + b_1 age_j + b_2 site_j + \sum_{i=1}^{k} b_{i+2} X_{ij} \quad \text{(M9)}$$

where $X_{1j}, \ldots, X_{kj}$ are the k indicators for the LV. This approach was examined by Jakubowski (2015), Nguyen et al. (under review), and Raykov (2012) and represents an approach recommended by Steiner, Cook, and Shadish (2011) to improve propensity score estimation with potentially unreliable variables. This technique might be typically employed when several observed variables seem to measure the same construct but may not clearly fall within a known factor structure. To compute the balance summary for this method, I computed the weighted SMD for each item between the treated and control groups and then used the mean of these SMDs to serve as the balance measure.

Method 3: Simple FS. In this method, an unconditional generalized factor model (i.e., a 2-PL IRT model, not including the observed covariates) was fit to generate the factor scores. Figure 3A depicts a path diagram corresponding to the fitted model. The estimated factor scores $\hat{\eta}$ were computed using EAPs and were used in the following logistic regression model:

$$\log\left(\frac{P(Z = 1)}{1 - P(Z = 1)}\right) = b_0 + b_1 age_j + b_2 site_j + b_3 \hat{\eta}_j \quad \text{(M10)}$$

the predicted values of which formed the estimated propensity scores. SMDs of these factor scores were used in computing the balance summary. This method corresponds to the approach recommended by Jakubowski (2015) and Raykov (2012) and to the simple factor score method described in Nguyen et al. (under review). This method assumed equal factor loadings across individuals and an unconditional normal distribution for the LV. To identify the model, the LV

mean and variance were fixed at 0 and 1, respectively, and all intercepts and loadings were freely estimated.

Method 4: MNLFA Simple FS. I fit a conditional (i.e., MNLFA) generalized factor model to the items, modeling mean impact on the LV from site, age, and their interaction; variance impact on the LV from age and site modeled with a log-linear function (Bauer & Hussong, 2009); and intercept and loading DIF from age and site on half of the items. Figure 3B depicts a path diagram corresponding to the fitted model. When DIF and impact were present, this MNLFA model corresponded to the data-generating model for the LV and items; otherwise, this model involved over-modeling nonexistent impact or DIF. DIF was modeled for the "correct" indicators, i.e., those indicators that in the "DIF present" condition had DIF (and therefore over-modeling non-existent DIF for those same items in the "DIF absent" condition). To identify the model, the mean and variance at the reference levels of the covariates (i.e., at 0) were fixed at 0 and 1, respectively, and all loadings were freely estimated. Factor scores $\hat{\eta}$ were estimated from this model using EAPs of the conditional LV distribution and measurement functions as in Curran et al. (2016) and used to estimate propensity scores from the logistic regression model in equation (M10) and to compute the balance summary.

Method 5: Fully Inclusive FS. I fit a structural equation model linking the covariates and the LV to the treatment (including the items as indicators), and then generated factor scores for the LV from this model. In this way, the treatment is used as an indicator of the LV, and linear covariances between the observed covariates and the LV are estimated in the model. Figure 3C depicts a path diagram corresponding to the fitted model. This model follows the same structure as the fully inclusive factor model proposed by Nguyen et al. (under review). The same identifying constraints were used as in the Simple FS method. Because the structural equation

model includes estimating covariances between the observed and LVs, any linear impact on the LV mean by the observed variables is modeled. The estimated factor scores $\hat{\eta}$ were estimated using EAPs and were used in the logistic regression specified above in equation (M10) and to compute the balance summary.

Method 6: MNLFA Fully Inclusive FS. I fit a MNLFA model as in MNLFA Simple FS method but included the relationship between the covariates, LV, and treatment as in the Fully Inclusive FS method. Figure 3D depicts a path diagram corresponding to the fitted model. When impact and DIF were present, this model corresponded to the data-generating model for the LV, items, and treatment, and involved over-modeling covariate relationships with the LV and items otherwise. As with the fully inclusive factor model, the treatment functions as an indicator for the LV. The same identifying constraints were used as in the MNLFA Simple FS method. The estimated factor scores $\hat{\eta}$ were used in the logistic regression specified above in equation (M10) and to compute the balance summary.

Method 7: True LV. Finally, these methods were compared to a model that uses the true LV in place of estimated factor scores in equation (M10). In practice, this method would not be accessible to researchers because the true values of the LVs are not available. This will serve as a benchmark for the other methods, given that this method uses true LV values and therefore is (in theory) the best a method relying on estimated LV values could achieve. Given that the propensity score model was correctly specified, this method was expected to yield the least biased and most precise estimates, even more precise than a model that were to use the true (i.e., model-generated) probabilities of treatment assignment (Lunceford & Davidian, 2004).

Model Estimation

The factor score models were fit in M*plus* using the *MplusAutomation* package (Hallquist & Wiley, 2018). The propensity score estimation and effect estimation were performed in R. LV models were fit with maximum likelihood estimation with adaptive quadrature and 15 quadrature points per dimension using default start values and convergence criteria; Bauer (2017) and Curran et al. (2016) found these specifications to be effective. In the Simple FS and Fully Inclusive FS methods, the LV was scaled to have mean 0 and variance 1 in order to identify the model. In the MNLFA Simple FS and MNLFA Fully Inclusive FS methods, the LV was scaled to have mean 0 and variance 1 for the "reference" group (i.e., when site and age are both 0) (Bauer, 2017).

Criterion Variables

Six criterion variables were examined and compared across conditions and method: mean factor score correlations with the true LV, mean propensity score correlations with the optimal propensity score, mean percent bias remaining (PBR), root mean squared PBR (RMSPBR), mean true balance, and root mean squared balance discrepancy (RMSBD).

Mean factor score correlations with the true LV were computed for all methods that involved estimating a factor score. The Pearson correlation for each factor score type with the true LV values was computed for each replication and then averaged within cells. High correlations between the factor scores and the true LV indicate that the estimated factor scores using the given method represent the true LV values well.

Mean propensity score correlations with the optimal propensity score were computed for all methods that involved estimating propensity scores. The "optimal" propensity scores are those estimated using the correct model and true LV values. Note that these will differ from the true model-generated treatment probabilities; propensity scores estimated using the correct model

yield unbiased effect estimates with smaller variance than do true assignment probabilities (Lunceford & Davidian, 2004), and therefore provide a better benchmark for comparison of the propensity scores estimated without access to the true LV values[3]. High correlations between the estimated and optimal propensity scores indicate that the estimated propensity scores should function similarly to the true propensity scores in arriving at covariate balance and therefore an unbiased estimate of the treatment effect. Although no hypothesis implied this criterion, I included it to help explain the patterns of results found.

The mean PBR was computed for all methods; for each method within each replication, PBR was computed using the following formula:

$$\text{PBR } = 100 * (\hat{\tau} - \tau_{pop})/(\hat{\tau}_{naïve} - \tau_{pop})$$

where $\hat{\tau}$ is the estimated treatment effect for a given method, τ_{pop} is the treatment effect in the population (here, 0), and $\hat{\tau}_{naïve}$ is the treatment effect estimated using the Naïve method (i.e., the raw difference in group means)[4]. A PBR of 100 means that the estimated treatment effect was as biased as the naïve estimate, and no bias was removed; a PBR of 0 means that the effect estimate was perfectly unbiased (i.e., equal to the population treatment effect); PBRs between 100 and 0 mean that not all bias was removed using the adjustment method; and PBRs less than 0 mean that there was overcorrection (i.e., bias in the direction opposite to that of the naïve estimate). The mean PBR was computed for each method and cell of the design.

The RMSPBR was computed as the square root of the mean of the squared PBRs for each

[3] The pattern of results was essentially unchanged when using the model-generated treatment probabilities as a benchmark.

[4] I used PBR rather than raw bias because the initial bias differed between cells due to the effects of impact on the covariance among the covariates; PBR ensures that all bias reduction is placed on the same scale. In addition, using proportion-based criteria (akin to measures of relative bias) allows for the interpretation of bias and variability in a way separated from the arbitrary scale of the variables used in this simulation.

method and design cell, and functions similarly to a traditional root mean squared error (RMSE), in that it indicates the typical distance of each effect estimate from the population effect and is on the same scale as the PBR. RMSPBRs close to 0 indicate that PBRs were typically low, and the effect estimate was often close to the population effect. For mean PBRs and RMSPBRs, I considered differences of greater than 1 between cells to be statistically meaningful on the context of this simulation[5]. All PBR and RMSPBR values are reported in Table 3 and in the subsequent text as whole numbers for clarity. I omit "%" in the reporting of the results, but the mean PBR should be interpreted as a percentage, and the RMSPBR, while not a percentage, should be interpreted on the same scale as the mean PBR (i.e., as percentage points).

For each method, estimated and true balance summaries were created. For the methods that involved estimating a factor score, the estimated balance summary was the weighted SMD of the factor scores, computed as in equation (9). For the Items method, which did not involve estimating a factor score, the average weighted SMD of the items was used as the estimated balance summary. For all methods, the true balance summary was the weighted SMD of the true LV using the weights estimated with that method. For each method and design cell, the average true balance summary was computed to examine the degree to which the estimated weights balanced the true LV; this are reported as the mean SMD of the true LV. Values close to 0 indicate good balance in the variable means between the treatment groups; authors recommend ensuring weighted SMDs below 0.1 before proceeding with effect estimation.

In addition, the RMSBD was computed as the square root of the mean squared difference

[5] For each method, I computed an analogue to the pooled standard error of the PBR and multiplied by 2; for all methods, this value was slightly less than 1. Differences in mean PBRs greater than 1 are thus greater than two standard errors away from each other, which can be considered evidence that the mean values in question differ statistically. In practice, mean PBR differences of 1 would not be meaningful; I consider them meaningful here to highlight the patterns observed in the results.

between the true and estimated balance summaries for each method and cell; this value represents the typical difference between the estimated balance summary, which would be accessible to researchers, and the true balance summary, which is normally inaccessible but which is desired. Values of the RMSBD close to 0 indicate that the estimated balance summary is similar to the true balance summary and might be used as a proxy for measuring balance on the true LV. Values of the RMSBD larger than 0.1 are particularly problematic because they indicate that even after achieving perfect balance based on the estimated balance summary, problematic imbalances in the true LV are likely to remain.

To my knowledge, these criteria have not been used to evaluate simulation results in the context of propensity score analysis and were developed for this study in particular, though similar values to the mean PBR have appeared sparingly (e.g., Shadish, et al., 2008; Hall, Steiner, & Kim, 2015). They reflect a compromise to satisfy the desire to provide familiar simulation criteria (e.g., bias, RMSE) while addressing the issues of differing baselines across cells and criteria not on typical or inherently meaningful scales. Although the overarching patterns hold when more standard simulation criteria are used, for the purposes of a fair comparison across methods and conditions, I have chosen to describe the results of this simulation using these new criteria.

CHAPTER 2: RESULTS

First, I discuss issues of model convergence and aberrant estimates. Next, I discuss the quality of the factor scores and propensity scores. Next, I discuss the bias and variability for each method in terms of mean PBR and RMSPBR across conditions. Finally, I discuss the balancing performance of the methods with respect to the mean true balance and RMSBD.

Convergence and Aberrant Estimates

I fit a total of 32,000 factor models across all replications and conditions: four scoring models (simple FA model, simple MNLFA model, fully inclusive FA model, and fully inclusive MNLFA model) fit to 1,000 replications within each of 8 cells[6]. Fourteen models failed to converge, yielding model parameter estimates that were not true maximum likelihood solutions: nine occurred in estimating the simple MNLFA models, and five occurred in estimating the fully inclusive MNLFA models. The failures only occurred when DIF was present and with six items. I omitted results originating from these models from all subsequent analyses.

All replications were checked for impossible or unusual parameter values, outlying data points, and extreme effect estimates. Although there were some unusual parameter estimates and effect estimates, these were not deemed to be worthy of removal because they appeared to be the result of natural randomness or estimation uncertainty that was relevant to the analysis.

[6] Naïve, item, and true score models were also fit but were not at risk for failure to converge.

Score Quality

I computed correlations between the estimated factor score and the true LV values and between the estimated propensity scores and the optimal propensity scores to inform variations in other criteria presented below. Correlations close to 1.00 indicate high score quality, suggesting performance of the scores close to their respective benchmarks. First, I discuss the quality of the estimated factor scores, and then I discuss the quality of the estimated propensity scores. Across both sets scores, patterns were similar regardless of the number of items, so only the six-item conditions will be discussed.

Factor Scores. The distributions of correlations between the estimated factor scores and the true values of the LV are displayed in the top panel of Figure 4. The largest differences were between the unconditional factor scores and the MNLFA-based factor scores. In the absence of impact and DIF, all factor score performed similarly with mean correlations around .83, indicating fairly good agreement with the true LV values. In the presence of impact and DIF, the unconditional factor scores had mean correlations around .77, while the MNLFA-based factor scores had mean correlations around .84. Similar results were found in Curran et al. (2016), in which MNLFA-based factor scores were more highly correlated with true LV values than were unconditional scores in the presence of impact and DIF, especially when impact and DIF effects were strong. The fully inclusive factor scores were consistently slightly more correlated with the true factor scores on average (by a small margin of approximately .005), as these factor scores were effectively computed with an additional correctly modeled indicator (i.e., the treatment).

Propensity Scores. The distributions of correlations between the propensity scores estimated with the approximate methods and the propensity scores estimated with true values of the LV are displayed in the bottom panel of Figure 4. Mean correlations ranged from .94 to .97, indicating excellent agreement with the optimal propensity scores. There was little variability in this range,

though some patterns emerged: overall, the simple FS and MNLFA simple FS methods yielded higher mean correlations than did using the items or the fully inclusive methods. Across conditions, the MNLFA simple FS yielded higher mean correlations than the other methods, ranging from .96 in the absence of impact and DIF to .97 in the presence of both. The unconditional fully inclusive FS consistently yielded the lowest mean correlations, ranging from .94 in the presence of impact and DIF to .95 in the absence of both, despite the method's high performance on the other metrics described below.

Bias and Variability of Effect Estimates

Mean PBR and RMSPBR values are displayed in Figure 5 and detailed in Table 3. Across cells and models, there was substantial variability in mean PBRs: values ranged from -2 to 13 (i.e., between -2 and 13 percent of the original bias remaining). On the other hand, RMSPBRs were not highly variable: values ranged from 12 to 19 (i.e., the typical PBR was between 12 and 19 percentage points from 0 percent). Given that mean PBRs for some methods and cells were close to 0, these methods were able to yield mostly unbiased estimates. Performance of each method on these metrics in the various conditions of impact, DIF, and number of items is described below. First, I describe the performance of the simple methods (those that do not involve including the treatment in the factor score model), and next I describe the fully inclusive methods.

Items. Using the items involved eschewing a factor score model and simply including the items in the propensity score model along with the observed covariates. This model was incorrect in assuming the true LV can be adequately represented by an optimally predictive weighted sum of the items. When impact or DIF were present, some degree of the induced covariance between the items and the covariates was accounted for in the propensity score model, which freely allowed the items and covariates to covary as exogenous predictors of

treatment.

With six items and no impact or DIF, using the items directly yielded a mean PBR of 13, indicating significantly biased effect estimates. Surprisingly, the presence of impact or DIF slightly *improved* bias removal; with either impact or DIF present, the mean PBR was 11, and with both, the mean PBR was 12. The opposite pattern emerged with 12 items: the least bias was observed when neither impact nor DIF were present (mean PBR = 7), and the greatest bias was observed when both impact and DIF were present (mean PBR = 9). Regardless of the presence of impact or DIF, using the items yielded biased effect estimates, especially with fewer items.

The RMSPBR largely remained the same or decreased in the presence of impact or DIF. With six items, RMSPBR was 19 in the absence of impact and DIF, 18 in the presence of only DIF, 17 in the presence of only impact, and 17 in the presence of both impact and DIF, indicating large typical differences between the estimated treatment effect and the true effect. With 12 items, RMSPBR was 16 in the absence of impact (regardless of DIF) and 15 in the presence impact (regardless of DIF). The bias and RMSPBR patterns were in approximately the same direction with six items, but in opposite directions with 12 items.

Simple Factor Score. The simple FS method involved estimating factor scores from an unconditional measurement model that included the items as indicators of the LV (Figure 3A). In the presence of impact, this method was incorrect in assuming an unconditional normal distribution for the LV when in reality the distribution was conditional on the covariates[7]. In presence of DIF, this method was incorrect in assuming equal intercepts and loadings for each item across all units, when in reality these parameters of the measurement equations varied

[7] Note that in the "Impact absent" condition, the LV distribution was also normal only conditional on the covariates, but only linear covariances were present in contrast to the nonlinear relationships present in the "Impact present" condition.

across units based on their covariates values.

The bias and RMSPBR results for the simple FS method were almost identical to those from the items, echoing the results of Nguyen et al. (under review), and indicating similarly significant bias in the effect estimates. With six items, the mean PBR was 13 in the absence of impact and DIF, 12 in the presence of only DIF, 11 in the presence of only impact, and 13 in the presence of both impact and DIF. With 12 items, the mean PBR was 7 in the absence of impact and DIF, 9 in the presence of only DIF, 8 in the presence of only impact, and 9 in the presence of both impact and DIF. The RMSPBR results were identical to those using the items method, except that in the presence of only impact with six items, the simple FS method yielded an RMSPBR of 15. In general, using the simple factor score yielded similar bias and RMSPBR to using the items; when the methods differed, the simple FS method always yielded increased bias compared to using the items, but these differences were only by one point[8]. The relatively high mean PBR and RMSPBR results indicate significant bias in treatment effect estimates and that effect estimates were typically quite far from the true effect.

MNLFA Simple Factor Score. The MNLFA simple FS incorporated the covariate effects on the LV distribution and measurement equations into the scoring model but did not include the treatment variable as an indicator (Figure 3B). When both impact and DIF were present, this method matched the data-generating process for the LV and its indicators; when either were absent, this method over-modeled nonexistent relationships between the covariates and the LV distribution and measurement equations.

Overall, the MNLFA simple FS method yielded mean PBR and RMSPBR values similar to or slightly less than those from using the items or using the (unconditional) simple FS. In the

[8] Though Figure 5 depicts some other differences between the simple FS method and using the items, these differences were small enough not to be considered meaningful.

absence of impact and DIF, the mean PBR was 13 with six items and 7 with 12 items, equal to those found with the previous two methods, reflecting significantly biased effect estimates. In the presence of either impact or DIF (or both), the MNLFA simple FS method yielded a mean PBR of 11 of with six items and a mean PBR of 7 with 12 items, representing modest but consistent improvements of between 0 and 2 points over the unconditional simple methods, which ranged from 11 to 13 with six items and from 8 to 9 with 12 items. This pattern of results indicates slight improvements in bias removal using the MNLFA Simple FS relative to the unconditional simple methods, but also that effect estimates using this method remained biased, despite using a scoring model at least as flexible as the data-generating model.

As was seen in the unconditional methods, RMSPBR was lower when impact or DIF were present (even if incorrectly modeled): with six items, the RMSPBR was 19 in the absence of impact and DIF, 17 in the presence of only DIF, and 16 in the presence of impact (regardless of DIF), and with 12 items, RMSPBR was 16 in the absence of impact and DIF, 15 in the presence of only DIF, 14 in the presence of only impact, and 13 in the presence of both. These values were slightly but consistently lower than those of the unconditional simple methods by between 0 and 2 percentage points when either impact or DIF were present. All improvements over the unconditional simple methods were modest, and the MNLFA simple FS method still yielded biased effect estimates, especially with fewer items.

Fully Inclusive Factor Score. The fully inclusive FS method involved including the treatment variable as an indicator in the scoring model and modeling the covariances between the observed covariates, the LV, and the treatment; however, mean impact by the covariate interaction, variance impact, and DIF were not modeled (Figure 3C). Nguyen et al. (under review) found this method to be highly effective in removing bias in the absence of impact or

DIF.

Here, the fully inclusive FS method removed essentially all the bias, regardless of the presence of impact or DIF or the number of items. With six items, the mean PBR was 0 in the absence of DIF (regardless of impact) and 2 in the presence of DIF (regardless of impact). With 12 items, the mean PBR ranged from -1 to 2, with a slight increase in the presence of impact, regardless of DIF. The RMSPBR decreased when impact or DIF were present, regardless of the number of items. The presence of impact and DIF together lowered the RMSPBR from 17 to 14 with six items and from 16 to 13 with 12 items, with intermediate values in the presence of either one, reflecting typical effect estimates closer to the true effect, a pattern found also with the simple methods.

MNLFA Fully Inclusive Factor Score. The MNLFA fully inclusive FS incorporated the covariate effects on the LV distribution and measurement equations into the scoring model and also included the treatment variable as an indicator and covariate effects on the treatment (Figure 3D). When both impact and DIF were present, this method matched the data-generating process for the LV, the indicators, and treatment; when either were absent, this method over-modeled nonexistent relationships between the covariates and the LV distribution and measurement equations.

With either six or 12 items and with no impact (regardless of the presence of DIF), the MNLFA fully inclusive FS method removed almost all bias (mean PBR between -1 and 1). In the presence of impact, however, with six items the mean PBR was 5 (regardless of DIF) and with 12 items was 4 in the absence of DIF and 3 in the presence of DIF. With respect to RMSPBR, the MNLFA fully inclusive FS consistently yielded similar or smaller values than the unconditional fully inclusive FS, regardless of the presence of impact or DIF or the number of

items. With six items, the RMSPBR was 17 in the absence of impact and DIF, 16 in the presence of only DIF, and 13 in the presence of impact (regardless of DIF). With 12 items, the RMSPBR was 16 in the absence of impact and DIF, 15 in the presence of only DIF, 13 in the presence of only impact, and 12 in the presence of both.

Notably, in the presence of impact, the MNLFA fully inclusive FS method had similar RMSPBRs than using the true LV values (which yielded RMSPBRs of 13 in the presence of impact). In addition, the pattern of mean PBR was opposite to that of RMSPBR: in the presence of impact and DIF, for which the analysis model matched the data-generating model, mean PBR was higher but RMSPBR was lower than in the absence of impact and DIF.

True LV Values. The true LV values were free of measurement error and therefore were expected to yield unbiased and efficient estimates (relative to the other methods), with any error resulting from sampling error and the natural inefficiency of propensity score weighting. Indeed, mean PBRs ranged from -1 to 1, indicating unbiased effect estimates, and RMSPBRs ranged from 13 to 15. As with the methods above, RMSPBRs were lowest (RMSPBR = 13) in the presence of impact, which might provide some context for these patterns.

Covariate Balance

Given that applied researchers rely on balance statistics to make decisions about the appropriateness of their balancing procedure, it is important to understand the relationship between the balance statistics computed on the observed variables and the balance statistic that would be computed on the unobserved true variables were they to be observed. I examined whether balancing on the observed factor scores yielded balance on the true values of the LV and how far off the balance statistics computed on the observed variables were from the balance statistics computed on the true variable.

Balancing Performance. Without any adjustment, on average the true LV was imbalanced according to the SMD (by design). Though nearly all of the methods were able to achieve balance on the observed manifestation of the LV, only the methods that included the treatment as an indicator were able to also balance the true LV. Figure 6 displays the overall balancing performance of each method for the case with six items, stratified by the presence of impact. Discrepancies between the median estimated balance summary (left-facing densities) and the true balance summary (right-facing densities) indicate the failure of the simple methods to achieve satisfactory balance on the true LV.

In general, the pattern of true balance summaries (the LV mean SMDs) followed that of the mean PBRs in the effect estimates, which is to be expected given that imbalance in the true LV produces bias. The methods that did not include the treatment as an indicator yielded true mean SMDs ranging across cells from 0.181 to 0.232 with six items and from 0.122 to 0.170 with 12 items, indicating a failure to achieve satisfactory balance based on the usual criterion of 0.100. For these methods, the best balance was achieved when both impact and DIF were present and correctly modeled (i.e., using the MNLFA simple FS; mean SMD = 0.181 for six items) and when only impact was present but not modeled (i.e., using the items or simple FS; mean SMDs = 0.182 and 0.183, respectively, for six items). For each cell, increasing the number of items consistently decreased mean SMDs. Including MNLFA in the score model reduced mean SMDs by approximately 0.02 in the presence of DIF and had little effect in the absence of DIF (regardless of impact).

The methods that did include the treatment as an indicator yielded superior balance on the true LV. All SMDs were below .100, even when the incorrect scoring model was used. In general, using the fully inclusive FS yielded better balance than using the MNLFA fully

inclusive FS in the presence of impact (e.g., mean SMD = 0.017 vs. 0.063, for six items with impact and DIF both present), mirroring the mean PBR results (mean PBR = 2 vs. 5, for the same design cell). The MNLFA fully inclusive FS, however, did yield satisfactory balance on the true LV, and yielded slightly better balance than the unconditional fully inclusive FS when impact was not present.

Indicating Balance on the LV. Although balance on the LV was achieved on average to varying degrees with these methods, another question is the degree to which the observed balance statistics provide information about the true balance of the LV. Although the estimated balance summaries of the simple methods mostly fell between -0.1 and 0.1, the true balance summaries mostly fell outside this range and had little overlap (Figure 6). On the other hand, the estimated balance summaries of the fully inclusive methods largely overlapped with the true balance summaries.

RMSBDs represent the typical difference between the estimated and true balance summaries within each cell. Larger values imply that the estimated balance summary is not a good indicator of balance on the true LV, while small values imply that the estimated balance summary is an adequate proxy for balance on the true LV. For the simple methods, RMSBDs ranged across cells from 0.185 to 0.254 with 6 items and from 0.128 to 0.175 with 12 items, indicating large discrepancies between the estimated and true balance summaries. In general, the MNLFA simple FS yielded estimated balance summaries that were closest to the true balance summaries, especially when either impact or DIF were present.

In contrast, for the fully inclusive methods, RMSBDs ranged across cells from 0.054 to 0.086 with 6 items and from 0.041 to 0.066 with 12 items, indicating that estimated balance summaries from these methods are trustworthy indicators of the balance on the true LV. In general, the

MNLFA fully inclusive FS yielded estimated balance summaries that were closer to the true balance summaries than did the fully inclusive FS, especially when impact was present.

In sum, the fully inclusive methods both yielded better balance on the true LV and provided more information about balance on the true LV. Although the patterns of balance in the true LV were very in line with those of the mean PBR, conditional on the inclusiveness of the scoring model, the estimated balance summaries yielded the best information about the true LV balance when the scoring model matched or was more flexible than the data-generating model.

CHAPTER 3: DISCUSSION

The greatest distinction among methods was whether they included the treatment as an indicator or not; those methods that did so (i.e., the fully inclusive methods) performed almost universally better than the simple methods across the criteria studied (i.e., they had lower mean PBR, lower RMSPBR, better mean balance, and lower RMSBD). Among the simple methods, the MNLFA simple FS method tended to yield lower mean PBRs, lower RMSPBR, better balance, and lower RMSBDs than using the items or the unconditional simple FS method. Although patterns of results varied in the presence of impact or DIF, the performance of MNLFA simple FS method on the above metrics was either unaffected or improved by the presence of impact or DIF. Of note, though, is that by some metrics, the performance of the unconditional methods improved in the presence of impact or DIF in some conditions relative to the absence of impact and DIF, though these same improvements were not manifested in the factor score or propensity score correlations.

Between the fully inclusive methods, the relative performance of the MNLFA-based and unconditional FS methods varied based on the presence of impact or DIF and the metric considered. In general, the unconditional fully inclusive FS method yielded lower mean PBRs and better mean balance than the MNLFA fully inclusive FS method, but the MNLFA-based method consistently yielded lower RMSPBRs and lower RMSBDs than the unconditional method, in some cases performing as well or slightly better than using the true LV values. The performance of the unconditional method was largely unaffected by the presence of impact or DIF; to the degree it was affected, impact in particular was associated with slightly increased

mean PBRs, decreased RMSPBR, worse balance, and greater RMSBDs. Similar patterns were found for the MNLFA-based method, except that some of these patterns were amplified, and RMSBDs decreased in the presence of impact. Factor score correlations were very slightly higher with the fully inclusive methods, but propensity score correlations were slightly lower.

In the following sections, I evaluate the hypotheses proposed earlier given the results described above. In addition, I discuss some possible reasons for the observed variation, especially between the simple and fully inclusive methods. Finally, I provide recommendations, note limitations of the study, and discuss future avenues of research.

Hypothesis 1: The presence of impact or DIF will yield biased effect estimates when using the standard estimators

My first hypothesis followed from conventional knowledge about model misspecification. I expected the standard (i.e., unconditional) estimators to be more biased in the presence of unmodeled impact or DIF. It is known in the factor score regression literature that estimated factor scores can yield unbiased regression coefficients when estimated with the correct measurement model (Lu & Thomas, 2008; Skrondal & Laake, 2001), but impact and DIF can yield degraded factor scores when estimated with conditional (i.e., incorrect) models (Curran et al., 2016). Millsap's (1995) Duality Theorem implies that unmodeled variance impact on the LV (at least in the absence of DIF) should yield biased estimated relationships between factor scores and its sequelae (i.e., the treatment). It is thus surprising that in some conditions and with some metrics, the simple methods were not negatively affected by the presence of impact or DIF. It is possible that the DIF and impact effect sizes used in this simulation were too small to elicit large effects that may have been in a more predictable direction, but the pattern of factor score correlations (Figure 4, top panel) does indeed indicate significant degradation of the estimated factor scores in the presence of the studied levels of impact and DIF.

Between the simple unconditional methods, there was little advantage to using the simple factor scores over the items, regardless of the presence of impact or DIF or the number of items. This may result from the increased flexibility afforded by using the items at the expense of providing incorrect regression coefficient estimates due to measurement error. It is possible that the inclusion of the covariates in the propensity score model along with the misspecified factor scores reduced or eliminated the effects due to the misspecification of the factor model, perhaps by compensating for the otherwise under-modeled covariances between the factor scores and covariates. This might explain the slightly improved performance of the items-based estimator relative to the simple factor score FS estimator in the presence of DIF: some of the DIF-induced covariance between the indicators and covariates is implicitly modeled in the propensity score model when the items are included.

One of the most surprising findings was that the unconditional fully inclusive FS method was largely unbiased in the presence of impact or DIF. Despite using an incorrect factor model that yielded highly degraded factor scores and propensity score slightly further from the optimal propensity scores, the fully inclusive FS method yielded unbiased effect estimates with relatively lower variability and fairly reliable information about true balance in the presence of impact or DIF, at least under the conditions studied here. This is good news for those employing this method, which fared very well in the simulation of Nguyen et al. (under review) as well. Possible reasons for this are discussed later, though it is clear that including the treatment as an indicator increased this method's robustness to misspecification due to impact and DIF, more so than did including covariate relationships with the LV in the scoring model (based on the performance of the MNLFA-based simple FS method, described below).

In sum, my hypothesis about the standard estimators in the presence of impact and DIF was

mostly incorrect. Unmodeled impact and DIF had small effects on the quality of the estimators despite reductions in the quality of the factor scores, indicating that individual differences in the measurement of the LV are not a major source of error in effect estimation. Although model misspecification can have serious consequences in structural equation modeling, it may be that the propensity score weighting procedure provided enough opportunities for bias cancellation or compensation by the included covariates that the misspecifications studied here were largely irrelevant.

Hypothesis 2: Incorporating MNLFA into standard estimators will yield improved estimates when impact or DIF are present.

Given that correctly modeling the indicator relationships with the LV will improve the correlation between the estimated factor scores and true LVs (Curran et al., 2016), I expected improved performance over the standard methods when incorporating those relationships with MNLFA. Correctly modeling these relationships allows the estimated posterior of the conditional LV distribution to center nearer to the true distribution, which in turn should improve factor score recovery. Indeed, these results were found here: while the unconditional methods yielded degraded score recovery, the MNLFA-based methods yielded excellent score recovery, especially when impact and DIF were both present and modeled, but also even when either one was absent (Figure 4, top panel).

Though we have seen that factor score quality as measured by its correlation with the true LV is not necessarily a good indicator of its success in yielding unbiased effect estimates, among the simple methods, the MNLFA simple FS saw reductions in bias and variability over the unconditional simple methods. Though impact and DIF did not negatively affect the performance of the unconditional simple methods, they instead provided an opportunity for improvement when correctly modeled with MNLFA. Correctly modeling the covariate relationships with the

LV and its indicators reduces the indeterminacy of the factor because a larger proportion of the variance in the indicators and the LV itself are explained by the included covariates (Bollen, 2002). Given that the two elements of the scoring algorithm are the (conditional) expectations of the item responses and the (conditional) LV distribution, correctly modeling these by conditioning on the covariates should improve precision and reduce bias, in the same way including control variables improves the precision and reduces the bias of the estimation of a focal variable on an outcome in regression by reducing the mean squared error. Similarly, the link between LV methods and missing data procedures described by Mislevy et al. (1995) would imply that including auxiliary variables in factor models would improve parameter and score recovery.

Unexpectedly, the pattern of decreasing bias when modeling impact and DIF was reversed with the fully inclusive methods. Modeling impact and DIF when it was present yielded effect estimates with somewhat *greater* bias than those estimated with the unconditional fully inclusive method. The reasons for this are not immediately clear; both factor score and propensity score correlations improved in the presence of impact and DIF when using the MNLA fully inclusive FS method over the unconditional fully inclusive FS method (Figure 4), and yet the true LV retained more imbalance and therefore more bias remained. One clue is that the *estimated* factor scores were not well balanced by the propensity score weights, especially in the presence of impact (Figure 6), even though weighting typically balances the (observed) variables that entered the propensity score model. It may be that irregularities in the estimated propensity scores not observable from their correlations with the optimal propensity scores were responsible for their failure to balance the scores and true LV; the effects of these irregularities may have been (at least partially) masked when impact and DIF were not modeled along with the fully inclusive

factor model (e.g., due to bias cancellation). Perhaps enforcing exact balance on the MNLFA fully inclusive factor scores (using, e.g., entropy balancing; Hainmueller, 2012) would yield better balance on the true LV, thereby yielding less biased effect estimates.

Overall, there were benefits to modeling impact and DIF with MNLFA when they were present due to improved score recovery caused by increased precision (i.e., decreased indeterminacy) in the scores. Despite being more precise, the MNLFA fully inclusive FS estimator was more biased than its unconditional counterpart, though the reasons were not clear based on standard methods of score quality assessment and require further investigation.

Hypothesis 3: Using MNLFA when impact and DIF are not present will yield unbiased but imprecise results due to over-modeling

In this study, over-modeling occurred when either bias or DIF were absent but were modeled with MNLFA. Although it is often possible to identify the presence of impact and DIF using DIF detection techniques (e.g., Bauer, 2017), no technique will perfectly capture all such relationships in the population. Thus, it was important to determine whether there were consequences of being overly liberal when modeling impact and DIF (while maintaining a tractable scoring model), especially in the attempt to uncover a general recommendation for researchers that allows for agnosticism about the data-generating process.

Overall, the MNLFA-based methods performed no worse than the unconditional methods in the absence of impact, DIF, or both in terms of bias, variability, and validity of the balance summary. Although overfitting can reduce the precision of estimates by inappropriately responding to noise in the sample (Forster, 2000), there were no such problems here. Overfitting is often maligned for its failure to generate replicable estimates or good out-of-sample predictions (e.g., Camstra & Boomsma, 1992), but because the step in the analysis that involves overfitting is one in which inferences are not made and predictions are made only in-sample,

capturing the unique qualities of the sample at hand did not impede the inferences made in the final step.

This is encouraging to those who might otherwise avoid a needlessly complex model in favor of parsimony; in fact, there were no gains to doing so given the simulation conditions. When a more complex model excludes (i.e., constrains) relationships that would be otherwise present in a simpler model (i.e., in order to estimate other relationships), bias can indeed occur (Kaplan, 1988), though this simulation did not examine these types of misspecification. All data-generating models that excluded at least one of impact and DIF were nested within the MNLFA-based factor models; future work is needed to examine the effects of model-misspecification when this is not the case.

Simple vs. Fully Inclusive Methods

Though there were some improvements in effect estimation that came with correctly modeling impact and DIF, clearly the most important factor was whether the factor score model included the treatment as an indicator. Although such effects were found by Nguyen et al. (under review), it was unexpected that the unbiasedness of this method would extend to the case of unmodeled impact and DIF. Indeed, deeper examination of the performance of the fully inclusive methods at various steps yields somewhat paradoxical conclusions, described subsequently.

There are a number of reasons why the fully inclusive method would be expected to perform well; these include that an additional indicator is used to measure the LV, thereby slightly increasing the LV's reliability, and that the (linear) covariate relationships with the LV are also modeled. Nguyen et al. (under review) found that it was important for the covariate relationships with the LV and treatment to be correctly modeled; failing to model these relationships (what the authors called the "partially inclusive" method) yielded biased effect estimates when the covariates were related to the LV, likely due to remaining covariance between the LV and the

treatment in the factor model. Correctly modeling the covariate relationships with LV was not enough alone, though, as demonstrated here by the poorer performance of the MNLFA simple FS method, which did correctly model these relationships, relative to the unconditional fully inclusive FS method when DIF and impact were present. In addition, the gains of an additional indicator do not explain these differences in performance: even with six *more* indicators and with covariate effects correctly modeled, the MNLFA simple FS estimator with 12 items was more biased than the unconditional fully inclusive FS estimator with only six items (although it should be noted that the former had less variability as measured by the RMSPBR). The score quality of the unconditional fully inclusive factor scores was lower than that of the MNLFA simple factor scores, indicating that it is not score quality that yielded these effects.

One possibility is the increased role of the propensity score model in the posterior distribution of the LV. By modeling the treatment as an indicator, one is using the propensity score model as one of the indicator measurement equations; the propensity score is the conditional probability of "item endorsement" (i.e., treatment receipt) given the covariates and LV value. The factor scores are computed including the propensity score function in the posterior distribution of the LV and are then used again in the observed-score propensity score model. The outcome (i.e., of the propensity score model; the treatment) is used as a component of its own predictor, artificially increasing the covariance between the predictor and the outcome. For example, in predicting loneliness from depression, using a depression scale that included an indicator related to loneliness would artificially increase the relationships between the scale and loneliness. One might expect better predictions of the outcome from such a model. In this case, this would amount to the fully inclusive factor scores producing propensity scores more highly correlated with true or optimal propensity scores.

In fact, this is not the case: the propensity scores estimated from the fully inclusive factor scores were *less* correlated with the optimal propensity scores than were the propensity scores estimated from MNLFA simple factor scores (or from any other model; Figure 4). Given that the estimation of the propensity scores is the penultimate step to arriving at an effect estimate, it remains unclear why the fully inclusive method would perform so well, especially with a misspecified model, given the low quality of its factor scores and propensity scores. Likewise, it remains unknown why this performance was not seen to the same extent with the MNLFA fully inclusive FS method, in which the factor score model matched the data-generating model in the presence of impact and DIF, but for which bias remained in the effect estimate.

Limitations

This study was one of few to examine a LV solution to the problem of measurement error in propensity score analysis. It is impossible to fully represent all possible data scenarios that might arise using a simulation study, and only a few such scenarios were explored here. The magnitude and form of impact and DIF did not differ, but it is possible that the magnitude of impact and DIF can change the quality of the score estimates (Curran et al., 2016) and therefore possibly the effect estimates. Here, I focused on a single plausible value for each impact and DIF parameter primarily to explore their effects in scenarios applied researchers might find; the parameters were not pushed to their limits. In addition, a simple data-generating model was used, and, except in the cases of unmodeled impact and DIF, there was no structural model misspecification. Given that misspecification in some parts of the model can have broad-reaching effects (Bollen, Kirby, Curran, Paxton, & Chen, 2007), and the fully inclusive methods relied on correctly modeled structural relationships, it would be interesting to examine how these methods would fare when these relationships are incorrectly specified, such as when residual covariances between items are inappropriately excluded from the scoring models. The goal here was to present a best-case

scenario to demonstrate the adequacy of the proposed MNLFA-based methods, since their efficacy in this context had not previously been explored.

Although estimation bias and variability were examined, there was little focus on inference here, in that standard errors were not estimated for any methods and power and Type I error rates were not examined. Standard error estimation for propensity score-based estimators is itself an ongoing area of research, though many researchers have agreed that sandwich-based standard errors (White, 1980) are an adequate solution (Robins, 2000). Nguyen et al. (under review), however, recommend using bootstrapping for inference, but such an approach may be too computationally intensive to be feasible with the highly structured scoring models studied here. It may be possible to combine commonly used standard errors for propensity score-based estimators with those that account for the uncertainty at multiple steps of the analysis, including the factor score estimation, as in Skrondal and Kuha (2012).

Finally, the only method examined was propensity score weighting, though other propensity-score based and related methods exist for estimating treatment effects, including matching, stratification, propensity score ANCOVA, and combinations of these methods with regression. In addition, I did not examine more robust methods of estimating balancing weights, including those that provide exact balance on covariates (e.g., entropy balancing; Hainmueller, 2012); these methods may allow for the separation between issues of estimating valid factor scores and estimating propensity scores that balance covariates.

Despite these limitations, my study was characterized by several key strengths, including using new standardized measures of estimation performance and examining simulation criteria related to how applied researchers would actually use the methods (i.e., with respect to covariate balance), beyond simply distributional qualities of the estimators. In addition, this study was the

first to consider differential measurement error in the context of causal inference.

Recommendations

Given the assumption of a correctly specified propensity score model, I recommend employing the MNLFA fully inclusive FS method for estimating causal effects with covariate measurement error. Although the unconditional fully inclusive FS method yielded unbiased results even in the presence of impact or DIF, the MNLFA fully inclusive FS method yielded estimates that were closer to the population values more often (i.e., with lower RMSPBR), was robust to over-modelling without increasing the variability of the estimate, and provided more accurate information about the balance of the true LV.

The preferred approach, as with any LV scoring procedure, is to gather as much information about the LV by using impact- and DIF-detection methods, as in Bauer (2017) and Curran et al. (2014). Substantive theory and evidence should drive the selection of potential moderating variables, although automated methods exist and others are in development (Cole, Gottfredson, & Giordano, 2018). Once these relationships have been identified, factor scores should be estimated as EAPs using the fully inclusive MNLFA model, erring on the side of over-modeling as opposed to under-modeling. Balancing weights should be estimated, and balance should be assessed on the factor scores. If balance is acceptable, effect estimation should proceed by using weighted regression of the outcome on treatment, perhaps also including some of the covariates in the regression model (as recommended by Nguyen et al., under review). If possible, a bootstrap confidence interval should be generated; otherwise, a sandwich standard error-based confidence interval may be appropriate. Researchers should report the fit of the factor model, balance on the covariates, and the results of a sensitivity analysis to examine the robustness of the estimate to unmeasured confounding.

Although the fully inclusive methods performed well in the present simulation and in that of

Nguyen et al. (under review), it is reasonable to be cautious about using it given its non-standard form (i.e., including the treatment as an indicator) and potential for bias due to misspecification (as was hinted at in Nguyen et al., under review). If a simple method is used instead, a MNLFA-based FS should be used, given its universally superior performance to the unconditional FS method and to using the items across all conditions and metrics. With many items, which may be common in some scales, the bias due to the indeterminacy of the factor will be small (Bollen, 2002).

Future Directions

This study was the first to examine the performance of LV methods in the context of propensity score analysis with differential measurement error in a covariate. Much uncertainty still remains about these techniques; in particular, their robustness to model misspecification remains to be seen. This model misspecification may come in the form of incorrectly specified impact and DIF relationships in the measurement model as well as ignored nonlinear relationships between the latent confounder and the treatment (especially in the context of the fully inclusive methods). This is a critical avenue of future research given that the value of propensity score-based methods is the removal of the necessity of correct functional form for modeling an outcome on treatment and confounders. Broadly, more research is required on the decision-making with MNLFA and the consequences of failing to identify relationships that exist in the population. In particular, future research should examine the effects of incorrectly modeled nonlinear relationships between the latent variable and the treatment, model misspecification both in the propensity score model and measurement model, and extensions to more covariates, including those measured with error.

Another important direction is the examination of LV methods in this context with the application of other causal inference methods, including matching, nonparametrically estimated

balancing weights, time-varying treatments, and doubly robust methods. For example, are there benefits to estimating the treatment effect with a weighted structural equation model over weighted factor score regression? Do matching and non-parametrically estimated balancing weights provide additional robustness against bias due to measurement error given their robustness to unmeasured confounding and model misspecification (Zubizarreta, Paredes, & Rosenbaum, 2014)? To date, no study on measurement error in causal inference has considered any techniques beyond logistic regression-based propensity scores, which, though unbiased asymptotically, are not always optimal in finite samples.

Methodological innovation and advancement often comes at the intersection of multiple scientific disciplines; by combining the state of the art in measurement from psychology with well-studied and robust causal inference methods in biostatistics and econometrics, new possibilities emerge for discussion among these disciplines and mutual adoption of techniques. As the problems of measurement error in the human sciences become increasingly acknowledged, finding ways to bring advances in LV modeling techniques into the methodological toolbox of causal inference researchers will create incremental progress in the attempt to create robust, reliable, and replicable sciences.

Figures and Tables

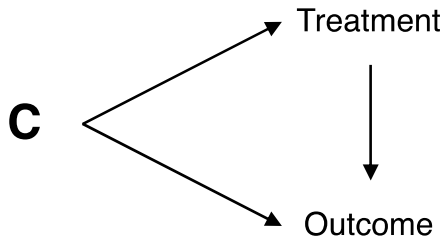

Figure 1. Nonparametric path diagram illustrating the basic structure of confounding. A set of variables **C** causes both selection into treatment and variation in the outcome. This situation, common in observational studies, is known as confounding.

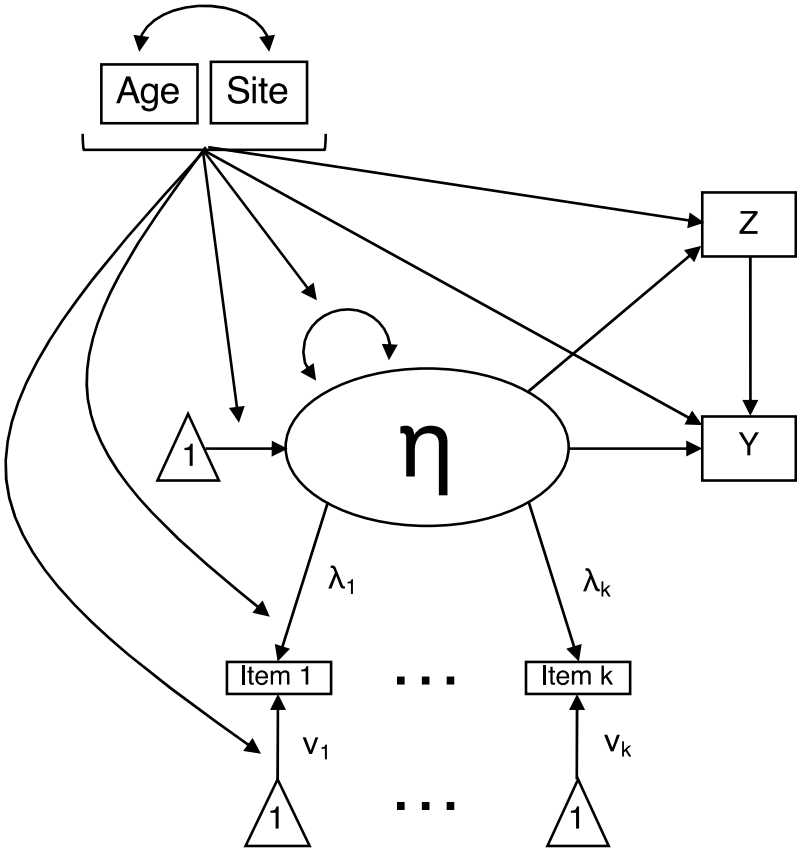

Figure 2. Path diagram depicting the data-generating model. η is the latent variable (LV) whose mean and variance depend on the observed variables age and site. Z is the treatment. Y is the outcome. Not depicted are the age-site interaction on the LV mean and the age-LV and site-LV interactions on the outcome. See Tables 1 and 2 for specific parameter values.

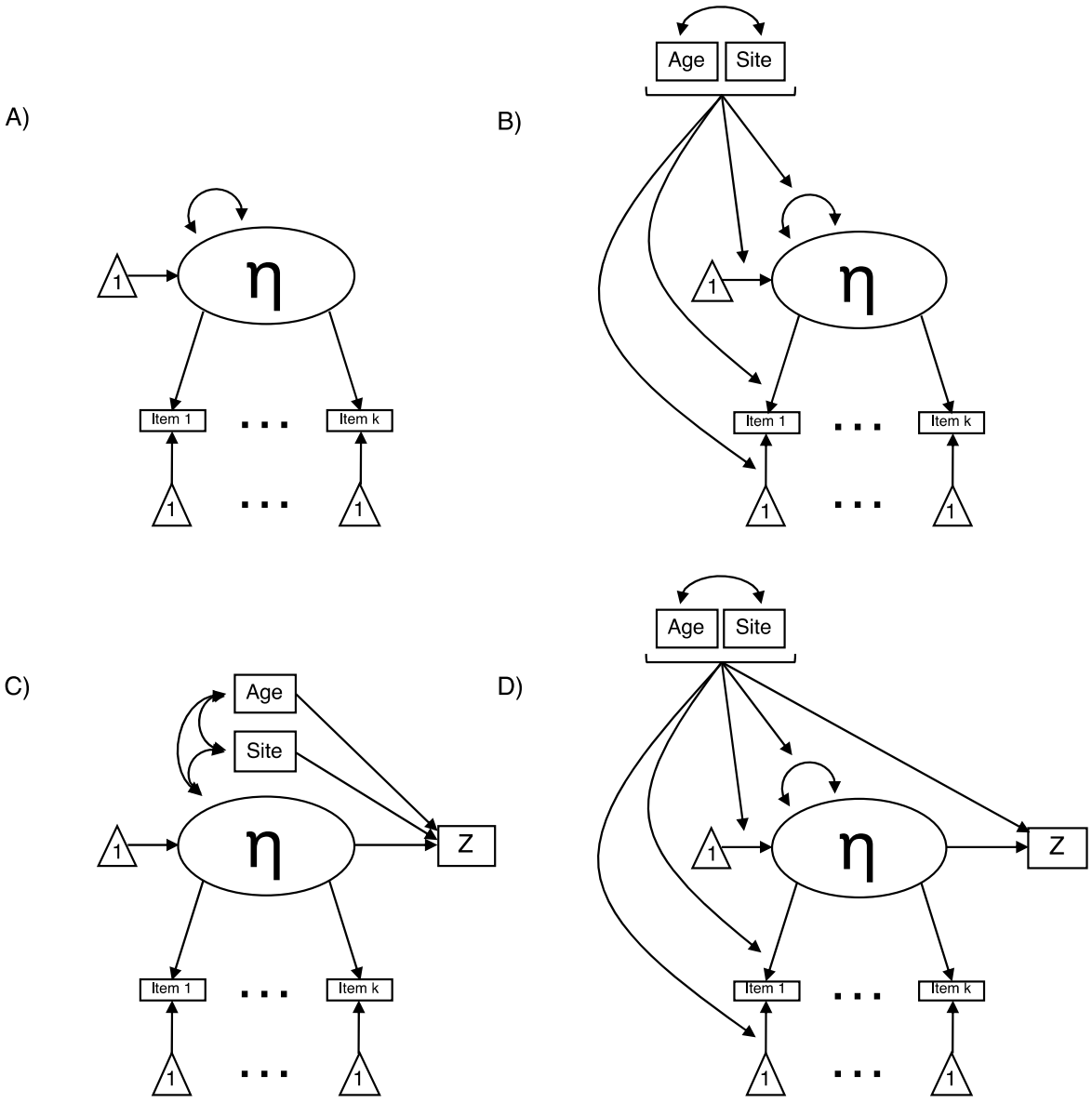

Figure 3. Path diagrams corresponding to the four factor models fit. A) Simple FS; B) MNLFA Simple FS; C) Fully Inclusive FS; D) MNLFA Fully Inclusive FS. η is the latent variable; Z is the treatment.

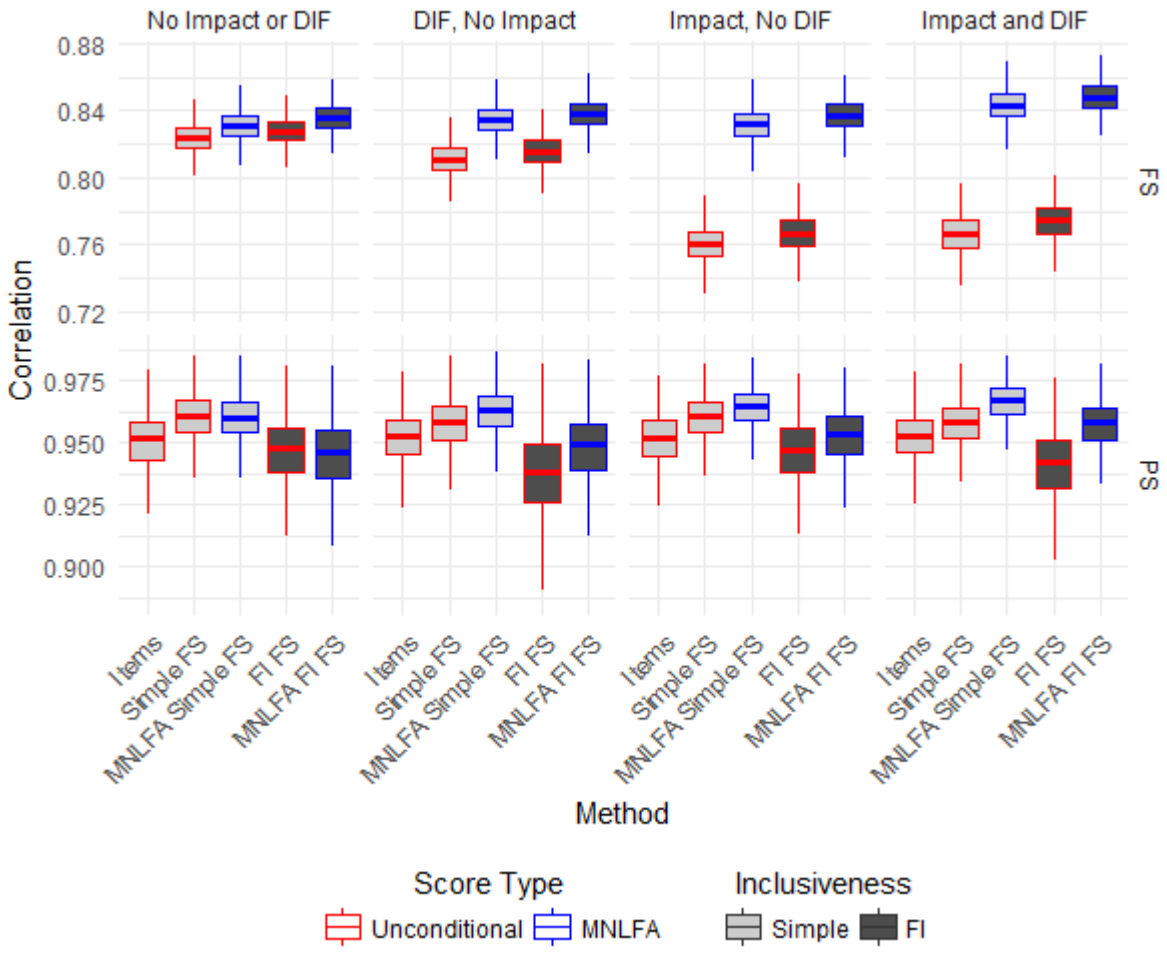

Figure 4. Correlations between estimated factor scores and true values of the latent variable (upper plot) and between estimated propensity scores and optimal propensity scores (lower plot) for each method with six items. FI = fully inclusive; FS = factor score; PS = propensity score. FS quality is not perfectly in line with PS quality, and neither align perfectly with bias performance.

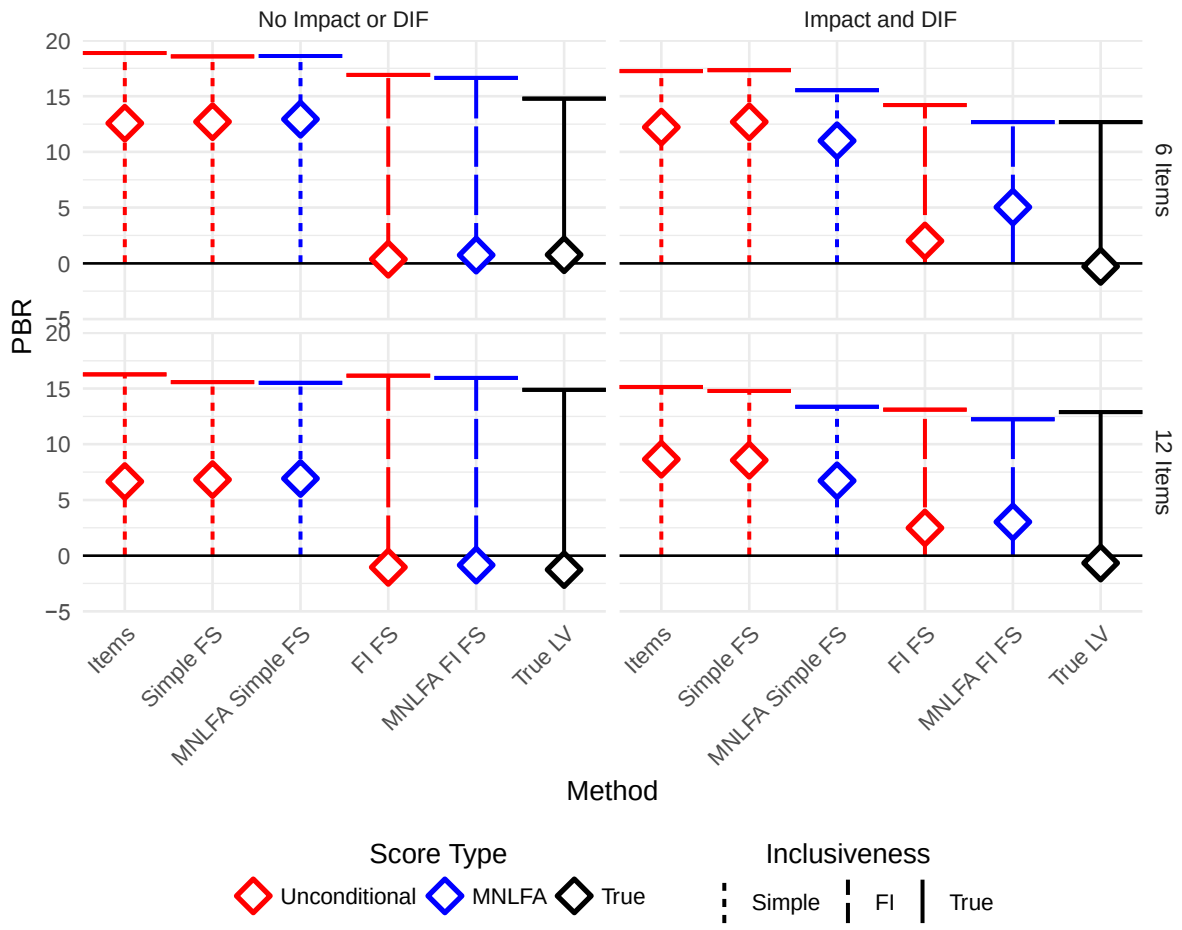

Figure 5. Percent bias remaining (PBR) of each method in the "Impact absent, DIF absent" and "Impact present, DIF present" conditions. Diamonds represent the mean PBR for each method. The horizontal bars represent the RMSPBR for each method. FI = fully inclusive.

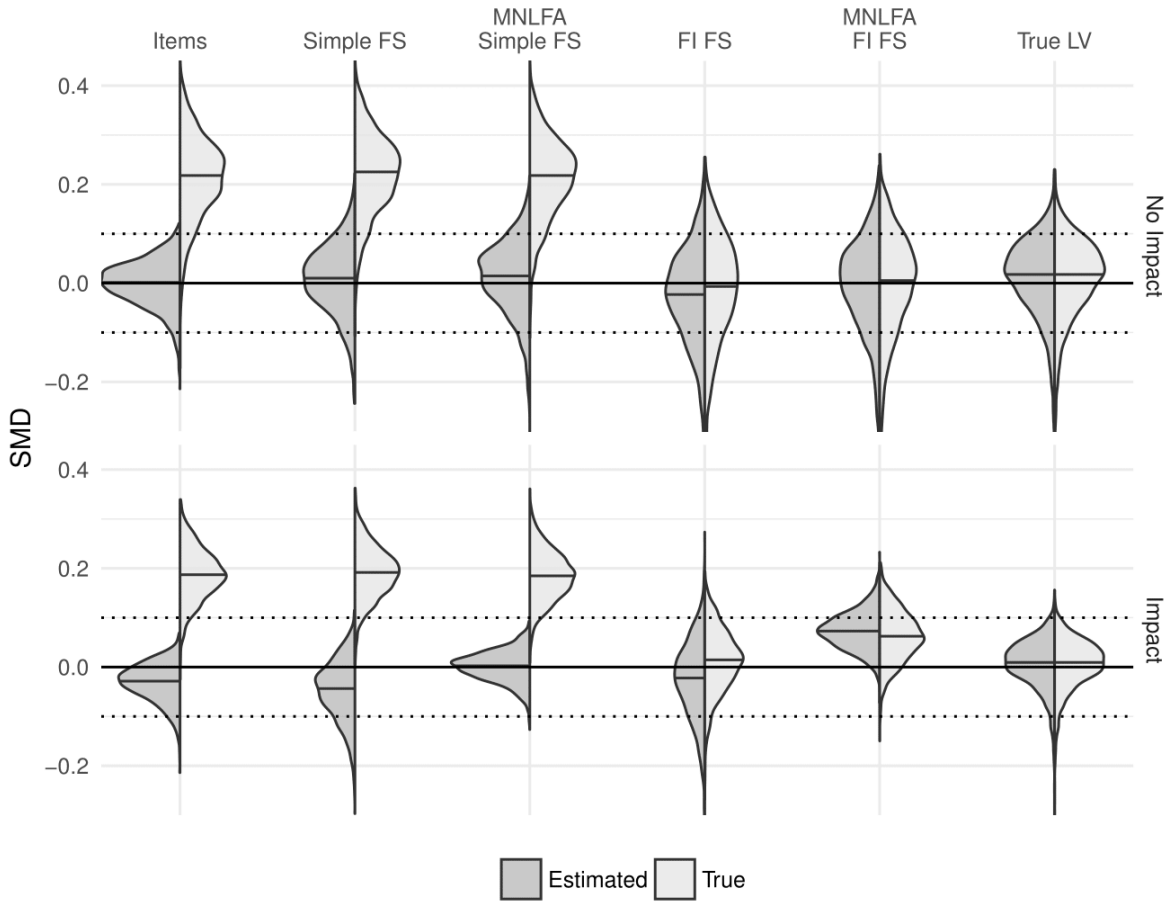

Figure 6. Split violin plots of balance across replications. Dark densities, facing left, are the weighted standardized mean differences (SMDs) of the estimated factor score (or the mean of the SMDs for the items). Light densities, facing right, are those of the true latent variable. The solid line within each density is the median. The line at zero indicates perfect balance, and the dotted lines at -0.1 and 0.1 indicate the boundaries of acceptable balance. Densities are shown for the six-item condition stratified by the presence or absence of impact and ignoring the presence or absence of DIF. FI = fully inclusive; FS= factor score. There is greater agreement between the estimated and true balance summaries when using the FI scores. Notably, when impact is present (lower panel), the MNLFA FI FS method produces some imbalance in the estimated and true balance summaries, though they remain in agreement.

Table 1. Data-Generating Model Parameters for Structural Equations

Parameter	Function	Value	
		Impact Absent	Impact Present
M2: Mean Impact			
γ_0	Intercept	-0.04	-0.04
γ_1	Age slope	0.18	0.19
γ_2	Site slope	0.21	0.22
γ_3	Age-Site interaction slope	0.05	0
M3: Variance Impact			
β_0	Intercept	0.52	0.85
β_1	Age slope	0.36	0
β_2	Site slope	0.63	0
M6: Treatment Selection			
a_0	Intercept	-1.05	
a_1	Age slope	0.26	
a_2	Site slope	0.93	
a_3	LV slope	0.64	
M7: Outcome			
τ	Treatment effect	0	
b_1	Age slope	0.50	
b_2	Site slope	0.88	
b_3	LV slope	1.22	
b_4	Age-LV interaction slope	0.15	
b_5	Site-LV interaction slope	0.46	
σ^2	Residual variance	9	

Note. LV = Latent Variable.

Table 2. Data-Generating Model Parameters for Measurement Equations

	Intercept (Equation M4)						Loading (Equation M5)					
	Baseline		Age		Site		Baseline		Age		Site	
6 Items												
Item 1	κ_{01}	0	κ_{11}	-	κ_{21}	-	ω_{01}	1.00	ω_{11}	-	ω_{21}	-
Item 2	κ_{02}	1	κ_{12}	-	κ_{22}	-	ω_{02}	1.75	ω_{12}	-	ω_{22}	-
Item 3	κ_{03}	2	κ_{13}	-	κ_{23}	-	ω_{03}	2.50	ω_{13}	-	ω_{23}	-
Item 4	κ_{04}	0	κ_{14}	0.6	κ_{24}	0.7	ω_{04}	1.00	ω_{14}	0.3	ω_{24}	0.9
Item 5	κ_{05}	1	κ_{15}	-0.6	κ_{25}	0.7	ω_{05}	1.75	ω_{15}	-0.3	ω_{25}	0.9
Item 6	κ_{06}	2	κ_{16}	0.6	κ_{26}	0.7	ω_{06}	2.50	ω_{16}	0.3	ω_{26}	-0.9
12 Items												
Item 1	κ_{01}	0.0	κ_{11}	-	κ_{21}	-	ω_{01}	1.0	ω_{11}	-	ω_{21}	-
Item 2	κ_{02}	0.4	κ_{12}	-	κ_{22}	-	ω_{02}	1.3	ω_{12}	-	ω_{22}	-
Item 3	κ_{03}	0.8	κ_{13}	-	κ_{23}	-	ω_{03}	1.6	ω_{13}	-	ω_{23}	-
Item 4	κ_{04}	1.2	κ_{14}	-	κ_{24}	-	ω_{04}	1.9	ω_{14}	-	ω_{24}	-
Item 5	κ_{05}	1.6	κ_{15}	-	κ_{25}	-	ω_{05}	2.2	ω_{15}	-	ω_{25}	-
Item 6	κ_{06}	2.0	κ_{16}	-	κ_{26}	-	ω_{06}	2.5	ω_{16}	-	ω_{26}	-
Item 7	κ_{07}	0.0	κ_{17}	0.6	κ_{27}	0.7	ω_{07}	1.0	ω_{17}	0.3	ω_{27}	0.9
Item 8	κ_{08}	0.4	κ_{18}	-0.6	κ_{28}	0.7	ω_{08}	1.3	ω_{18}	-0.3	ω_{28}	0.9
Item 9	κ_{09}	0.8	κ_{19}	0.6	κ_{29}	0.7	ω_{09}	1.6	ω_{19}	0.3	ω_{29}	-0.9
Item 10	$\kappa_{0,10}$	1.2	$\kappa_{1,10}$	0.6	$\kappa_{2,10}$	0.7	$\omega_{0,10}$	1.9	$\omega_{1,10}$	0.3	$\omega_{2,10}$	0.9
Item 11	$\kappa_{0,11}$	1.6	$\kappa_{1,11}$	-0.6	$\kappa_{2,11}$	0.7	$\omega_{0,11}$	2.2	$\omega_{1,11}$	-0.3	$\omega_{2,11}$	0.9
Item 12	$\kappa_{0,12}$	2.0	$\kappa_{1,12}$	0.6	$\kappa_{2,12}$	0.7	$\omega_{0,12}$	2.5	$\omega_{1,12}$	0.3	$\omega_{2,12}$	-0.9

Note. A dash (-) indicates zero. In the "DIF absent" condition, all parameters other than those under "Baseline" are zero.

Table 3. Mean Bias and Variability for Treatment Effect Estimates

No. Items	Impact	DIF	Method					
			Items	S FS	MS FS	FI FS	MFI FS	True LV
PBR								
6	Absent	Absent	13	13	13	0	1	1
		Present	11	12	11	-2	-1	-1
	Present	Absent	11	11	11	0	5	-1
		Present	12	13	11	2	5	0
12	Absent	Absent	7	7	7	-1	-1	-1
		Present	8	9	7	1	0	0
	Present	Absent	8	8	7	2	4	0
		Present	9	9	7	2	3	-1
RMSPBR								
6	Absent	Absent	19	19	19	17	17	15
		Present	18	18	17	17	16	14
	Present	Absent	17	16	16	15	13	13
		Present	17	17	16	14	13	13
12	Absent	Absent	16	16	16	16	16	15
		Present	16	16	15	15	15	14
	Present	Absent	15	15	14	13	13	13
		Present	15	15	13	13	12	13

Note. PBR = Percent Bias Remaining; RMSPBR = Root Mean Squared PBR; S FS = Simple Factor Score; MS FS = MNLFA Simple Factor Score; FI FS = Fully Inclusive Factor Score; MFI FS = MNLFA Fully Inclusive Factor Score; LV = Latent Variable. Values are rounded to nearest whole number.

REFERENCES

Achenbach, T. M., & Edelbrock, C. S. (1981). Behavioral problems and competencies reported by parents of normal and disturbed children aged four through sixteen. *Monographs of the society for research in child development*, 1-82.

Burt, R. S. (1976). Interpretational confounding of unobserved variables in structural equation modeling. *Sociological Methods & Research, 5*, 3–53.

Augurzky, B. & Schmidt, C. (2001). The propensity score: A means to an end. Discussion Paper 271, Institute for the Study of Labor (IZA).

Austin, P. C. (2011). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioral Research, 46*(3), 399–424.

Bauer, D. J. (2017). A More General Model for Testing Measurement Invariance and Differential Item Functioning. *Psychological Methods*.

Bauer, D. J., & Hussong, A. M. (2009). Psychometric approaches for developing commensurate measures across independent studies: Traditional and new models. *Psychological Methods*, *14*(2), 101–125.

Blackwell, M., Honaker, J., & King, G. (2015). A Unified Approach to Measurement Error and Missing Data: Overview and Applications. *Sociological Methods & Research*, 0049124115585360.

Bollen, K. A. (2002). Latent Variables in Psychology and the Social Sciences. *Annual Review of Psychology*, *53*(1), 605–634.

Bollen, K. A., Kirby, J. B., Curran, P. J., Paxton, P. M., & Chen, F. (2007). Latent variable models under misspecification: two-stage least squares (2SLS) and maximum likelihood (ML) estimators. *Sociological Methods & Research*, *36*(1), 48-86.

Bray, B. C., Lanza, S. T., & Tan, X. (2015). Eliminating Bias in Classify-Analyze Approaches for Latent Class Analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(1), 1–11.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Stürmer, T. (2006). Variable Selection for Propensity Score Models. *American Journal of Epidemiology*, *163*(12), 1149–1156.

Camstra, A., & Boomsma, A. (1992). Cross-validation in regression and covariance structure analysis: An overview. *Sociological Methods & Research, 21*(1), 89-115.

Cohen, J. (1998). Statistical power analysis for the behavioral sciences (2nd edition). Hillsdale, Lawrence Earlbaum Associates, 25, New Jersey.

Cole, S. R., Chu, H., & Greenland, S. (2006). Multiple-imputation for measurement-error correction. *International journal of epidemiology*, *35*(4), 1074-1081.

Cole, V., Gottfredson, N., & Giordano, M. (2018). aMNLFA: Automated Fitting of Moderated Nonlinear Factor Analysis Through the 'Mplus' Program. R package version 0.1.

Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*(4), 330–351.

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*, *0*(0), 1–18.

Curran, P. J., Cole, V., Giordano, M., Georgeson, A. R., Hussong, A. M., & Bauer, D. J. (2017). Advancing the Study of Adolescent Substance Use Through the Use of Integrative Data Analysis. *Evaluation & the Health Professions*, 0163278717747947.

Curran, P. J., McGinley, J. S., Bauer, D. J., Hussong, A. M., Burns, A., Chassin, L., … Zucker, R. (2014). A Moderated Nonlinear Factor Model for the Development of Commensurate Measures in Integrative Data Analysis. *Multivariate Behavioral Research*, *49*(3), 214–231.

Diamond, A., & Sekhon, J. S. (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics*, *95*(3), 932–945.

Fava, J. L., & Velicer, W. F. (1992). An empirical comparison of factor, image, component, and scale scores. *Multivariate Behavioral Research*, *27*(3), 301–322.

Forster, M. R. (2000). Key concepts in model selection: Performance and generalizability. *Journal of mathematical psychology*, *44*(1), 205-231.

Foster, E. M. (2010). Causal inference and developmental psychology. *Developmental Psychology*, *46*(6), 1454–1480.

Freedman, D. A., & Berk, R. A. (2008). Weighting Regressions by Propensity Scores. *Evaluation Review*, *32*(4), 392–409.

Golinelli, D., Ridgeway, G., Rhoades, H., Tucker, J., & Wenzel, S. (2012). Bias and variance trade-offs when combining propensity score weighting and regression: with an application to HIV status and homeless men. *Health Services and Outcomes Research Methodology*, *12*(2-3), 104-118.

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods*, *6*(4), 430–450.

Hall, C. E., Steiner, P. M., & Kim, J. S. (2015). Doubly Robust Estimation of Treatment Effects from Observational Multilevel Data. *Quantitative Psychology Research*, *140*, 321.

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R Package for Facilitating Large-Scale Latent Variable Analyses in M plus. *Structural Equation Modeling: A Multidisciplinary Journal*, 1-18.

Hernán M. A. & Robins, J. M. (2018). *Causal Inference*. Boca Raton: Chapman & Hall/CRC, forthcoming.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis*, *15*(3), 199–236.

Holland, P. W., & Thayer, D. T. (1988). Differential Item Performance and the Mantel-Haenszel Procedure. In H. Wainer, and H. I. Brown (Eds.), *Test Validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Jakubowski, M. (2015). Latent variables and propensity score matching: a simulation study with application to data from the Programme for International Student Assessment in Poland. *Empirical Economics, 48*(3), 1287–1325.

Jones, D. S., & Podolsky, S. H. (2015). The history and fate of the gold standard. *The Lancet*, *385*(9977), 1502-1503.

Kainz, K., Greifer, N., Givens, A., Swietek, K., Lombardi, B. M., Zietz, S., & Kohn, J. L. (2017). Improving Causal Inference: Recommendations for Covariate Selection and Balance in Propensity Score Methods. *Journal of the Society for Social Work and Research*, 000–000.

Kaplan, D. (1988). The impact of specification error on the estimation, testing, and improvement of structural equation models. *Multivariate Behavioral Research*, *23*(1), 69-86.

King, G., & Nielsen, R. (2016). Why propensity scores should not be used for matching.

Lee, B. K., Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine*, *29*(3), 337–346.

Leyrat, C., Seaman, S. R., White, I. R., Douglas, I., Smeeth, L., Kim, J., … Williamson, E. J. (2017). Propensity score analysis with partially observed covariates: How should multiple imputation be used? *Statistical Methods in Medical Research*, 0962280217713032.

Liu, W., Kuramoto, S. J., & Stuart, E. A. (2013). An Introduction to Sensitivity Analysis for Unobserved Confounding in Nonexperimental Prevention Research. *Prevention Science*, *14*(6), 570–580.

Long, J. S. (1997). Regression Models for Categorical and Limited Dependent Variables. Thousand Oaks, CA: Sage Press.

Lu, I. R. R., & Thomas, D. R. (2008). Avoiding and Correcting Bias in Score-Based Latent Variable Regression With Discrete Manifest Items. *Structural Equation Modeling: A Multidisciplinary Journal*, *15*(3), 462–490.

Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, *23*(19), 2937–2960.

Maraun, M. D. (1996). Metaphor Taken as Math: Indeterminancy in the Factor Analysis Model. *Multivariate Behavioral Research*, *31*(4), 517–538.

Masyn, K. & Waldman, M. (2016, May). *Latent Class Covariate Balancing for Causal Inference*. Paper presented at the Modern Modeling Methods Conference, Storrs, CT.

McCaffrey, D. F., Lockwood, J. R., & Setodji, C. M. (2013). Inverse probability weighting with error-prone covariates. *Biometrika*, *100*(3), 671–680.

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity Score Estimation With Boosted Regression for Evaluating Causal Effects in Observational Studies. *Psychological Methods*, *9*(4), 403–425.

Meng, X. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, *9*(4), 538–558.

Michaelides, M. P. (2008). An illustration of a Mantel-Haenszel procedure to flag misbehaving common items in test equating. *Practical Assessment Research & Evaluation*, *13*(7).

Millsap, R. E. (1995). Measurement Invariance, Predictive Invariance, and the Duality Paradox. *Multivariate Behavioral Research*, *30*(4), 577–605.

Moons, K. G., Donders, R. A., Stijnen, T., & Harrell, F. E. (2006). Using the outcome for imputation of missing predictor values was preferred. *Journal of clinical epidemiology*, *59*(10), 1092-1101.

Nguyen, T. Q., Ebnesajjad, C., Hong, H., Stuart, E. (under review). A fix for bias due to a mismeasured/latent covariate in propensity score weighting analysis: Factor scores from models combining multiple measurements with exposure and other variables. Under review.

Nguyen, T. Q., Ebnesajjad, C., Stuart, E. A., Kennedy, R. D., & Johnson, R. M. (2018). Does Marijuana Use at Ages 16–18 Predict Initiation of Daily Cigarette Smoking in Late Adolescence and Early Adulthood? A Propensity Score Analysis of Add Health Data. *Prevention Science*, 1–11.

Odgers, C. L., Caspi, A., Nagin, D. S., Piquero, A. R., Slutske, W. S., Milne, B. J., … Moffitt, T. E. (2008). Is it important to prevent early exposure to drugs and alcohol among adolescents? *Psychological Science*, *19*(10), 1037–1044.

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling*, *8*(2), 287-312.

Raykov, T. (2012). Propensity Score Analysis With Fallible Covariates A Note on a Latent Variable Modeling Approach. *Educational and Psychological Measurement*, *72*(5), 715–733.

Robins, J. M. (2000). Marginal Structural Models versus Structural nested Models as Tools for Causal inference. Statistical Models in Epidemiology, the Environment, and Clinical Trials The IMA Volumes in Mathematics and Its Applications, 95-133.

Robins, J. M., Hernan, M. A., & Brumback, B. (2000). Marginal Structural Models and Causal Inference in Epidemiology. *Epidemiology*, *11*(5), 550-560.

Rodríguez De Gil, P., Bellara, A. P., Lanehart, R. E., Lee, R. S., Kim, E. S., & Kromrey, J. D. (2015). How Do Propensity Score Methods Measure Up in the Presence of Measurement Error? A Monte Carlo Study. *Multivariate Behavioral Research*, 50(5), 520–532.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, *79*(387), 516-524.

Rothwell, P. M. (2005). External validity of randomised controlled trials: "to whom do the results of this trial apply?". *The Lancet*, *365*(9453), 82-93.

Rubin, D. B. (2001). Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation. *Health Services and Outcomes Research Methodology*, *2*(3–4), 169–188.

Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods*, *13*(4), 279–313.

Schonemann, P. H. (1996). The psychopathology of factor indeterminancy. *Multivariate behavioral research*, *31*(4), 571-577.

Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can Nonrandomized Experiments Yield Accurate Answers? A Randomized Experiment Comparing Random and Nonrandom Assignments. *Journal of the American Statistical Association*, *103*(484), 1334–1344.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Shah, B. R., Laupacis, A., Hux, J. E., & Austin, P. C. (2005). Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. *Journal of Clinical Epidemiology*, *58*(6), 550–559.

Skrondal, A., & Kuha, J. (2012). Improved regression calibration. *Psychometrika, 77,* 649–669.

Skrondal, A., & Laake, P. (2001). Regression among factor scores. *Psychometrika*, *66*(4), 563–575.

Steiner, P. M., Cook, T. D., & Shadish, W. R. (2011). On the Importance of Reliable Covariate Measurement in Selection Bias Adjustments Using Propensity Scores. *Journal of Educational and Behavioral Statistics*, *36*(2), 213–236.

Stuart, E. A. (2010). Matching Methods for Causal Inference: A Review and a Look Forward. *Statistical Science, 25*(1), 1–21.

Stuart, E. A., & Green, K. M. (2008). Using full matching to estimate causal effects in nonexperimental studies: Examining the relationship between adolescent marijuana use and adult outcomes. *Developmental Psychology, 44*(2), 395–406.

Stuart, E. A., & Rubin, D. B. (2008). Best practices in quasi-experimental designs. *Best practices in quantitative methods*, 155-176.

Thoemmes, F. J., & Kim, E. S. (2011). A Systematic Review of Propensity Score Methods in the Social Sciences. *Multivariate Behavioral Research*, *46*(1), 90–118.

Webb-Vargas, Y., Rudolph, K. E., Lenis, D., Murakami, P., & Stuart, E. A. (2015). An imputation-based solution to using mismeasured covariates in propensity score analysis. *Statistical Methods in Medical Research*, 0962280215588771.

White H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society*, 48, 817–838.

Zubizarreta, J. R., Paredes, R. D., & Rosenbaum, P. R. (2014). Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in Chile. *The Annals of Applied Statistics*, *8*(1), 204–231.

Zwick, R. (2012). A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement. *ETS Research Report Series*, *2012*(1), i–30.