Abstract Title Page

**Title:** Can You Correct a Propensity Score Analysis for Covariate Measurement Error?

**Authors and Affiliations:** J.R. Lockwood and Daniel F. McCaffrey, Educational Testing Service
Background

In many research areas including education, there is a need to use observed variables to correct for missing information. Such applications include adjusting for non-response in which some units do not have observed values of quantities of interest, and observational studies of causal effects in which each unit’s potential outcome under one alternative treatment assignment is observed and the remaining potential outcomes are missing (Kang & Schafer, 2007; Imbens, 2000; Lunceford & Davidian, 2004; Scharfstein, Rotnitzky, & Robins, 1999; Robins & Rotnitzky, 1995; Robins, Rotnitzky, & Zhao, 1995; Bang & Robins, 2005; McCaffrey, Ridgeway, & Morral, 2004). A common approach to these problems is to use covariates observed for all units, along with ignorability assumptions about response or treatment assignment for outcomes of interest conditional on those covariates, to construct consistent estimators in the presence of the missing data. The approaches do not require assumptions about the functional form of the relationships between covariates and outcomes, making them attractive in some settings. Examples include inverse-probability-weighted estimators in non-response settings (Kang & Schafer, 2007), analogous estimators based on the propensity score (Rosenbaum & Rubin, 1983) for causal inference problems, and matching or stratification estimators (Stuart, 2010) using the propensity score or full covariate vectors for causal inference problems.

The success of these methods requires the covariates that are necessary for missing data to be ignorable to be free of measurement error. When the probability of response or treatment assignment depends on latent quantities that are observed only through error-prone surrogates, ignoring the measurement error and just using the surrogates can result in bias in weighting or matching estimators (D’Agostino & Rubin, 2000; McCaffrey, Lockwood, & Setodji, 2013; Pearl, 2010; Steiner, Cook, & Shadish, 2011; Raykov, 2012; Stuart, 2013; Yi, Ma, & Carroll, 2012).

The need for covariates to be free of measurement error is problematic for the growing application of propensity score methods in applied education research (Hallberg, Cook, & Figlio, 2013; Kaplan & Chen, 2012; Steiner et al., 2011; Stuart, 2007). For example, in observational studies of educational interventions, scores from standardized tests are commonly used as covariates to adjust for differences among non-equivalent groups. However, test scores are error-prone measures of latent achievement (Lord, 1980). It is clear that outcomes of interest, such as future achievement, depend on these latent achievement attributes, and thus observational treatment groups need to be balanced on the latent attributes in order to estimate treatment effects unbiasedly. In general, balancing the error-prone scores is not sufficient to balance the latent attributes when groups differ on those latent attributes (Yi et al., 2012). This means that education studies using test scores as covariates in propensity score models to adjust for latent achievement attributes could lead to biased inferences if the measurement error in the test scores is not addressed. Similar arguments apply to education studies using other types of error-prone measures such as scores from teacher knowledge tests (The Learning Mathematics for Teaching Project, 2013), measures of teaching practices from classroom observations (Kelcey, McGinn, & Hill, 2013; Bill and Melinda Gates Foundation, 2013), or indicators from surveys.

Given the prevalence of both error-prone covariates and the application of propensity score approaches in applied education research, there is clear need for guidance about strategies that are or are not appropriate for correcting for covariate measurement error in propensity score analyses.

Objective

The goals of this presentation are to 1) synthesize emerging research about what approaches to addressing covariate measurement error in propensity score applications are valid and which are not; 2) to provide practical advice about how education researchers can address this problem in their analyses; and 3) discuss areas of future research. We will discuss theoretical and practical issues for
both weighting and matching estimators using propensity scores with error-prone covariates.

**Novelty of Study**

The literature has widespread acknowledgment that covariate measurement error is problematic for propensity score analysis (Steiner et al., 2011; Raykov, 2012) but there has been significantly less work done on providing theoretically defensible and practical solutions. One of our key messages in this presentation is that covariate measurement error introduces asymmetry between weighting and matching applications. In standard (i.e., no measurement error) settings, many analysts base the choice between weighting and matching on criteria such as preference, disciplinary tradition, transparency or statistical efficiency. Covariate measurement error makes the choice more critical. Whereas it is theoretically possible to fix propensity score weighting estimators for covariate measurement error, and methods for doing so exist, we will present novel results that we derived indicating that in most practical cases, it is not possible to similarly correct matching estimators. That is, it is possible to construct a function of the error-prone data that serves as a “corrected” propensity score for the purposes of weighting, but that function is generally not appropriate for matching or stratification. Furthermore, it is not generally possible to find any function of the observed covariates that is appropriate for matching. Thus correcting propensity score analyses for covariate measurement error introduces a fundamental distinction between weighting and matching that does not exist in standard cases without measurement error. To our knowledge these results are novel and are not only of theoretical interest, but also of interest to applied researchers seeking guidance about how to appropriately handle covariate measurement error in propensity score applications.

**Statistical Methods**

**Notation and Assumptions:** We let \((T, Y_0, Y_1, X, S)\) be random quantities with a joint distribution on a probability space. We assume all conditional distributions required in the development are well-defined. We assume we are dealing with IID samples from this joint distribution. We also assume that \(T\) is a dichotomous treatment indicator, observed for all units and that \(Y_0\) and \(Y_1\) are potential outcomes: \(Y_0\) is the outcome for a unit if it is assigned to \(T = 0\) and \(Y_1\) is the outcome for a unit if it assigned to \(T = 1\). \(Y_0\) is observed for units only when \(T = 0\) and \(Y_1\) is observed for units only when \(T = 1\). We assume \(X\) is a random vector of latent variables that are relevant to the analysis but not observed for any unit. In practice, \(X\) might be partitioned as \((X_u, Z)\) where \(X_u\) are components of \(X\) that are unobserved and \(Z\) are components that are observed, but this distinction will not be necessary in our development and our results extend to that case. Finally we assume \(S\) is another random quantity observed for all units.

Throughout, we make the same strong ignorability assumption of Rosenbaum and Rubin (1983):

**Assumption 1:** \((Y_0, Y_1)\) is independent of \(T\) given \(X\), and \(0 < \Pr(T = 1|X) < 1\) for all \(X\).

It is critical to note that the strong ignorability assumption applies to \(X\) which is not observed for any unit. The goal of the work is to understand what progress can be made in propensity score analysis using \(S\) rather than \(X\). Following the definition of surrogacy of Carroll et al. (2006) we make the following assumption regarding \(S\):

**Assumption 2:** \((Y_0, Y_1, T)\) is independent of \(S\) given \(X\).

The idea is that \(S\) would be irrelevant to the analysis if \(X\) were observed. There are two important examples of \(S\). The first is that any function \(g(X)\) of \(X\) is a surrogate by this definition, because \(g(X)\) is constant conditional on \(X\) and so, given \(X\), it is conditionally independent of any other random quantity. This allows us to prove general results about surrogates that extend existing results such as those provided by Rosenbaum and Rubin (1983) for propensity scores, which are functions of \(X\).
The second case is particularly relevant to the case of covariates with measurement error, where $S$ may represent error-prone measures $W$ of the latent variable $X$ or functions of such error-prone measures. For example, in education settings, $X$ might correspond to latent achievement attributes, $W$ might correspond to observed test scores, and the context of the application may justify making Assumption 2 so that any function of the observed test scores $W$ would be irrelevant to the study if the latent variables $X$ were observed instead.

**A Solution to Covariate Measurement Error for Propensity Score Weighting:** McCaffrey et al. (2013) show that under assumptions analogous to those previously stated, where $W$ is an error-prone measure of $X$, then it is possible to construct a consistent weighting estimator for various average treatment effects including the population average treatment effect. Letting $e(X) = \text{pr}(T = 1|X)$ denote the true propensity score, they demonstrate that a function $A(W)$ satisfying

\[
E[A(W)|X] = p(T = 1)/e(X)
\]

can be used as weights for units for which $T = 1$ to obtain a consistent estimate of $E[Y_1]$. An analogous function constructed for $T = 0$ units can be used to obtain a consistent estimate of $E[Y_0]$. Combining the two provides a consistent estimate of the population average treatment effect $E[Y_1 - Y_0]$. The intuition is that $A(W)$ is a function of the error-prone data such that on average, over the distribution of plausible values $w$ of $W$ that a unit with true value $x$ of $X$ might have, the unit receives the correct weight. In other words, $A(W)$ is conditionally unbiased for the true weight for every unit.

The challenge is finding such a function. McCaffrey et al. (2013) provide computational methods for the case of a single error-prone covariate. The methods require that $p(W|X)$ be treated as known and that any parameters of the true propensity score function $e(X)$ can be estimated from the error-prone data using methods for estimating parameters of nonlinear measurement error models such as those discussed by Carroll et al. (2006). The feasibility of the approach with multiple error-prone covariates, or in cases of heteroskedastic measurement error such as commonly present in standardized test scores (Lord, 1980), is currently unknown.

**A Solution to Covariate Measurement Error for Propensity Score Matching:** Since $A(W)$ defined above is conditionally unbiased for the true weight for each unit, might it be appropriate to construct a matching estimator using it? What about $B(W)$ such that $E[B(W)|X] = e(X)$, which is conditionally unbiased for the true propensity score for each unit? Or $E[e(X)|W]$, the conditional mean of the true propensity score given the error-prone measures? Can these variables be used for matching? More generally, is there any function of $W$ that is appropriate for matching? We prove a sequence of theorems indicating that in practical settings, the answer is no. Intuitively, matching on a function of $W$ fails because the conditional distribution of $X$ given that function and $T = 1$ does not equal the distribution of $X$ given the function and $T = 0$. That is, we show that in most practical situations there is no function of $W$ which we can match on to balance $X$.

To derive our result we first define a “surrogate balancing score” to be any surrogate $S$ that in addition to satisfying Assumption 2, satisfies

**Assumption 3:** $T$ is independent of $X$ given $S$.

This generalizes Rosenbaum and Rubin’s (1983) definition of a balancing score $b(X)$ that is a function of $X$. Because all functions of $X$ are surrogates, all Rosenbaum and Rubin balancing scores are surrogate balancing scores because they satisfy Assumption 3 by definition. In particular, the propensity score $e(X) = \text{pr}(T = 1|X)$ is a surrogate balancing score.

We then prove results establishing that matching on a surrogate balancing score is all that is needed for matching to yield unbiased estimates of the treatment effect. We also show that any function of a surrogate used in matching must be a surrogate balancing score to assure treatment
effect estimates are unbiased. However, we also prove that in common measurement error settings, including additive measurement error models for continuous variables (Carroll, Ruppert, Stefanski, & Crainiceanu, 2006; Wooldridge, 2002), misclassification models where $X$ is discrete (Buonaccorsi, 2010), and latent variable models such as factor and Item Response Theory models (Bollen, 1989; Joreskog & Moustaki, 2001; Lord, 1980; van der Linden & Hambleton, 1996), surrogate balancing scores will not exist under standard modeling assumptions made in practice, including in education research. The theorem indicates that several approaches that have been suggested in the literature for correcting propensity scores for covariate measurement error (Raykov, 2012; Sturmer, Schneeweiss, Avorn, & Glynn, 2005) will not generally lead to consistent matching estimators.

It is possible in some situations to match on functions of $(W, T)$ rather than $W$ alone and balance $X$ to obtain unbiased estimates of treatment effects. However, determining these functions is generally difficult or impossible. Finally, we also demonstrate that approaches using imputed values of $X$ from its conditional distribution given either $W$ or $(W, T)$ to construct matching estimators will not generally be appropriate unless $Y$ is also included in the imputation model. This undermines the original intent of weighting or matching methods, which is that they avoid assumptions about the functional form of the relationships between covariates and outcomes.

The combination of results imply that practically speaking it is not possible to fix matching estimators for covariate measurement error.

**Approximate Solutions:** Given these results it is important to consider what can be done in practice. One possibility is Simulation-Extrapolation (SIMEX), which provides approximately consistent estimators in models where other measurement error correction methods would be intractable or computationally infeasible (Carroll et al., 2006; Cook & Stefanski, 1994; Fung & Krewski, 1999; Shang, 2012). The key benefit of SIMEX compared to other approaches is that it requires only repeatedly fitting standard models that ignore covariate measurement error to simulated data containing additional measurement error, making it broadly accessible. Lockwood and McCaffrey (2013b) evaluated the potential for SIMEX to correct for covariate measurement error for weighting estimators and find that the approach can greatly mitigate bias relative to ignoring covariate measurement error, even in very complex models, but may require modest to large sample sizes to improve on mean squared error.

Another possibility is to rely on results by Steiner et al. (2011), who demonstrate through simulation that when multiple correlated error-prone covariates are available, including them all in the propensity score model and ignoring measurement error can help to mitigate bias. However, Lockwood and McCaffrey (2013a) find in a case study involving regression models for student test score outcomes using past test scores as covariates, even including up to 12 past test scores and ignoring measurement error was not sufficient to remove all of the evident bias due to omitted latent achievement attributes. Therefore additional studies in diverse education research settings are required to understand how much we can generally rely on multiple correlated proxies to adequately adjust for latent attributes.

**Conclusions**

Addressing covariate measurement error in propensity score applications is very challenging. The problem is of particular importance to applied education research given how many observed quantities we commonly use in analysis can be considered error-prone measures of latent quantities of interest. The presentation will provide a much-needed overview of the landscape of the problems and possible solutions to handling error-prone covariates in propensity score applications. The content will be of interest to both applied education researchers and methodologists.
Appendix A: References


Bill and Melinda Gates Foundation (2013). Ensuring Fair and Reliable Measures of Effective Teaching: Culminating Findings from the MET Project’s Three-Year Study.


