Peabody Journal of Education
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hpje20

Identifying Baseline Covariates for Use in Propensity Scores: A Novel Approach Illustrated for a Nonrandomized Study of Recovery High Schools

Emily E. Tanner-Smith[a] & Mark W. Lipsey[a]
[a] Vanderbilt University
Published online: 14 Apr 2014.

To cite this article: Emily E. Tanner-Smith & Mark W. Lipsey (2014) Identifying Baseline Covariates for Use in Propensity Scores: A Novel Approach Illustrated for a Nonrandomized Study of Recovery High Schools, Peabody Journal of Education, 89:2, 183-196, DOI: 10.1080/0161956X.2014.895647

To link to this article: http://dx.doi.org/10.1080/0161956X.2014.895647

PEABODY JOURNAL OF EDUCATION, 89: 183–196, 2014

ISSN: 0161-956X print / 1532-7930 online
DOI: 10.1080/0161956X.2014.895647

Routledge
Taylor & Francis Group

Identifying Baseline Covariates for Use in Propensity Scores: A Novel Approach Illustrated for a Nonrandomized Study of Recovery High Schools

Emily E. Tanner-Smith and Mark W. Lipsey
Vanderbilt University

There are many situations where random assignment of participants to treatment and comparison conditions may be unethical or impractical. This article provides an overview of propensity score techniques that can be used for estimating treatment effects in nonrandomized quasi-experimental studies. After reviewing the logic of propensity score methods, we call attention to the importance of the strong ignorability assumption and its implications. We then discuss the importance of identifying and measuring a sufficient set of baseline covariates upon which to base the propensity scores and illustrate approaches to that task in the design of a study of recovery high schools for adolescents treated for substance abuse. One novel approach for identifying important covariates that we suggest and demonstrate is to draw on the predictor-outcome correlations compiled in meta-analyses of prospective longitudinal correlations.

The current issue of this journal focuses on recovery programs, including recovery high schools and collegiate recovery communities. Schools are important social institutions in the lives of youth and adolescents, and are particularly crucial social contexts for students in recovery from substance use disorders. Peer pressure, association with substance-using peers, and availability of drugs are important risk factors for youth substance abuse, risk factors that are largely situated in school contexts for high school aged students. As such, traditional school contexts may be particularly risky environments for youth immediately after receipt of substance abuse treatment. Recovery high schools (or recovery programs located in traditional high schools) attempt to address this need by providing an educational environment designed to support recovery in a safe and protective setting that promotes sobriety and academic success. These recovery high schools both support the academic and therapeutic needs of students and work to build peer and family support structures that are supportive of recovery. Additional details about the recovery high school model and its theory of change are provided in other articles in this issue (Finch & Frieden, 2014/this issue; Finch & Karakos, 2014/this issue; Moberg & Finch, 2014/this issue).

Given the therapeutic and continuing care focus of recovery high schools, students with substance use disorders may benefit behaviorally and academically from attending these schools during and after substance abuse treatment. Despite their theoretical and intuitive appeal, however,

Correspondence should be sent to Emily E. Tanner-Smith, Vanderbilt University, Peabody Research Institute, Box 0181 GPC, Vanderbilt University, Nashville, TN 37203-5721. E-mail: e.tanner-smith@vanderbilt.edu

there have been no well-controlled studies to assess whether recovery high schools are more effective than traditional or other nonrecovery high schools in promoting positive outcomes among youth with substance use disorders. This is in large part due to the circumstances of participation in recovery high schools. There are many situations where randomly assigning participants to treatment and comparison conditions is difficult for practical or ethical reasons, and recovery high schools present one such situation. Many recovering students provided with the option to attend a recovery high school may not do so for reasons largely extraneous to their interest and motivation to avoid relapse. Transportation and location are often issues, for instance, or there may be special programs, academic course offerings, extracurricular opportunities, or other individually relevant circumstances at the high school they attended prior to treatment that override their interest in a recovery high school. Further, it would be difficult to convince many parents of youth with substance use disorders to agree to random assignment to a recovery or nonrecovery high school. Some parents, for instance, may be committed to sending their children to recovery high schools because of their presumed therapeutic benefits, and might therefore consider other options simply unacceptable. Other parents may perceive a stigma around a designated "recovery" high school and might therefore not consider this a viable option for fear it would limit their child's college or employment opportunities.

There are other situations in educational settings where random assignment would be difficult. For example, a state legislative mandate requiring that all eligible 4-year-olds have access to prekindergarten programs would require a randomized trial to assign some children to control conditions despite their legal right to enroll in prekindergarten—an arrangement the legislature and few parents would agree to. Similarly, an evaluation of the effects of school-level policies, programs, or interventions that are already entrenched or institutionalized would be difficult to conduct as a randomized experiment. A researcher interested in examining whether schools that employ school-resource officers have lower crime rates than schools without resource officers might, for instance, find that school administrators are unwilling to agree to randomization to such conditions because of political and safety concerns.

Fortunately, there are nonrandomized quasi-experimental research designs available to researchers that have the potential to produce valid estimates of intervention effects when well executed. Propensity score methods for equating nonrandomized intervention and comparison groups provide one family of useful techniques. In general, propensity score methods attempt to match treated and comparison units on a single composite score (the propensity score) that balances both groups on observed baseline characteristics. The goal of this balancing is to remove any selection bias that has made the groups different on those observed variables prior to intervention. Propensity scores are estimated for each individual participant in the sample and are calculated as the predicted probability of that participant being in the treatment group. Under certain stringent assumptions, propensity score techniques can provide unbiased estimates of treatment effects in nonrandomized studies and have, therefore, become increasingly popular among intervention researchers.

We are currently involved in a project funded by the National Institute on Drug Abuse that uses propensity score methods to compare behavioral and educational outcomes for students who attend recovery versus nonrecovery high schools after discharge from substance abuse treatment (Finch & Karakos, 2014/this issue). This quasi-experimental study was designed to collect data from a sample of adolescents with substance use disorders who had recently received treatment and then enrolled in high school in the greater Minneapolis/Saint Paul Twin Cities area (the study

was approved by the Institutional Review Boards at the University of Minnesota, University of Wisconsin, and Vanderbilt University). Because we knew random assignment of students to recovery versus nonrecovery high schools was not feasible but wanted to make causal inferences about the effects of recovery high school attendance on student outcomes, we designed this study to use propensity score methods to estimate the effects of recovery high school attendance.

This article first provides an overview of the use of propensity scores in nonrandomized quasi-experimental research designs with particular emphasis on the strong assumptions necessary for valid causal interpretation of the results. We then describe the importance of the identification and measurement of relevant baseline covariates and the associated implications for prospectively designing nonrandomized comparison group studies. With that in mind, we describe the novel approach we used to identify covariates in the design of the recovery high school evaluation, where we incorporated results from a meta-analysis of prospective longitudinal studies to identify the key predictors of the expected outcomes.

COUNTERFACTUAL FRAMEWORK FOR CAUSAL INFERENCE

The Neyman–Rubin counterfactual framework (Neyman, 1923; Rubin, 1974) provides a way of defining the causal estimates of interest in treatment outcome studies and the strong assumptions needed to enable causal interpretation of treatment effects when using propensity score methods. Under this causal model framework, each participant has two potential outcomes—one under the treatment condition and another under the control/comparison condition. In the recovery high school study, for example, one outcome of interest is academic performance as measured by grade point average (GPA). In this instance, each adolescent has one potential GPA outcome if he or she attends a recovery high school and another potential GPA outcome if he or she attends a nonrecovery high school. The causal effect of attending recovery high school relative to not attending for each student is the difference between these potential outcomes. The problem, of course, is that participants can experience either the treatment or the control condition during the intervention period, not both simultaneously, so only one of these potential outcomes can actually be observed. This has been called the "fundamental problem of causal inference" (Holland, 1986).

Although both potential outcomes for any given participant cannot be observed, the average treatment effect for a group of participants can be estimated by comparing their outcomes with those of a comparable group that did not receive the treatment. For that estimate to be a valid representation of the average treatment effect, however, the two groups must be well matched at the beginning of the intervention period on *every* characteristic, other than treatment exposure, that is capable of influencing the outcome. Random assignment of participants to conditions accomplishes this by making it a matter of chance which individuals with which characteristics are in each group. With a sufficient number of participants, the probability of a difference between the groups on any given characteristic that might produce a spurious appearance of a treatment effect is kept small and, further, that probability can be estimated with a test of statistical significance.

In nonrandomized quasi-experimental studies there is no such protection against selection of participants with different characteristics into treatment and control conditions. This vulnerability makes the simple difference in the outcome means a potentially biased estimator of the treatment effect, a problem known as selection bias. Under such circumstances, it is possible to estimate the

causal effects of treatment, but doing so requires meeting a stringent assumption called the strong ignorability assumption (also called conditional independence, unconfoundedness, or selection on observables; Imbens, 2004; Rosenbaum & Rubin, 1983). The strong ignorability assumption requires that all participants have a nonzero probability of being in either the treatment or control group, that outcomes for each participant be independent of those for other participants, and that the potential outcomes be independent of the type of treatment received conditional on a set of covariates. The key implication of the strong ignorability assumption is that every initial difference between the treatment and control groups that, absent treatment effects, might result in differences on the outcomes must be accounted for by baseline covariates that are included in the design and analysis. To meet the strong ignorability assumption, therefore, *the researcher must identify, measure, and properly account for all the variables that may (a) potentially differ between the treatment and control conditions and (b) influence any of the outcomes being examined*.

Statistical matching procedures aim to meet the assumption of strong ignorability by equating treatment and control groups on relevant baseline characteristics. There are many procedures available to researchers for matching individual cases or accomplishing comparable results via statistical controls (see Steiner & Cook, 2013, for a succinct review). Propensity score methods have distinct advantages for this purpose and their use is increasingly common in studies using nonrandomized comparison groups. The common theme across matching and statistical control techniques is the attempt to remove selection bias from causal effect estimates by equating treatment and control units on a sufficient set of measured covariates.

PROPENSITY SCORE TECHNIQUES

In the simplest case of matching, a researcher might match one treatment unit with one control unit based on some covariate presumed to be correlated with potential outcomes; for instance, each recovery school student might be matched with a student in a regular high school who has the same baseline GPA. But it is unlikely that any such single covariate would fully satisfy the strong ignorability assumption in a nonrandomized comparison. It is more likely that numerous baseline covariates will be necessary to account for all the differences between the groups that might influence the outcomes. Given the large number of relevant covariates that might be required, Rosenbaum and Rubin (1983) proposed the use of propensity scores as a single composite variable that incorporated all the relevant covariates. They proved that if treatment selection is strongly ignorable using a particular set of covariates, then it is also strongly ignorable using a propensity score based on those covariates. The main appeal of using a one-dimensional propensity score to characterize a multidimensional vector of covariates, therefore, is that it simplifies the statistical matching; it is much easier to match participants on the values of a single propensity score than on a number of individual covariate variables.

Propensity scores are typically generated using binomial regression models to predict treatment assignment from a set of observed baseline covariates. For each participant, the resulting propensity score estimate is the predicted probability, ranging from 0 to 1, of that individual being in the treatment group, that is, the propensity to be a treatment case. The goal of this procedure is to create propensity score estimates that balance the treatment and control conditions on a set of baseline covariates sufficient to satisfy the strong ignorability assumption. Unfortunately there is no simple rule to determine when acceptable balance has been reached, and various techniques

can be used to assess the adequacy of the propensity scores for equating the groups on all the individual covariates incorporated in the propensity score estimate (Rubin, 2001; Steiner & Cook, 2013). It is also often necessary for the researcher to identify and remove from the analysis any participants in either comparison group with extreme propensity score estimates that cannot be matched to participants in the other group.

After the researcher estimates the propensity scores and is confident that covariate balance has been attained, they can be used in a variety of ways to estimate the treatment effect. The most common methods are propensity score matching, propensity score subclassification, inverse propensity score weighting, and inclusion of the propensity score as a covariate in a regression model (Guo & Fraser, 2010; Steiner & Cook, 2013). These are among the numerous analytic decisions that must be made when using propensity score techniques to estimate treatment effects in a quasi-experimental study. Other analytic decisions include the selection of baseline covariates on which to base the propensity scores, the method and assumed functional form for estimating the propensity scores, the procedure for assessing covariate balance, and the analysis model for estimating the treatment effect. Here we focus on the first of those decisions—selecting baseline covariates for the propensity score estimation model—which is one of the most important but often underdeveloped steps in propensity score techniques.

THE ROLE AND IMPORTANCE OF BASELINE COVARIATES

What Types of Baseline Covariates Are Needed?

With propensity score techniques, or any procedure for matching nonrandomized treatment and control cases, the key to satisfying the strong ignorability assumption rests on the adequacy of the identification and measurement of the covariates that will be used in the propensity score estimation model. If this process is not done well, the set of covariates is not likely to be sufficient to eliminate selection bias and no choice of analytic techniques after that can compensate for the omission of critical covariates (Steiner, Cook, Shadish, & Clark, 2010). The propensity scores can only remove bias due to the covariates that have been included in those scores; they cannot remove any additional bias due to unobserved covariates, that is, those that were not measured and thus not available for inclusion in the propensity score estimation model. Therefore, selection and measurement of the relevant covariates is the most crucial step in propensity score techniques.

A sufficient set of covariates for the propensity score estimation model includes measures of all those characteristics of the treatment and control conditions that have the following two properties:

1. The characteristic is independently predictive of any outcome variable of interest either directly or via a relationship with another characteristic such that baseline differences between the conditions on that characteristic would produce differences between the conditions on the outcome net of any effect of the treatment.
2. The characteristic, or one with which it interacts, differs between the conditions at baseline. A variable that is not related in any way to the eventual outcome will have no biasing influence on the treatment effect estimate even if it is different for the treatment and control conditions; such a variable is neutral for purposes of estimating the treatment

effect. Conversely, a variable that is related in some way to the eventual outcome but is not different at baseline for the treatment and control conditions is already matched and will not bias the treatment effect estimate.

Note that to be effective in adjusting for selection bias, a covariate must be measured reliably. If it is measured with substantial measurement error (e.g., reliability $< .80$), it does not fully represent the characteristic at issue and thus cannot support accurate matches or full statistical control of that characteristic. Note also that the critical covariates are those that have an independent influence on the outcome, that is, influence the outcome above and beyond the influence of any other covariates included in the propensity score or other form of matching or statistical control. Covariates that are so highly correlated with other covariates in the analysis that they do not add anything to the ability of the set of covariates to predict the outcome already have their influence accounted for by those other covariates and therefore do not need to be included in the propensity score.

Strategies for Selecting Baseline Covariates

These requirements present a considerable challenge to a researcher designing a nonrandomized quasi-experiment. Because it is not possible in practice to test the assumption of strong ignorability, researchers must carefully consider the question of how they will identify and measure a set of covariates sufficient to ensure that any appreciable selection bias will be matched or adjusted away. Unfortunately, researchers using propensity score techniques often ignore or inadequately discuss the strong ignorability assumption and its stringent implications for the selection of baseline covariates.

Researchers conducting secondary analysis of an existing data set will be inherently limited in their ability to identify and use appropriate covariates by what is available in that data set. The treatment effect estimates in such cases may be especially suspect unless the available data include a convincing set of baseline covariates that can be used in the analysis. Researchers prospectively planning quasi-experimental studies using propensity score techniques, however, are in a position to strategically design baseline data collection on covariates that will support a reasonable claim that strong ignorability was achieved.

Covariates That Predict Selection

One approach to identifying a sufficient set of covariates is to focus on the generally unknown implicit selection process that sorts participants into treatment and control conditions. If all the variables for which group differences are produced by that selection process can be identified and measured or, at least, all of those independently related to selection, the result should be a set of covariates sufficient to create propensity scores capable of removing selection bias from the treatment effect estimates. This may require pilot studies with the target population to investigate the selection mechanisms at work. For example, prior to a study where undergraduates were allowed to self-select into a mathematics or vocabulary training program, Shadish, Clark, and Steiner (2008) interviewed student counselors about variables that might be associated with students' choice of program, which helped identify potentially important variables such as general

student preferences for mathematics versus literature. In addition to pilot studies, there may also be prior research or theory that can provide guidance.

In general, however, relatively little is known about the factors that influence exposure to different kinds of interventions, and it can be difficult to identify all the variables that are related to differential exposure. Moreover, the variables related to selection into treatment conditions may be very specific to the circumstances of those conditions and the associated participants. The factors that influence self-selection into recovery high schools or nonrecovery high schools by adolescents recovering from substance abuse, for instance, might vary across participants and encompass a broad range of geographical, peer, family, and personal characteristics, few of which might be related to differential exposure to any other educational program (e.g., a vocabulary training program).

Covariates That Predict Outcomes

Another approach to identifying a sufficient set of covariates for use in propensity scores is to focus on variables that are predictive of the outcomes of interest. If all the baseline variables directly or indirectly related to those outcomes could be identified, that collection of covariates would necessarily include a set (possibly a subset) sufficient to account for selection bias. Only those that differed for the treatment and control conditions would be essential for that purpose. For many outcomes of interest in intervention studies, there is considerable prior research that reports correlates of those outcomes. With adolescent populations, for example, longitudinal studies have provided a great deal of information about the predictors of academic achievement, dropout, substance use, delinquent behavior, and other such outcomes that are often targeted by interventions. Indeed, the whole topic of identifying risk, protective, and promotive factors predictive of such outcomes is a major area of research in its own right. For developing adequate propensity scores in many intervention contexts, therefore, a focus on identifying a full set of independent predictors of the outcomes of interest may be more productive than a focus on identifying the full set of predictors of selection into the respective comparison conditions. Of course, these two approaches are not mutually exclusive, and both deserve attention when selecting baseline covariates for a nonrandomized quasi-experiment.

IDENTIFYING POTENTIALLY RELEVANT BASELINE COVARIATES IN THE DESIGN OF THE RECOVERY HIGH SCHOOL STUDY

In the quasi-experimental study designed to examine the effects of recovery high schools on student outcomes (Finch & Karakos, 2014), we recognized that valid effect estimates would be possible only if we collected good baseline data on a set of covariates fully capable of accounting for the selection bias inherent in comparing outcomes for youth who chose to attend recovery high schools with those who did not. We were therefore very deliberate, systematic, and expansive in our procedures for identifying candidate covariates to include in our baseline measurement battery. Although we might not go so far as to offer our approach to this challenge as a model for other researchers, we do believe it illustrates the nature of the effort that is required and stands in

contrast to the rather superficial attention paid to this matter in the design of many nonrandomized quasi-experiments.

We began with consideration of the possible selection mechanisms for attending recovery versus other high schools after discharge from substance abuse treatment. This is a particularly difficult question because attending a recovery high school is largely a matter of self-selection by the students and their families under circumstances where they have some awareness of, and potential reactions to, what each of the options entails. In particular, there are almost certainly differences in the attitudes, motivation, knowledge, personality, and other such personal characteristics of the participants and their parents that are very specific to their orientation to recovery from substance abuse and their perceptions of recovery high schools. Such highly contextualized self-selection factors are especially difficult to identify and measure well. Other potential selection factors are easier to apprehend. The location of the recovery high schools relative to participants' homes and the accessibility of transportation to school, for instance, are likely to be relevant and are fairly straightforward.

To identify covariates potentially related to selection into recovery high schools, we drew upon the insights of members of our team with substantial experience with recovery high schools and the students who attend. Those experiences include prior employment in a recovery high school, ongoing participation in the professional association for recovery high school personnel, and extensive prior site visits and interviews with recovery high school personnel and students in the geographic region where the current quasi-experimental study is being conducted (Moberg & Finch, 2008). These key informants identified a number of factors they believed to be related to the decision about whether to attend a recovery high school, including students' and parents' awareness and knowledge of recovery high schools, students' and parents' interest in attending a recovery high school, proximity of a recovery school, cost (some programs are private schools), readiness to change substance use behaviors, interest in obtaining continuing care services, denial of an ongoing substance use disorder, and willingness to accept recovery support. These motivational, attitudinal, and practical variables were then represented in questionnaire measures that were included in the baseline survey of study participants and their parents.

The difficulty of identifying all the covariates that might be related to the selection of recovery versus nonrecovery high school, however, made this approach unsatisfactory as our sole method for identifying baseline covariates for inclusion in the propensity score estimation model. We therefore put our primary emphasis on the identification of variables predictive of the major outcomes of interest, which were students' academic performance and substance use. If a relatively complete set of such predictors could be identified and measured, we can expect one or more of those predictors to be correlated with any selection variable related to outcomes that we failed to include. For this purpose, we turned to longitudinal research on the predictors of academic performance and/or substance use for high school aged youth. An effective approach to identifying these covariates, when possible, is to utilize meta-analytic findings that describe the magnitude of the predictive relationships with the outcomes of interest.

For the design of the recovery high school study, we were able to use information from a large meta-analysis on the risk and protective factors for delinquency, substance use, and academic success in adolescents to identify a wide range of known predictive factors for the outcomes of interest (Tanner-Smith, Wilson, & Lipsey, 2013). This meta-analysis was based on a comprehensive systematic review of literature published through 2002 that examined the longitudinal correlations between risk factors and substance use or academic performance during

adolescence and early adulthood. For the purpose of identifying baseline covariates for inclusion in the quasi-experimental study of recovery high schools, we examined correlation coefficients indexing the longitudinal relationships between risk factors measured between ages 11 and 14 and subsequent substance use or academic performance outcomes measured at least 6 months later between the ages of 14 and 18.

The meta-analytic database included more than 5,000 correlation coefficients from 119 longitudinal studies that indexed relationships between risk factors and subsequent substance use measured in the target age ranges. The database also included more than 2,900 correlation coefficients from 416 longitudinal studies indexing the relationships between risk factors and subsequent academic performance measured in the target age ranges. These predictive risk factors were sorted into more than 70 narrow construct categories in the meta-analytic database (e.g., teacher–student relations, family structure, academic anxiety, quality of peer relationships, association with delinquent peers). To identify important baseline covariates for use in the study of recovery high schools, we further categorized these risk factors into 18 broad macrolevel construct families (e.g., impulsivity, peer behavior and influences, school motivation and attitudes). We then identified any risk factor families with an average predictive correlation for later substance use or academic performance that was greater than or equal to .30. Specifically, we used multilevel inverse-variance weighted mixed-effects metaregression models to estimate the average longitudinal correlation between a given risk factor construct family and subsequent substance use or academic performance. All these estimated mean correlations additionally adjusted for the age of participants, the interval between the two longitudinal waves of data collection, the scaling of each construct, the reporting source of each construct, and the form of data collection (e.g., self-report, parent-report).

These analyses therefore identified the strongest groups of predictors for adolescent substance use and/or academic performance found in an extensive set of longitudinal studies. As would be expected, the strongest predictors for adolescent substance use overlapped to a significant degree with those for academic performance. Overall, the predictor families with the largest mean correlations with later substance use and academic performance were as follows:

- Antisocial attitudes;
- Antisocial behavior (problem behavior, delinquency);
- Drug exposure and attitudes (attitudes toward substance use, intentions to use drugs or alcohol);
- Family antisocial behavior, substance use;
- Family and household characteristics (socioeconomic status);
- Impulsiveness, hyperactivity;
- Internalizing behavior/symptoms;
- Parenting, parental practices (negative parenting, poor parent skills, weak family cohesion);
- Peer behaviors and influences (peer antisocial behavior, peer substance use, peer attitudes toward substance use; peer antisocial behavior, availability of drugs from peers);
- Prior school performance;
- Prior substance use;
- Religiosity;
- School motivation and attitudes (school bonding, engagement, effort);

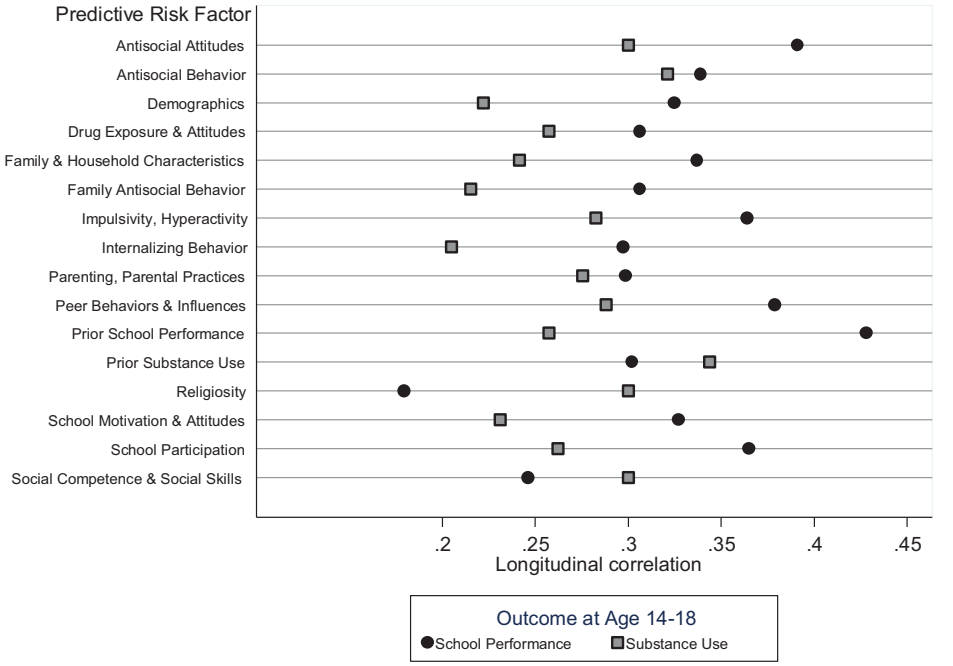

FIGURE 1 Mean longitudinal correlations between risk predictors measured at ages 11 to 14 and substance use or school performance outcomes measured at ages 14 to 18 (color figure available online).

- School participation (truancy, attendance);
- Social competence and social skills; and
- Demographic characteristics (gender, ethnicity, age).

Figure 1 shows the average longitudinal correlations with substance use and academic performance for these families of predictive risk factors. Overall, the strongest predictors of adolescent substance use were prior substance use and antisocial behavior; the strongest predictors of academic performance were prior school performance and antisocial attitudes.

The large number of studies and the extensive evidence from this meta-analysis provided a firm basis for identifying a set of baseline covariates that, arguably, was sufficient to account for any selection bias in the planned quasi-experimental study of recovery high schools. Though important variables that were independently related to either of the outcomes of interest might have been overlooked in the longitudinal studies included in this meta-analysis, it is difficult to imagine what they might be. Such variables not only would have to have been neglected by the researchers studying predictive factors for these outcomes but also would have to be substantially uncorrelated with the variables already identified.

With confidence that we had a sufficient set of covariates to meet the strong ignorability assumption for propensity scores, we developed the baseline assessment battery for the recovery high school study to include measures of each of the strongest overall predictive factors identified above and in Figure 1. As shown in Table 1, each of these predictive factors was measured at

TABLE 1
Summary of Baseline Assessment Battery Measures in the Recovery High School Project, Identified as Correlates of Treatment Outcomes of Interest

Construct	Instrument	Example Measures
Antisocial Attitudes	Interview, MINI-SCID, PEI	Attitudes toward crime/delinquency
		Perceived benefits of substance use
		Antisocial personality disorder symptoms
Antisocial Behavior	Interview	Antisocial and criminal behavior
		Juvenile justice system involvement
Demographics	Interview	Age, race/ethnicity, gender
Drug Exposure & Attitudes	ADI; MTF; PEI	Perceived availability of substances
		Attitudes toward substances
		Perceived benefits/harms of substances
		Exposure to substances
Family & Household Characteristics	Interview	Parent occupation
		Parent education level
		Household income
Family Antisocial Behavior	ADI	Family history of substance use, substance use treatment
		Family history of mental health problems
Impulsivity, Hyperactivity	Interview	Attention deficit hyperactivity disorder symptoms
Internalizing Behavior	ADI, GAIN, MINI-SCID, PSI, Interview	Life satisfaction
		Life stressors
		Problem solving strategies
		Behavioral problems
		Psychiatric disorder symptoms/diagnoses (e.g., depression, suicidality, social phobia)
Parenting, Parental Practices	APQ, YHPS	Harsh, negative parenting
		Consistent parenting practices
		Parental warmth, support
Peer Behaviors & Influences	PEI, PSUT	Peer attitudes toward drugs
		Peer antiestablishment attitudes
Religiosity	GAIN	Religious identity
		Religious service attendance
		Perceived strength and importance of religious beliefs
School Motivation & Attitudes	BASC, HSQ	Perceived academic abilities
		Attitudes toward school/learning
School Participation	GAIN, HSQ	School attendance/absenteeism, tardiness, truancy
School Performance	Interview, HSQ	Perceived problems with school grades
		Grade point average
		Self-reported grades
Social Competence & Social Skills	GAIN, HFL, Interview	Access to social support from friends
		Time spent with friends/romantic partners
		Time spent participating in social activities
		Perceived social competence
Substance Use	Interview, ADI, MINI-SCID, TLFB	Substance use (alcohol, cannabis, other specific substances), substance use treatment history

Note. ADI = Adolescent Diagnostic Interview (Winters & Henly, 1993); APQ = Alabama Parenting Questionnaire (Shelton, Frick, & Wootton, 1996); BASC = Behavior Assessment System for Children (Reynolds & Kamphaus, 1992); GAIN = Global Appraisal of Individual Needs (Dennis, Titus, White, Unsicker, & Hodgkins, 2003); HFL = Healthy for Life (Piper, Moberg, & King, 2000); HSQ = High School Questionnaire (Moberg & Finch, 2008); MTF = Monitoring the Future (Johnston, O'Malley, Bachman, & Schulenberg, 2011); MINI-SCID (Sheehan, Shytle, & Milo, 2006); PEI = Personal Experiences Inventory (Winters & Henly, 1989); PSI = Problem Solving Inventory (Latimer, Winters, D'Zurilla, & Nichols, 2003); PSUT = Peer Substance Use Test (Chassin, Pillow, Curran, & Molina, 1993); TLFB = Timeline Follow Back (Sobell & Sobell, 1995); YHPS = Youth Happiness with Parent Scale (De Cato, Donohue, Azrin, & Teichner, 2001).

baseline with at least one instrument or scale and in many instances multiple measures were used. Moreover, each predictive factor was measured with both youth and parent interviews or questionnaires, providing another layer of multiple measures. Multiple measures can be combined to improve the reliability of measurement and thus further support the effectiveness of the resulting propensity scores for reducing selection bias.

SUMMARY

Propensity score methods can be used to estimate the causal effects of intervention in nonrandomized quasi-experimental designs, but doing so requires meeting the relatively stringent assumption of strong ignorability. To meet that assumption, the researcher must identify and reliably measure at baseline a set of covariates on which to base the propensity scores that is sufficient to fully account for any selection bias. It is therefore crucial for researchers prospectively planning nonrandomized intervention studies that will use propensity scores to carefully consider how best to identify these key covariates and ensure their reliable measurement during baseline data collection. Relevant covariates are those that differ at baseline between the intervention and comparison groups in the study and are related to the outcome variables. Identification of those covariates can thus focus on the factors that characterize the selection process or the factors that are predictive of the outcome or, ideally, both.

In this article we emphasized the difficulty of meeting the strong ignorability assumption when using propensity scores and the critical importance of a thorough and systematic effort to identify a sufficient set of covariates for use in those scores. We illustrated the various approaches to this task for the design of a study of recovery high schools. In that context, we described a novel strategy for identifying covariates correlated with the outcomes of interest—using results from a meta-analysis of prospective longitudinal correlations between predictive factors and later outcomes. This approach allowed us to identify the strongest predictors of academic and substance use outcomes among adolescents. Those predictors, along with variables we hypothesized were related to the selection process, were then included in an extensive baseline assessment battery that arguably will support propensity scores that meet the stringent strong ignorability assumption and allow us to generate unbiased treatment effect estimates.

Researchers intending to use propensity score techniques to estimate causal treatment effects must carefully consider the importance of identifying and reliably measuring an adequate set of covariates for use in their propensity scores. Although researchers conducting secondary analyses of existing data sets will be inherently limited in this regard, there are several strategies available to researchers prospectively planning quasi-experimental studies. First, researchers should always use prior theory to guide covariate identification. Pilot research can also be used to identify variables correlated with the selection process and/or outcome constructs (e.g., attitudinal and motivational measures). Another promising strategy, detailed in this article, is to identify covariates based on prior empirical research. This might involve reviewing findings from systematic reviews or meta-analyses on the topic or, at minimum (when a systematic review may be unavailable), reviewing findings from individual research studies that have examined correlates of treatment selection or treatment outcomes relevant to the planned study. With that information in hand, researchers can then carefully plan their baseline data collection efforts in an attempt to

meet the stringent assumptions necessary for their propensity score analysis techniques to yield unbiased estimates of treatment effects.

AUTHOR BIOS

Emily E. Tanner-Smith is a Research Assistant Professor at the Peabody Research Institute and Department of Human and Organizational Development at Vanderbilt University. Her broad areas of expertise include the social epidemiology, prevention, and treatment of adolescent substance use. Her recent research appears in the *Oxford Handbook of Criminological Theory*, *Journal of Substance Abuse Treatment*, *Prevention Science*, and *Research Synthesis Methods*.

Mark W. Lipsey is the Director of the Peabody Research Institute and a Research Professor at Vanderbilt University. He specializes in program evaluation with a focus on programs for at-risk children and youth.

FUNDING

This publication was made possible by Grant Number R01DA029785-01A1 from the National Institute on Drug Abuse. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Institute on Drug Abuse or National Institutes of Health.

REFERENCES

Chassin, L., Pillow, D. R., Curran, P. J., & Molina, B. (1993). Relation of parental alcoholism to early adolescent substance use: A test of three mediating mechanisms. *Journal of Abnormal Psychology*, 102, 3–19.

De Cato, L. A., Donohue, B., Azrin, N. H., & Teichner, G. A. (2001). Satisfaction of conduct-disordered and substance-abusing youth with their parents. *Behavior Modification*, 25, 44–61.

Dennis, M. L., Titus, J. C., White, M. K., Unsicker, J. I., & Hodgkins, D. (2003). *Global appraisal of individual needs: Administration guide for the GAIN and related measures*. Bloomington, IL: Chestnut Health Systems.

Finch, A. J., & Frieden, G. (2014/this issue). The ecological and developmental role of recovery high schools. *Peabody Journal of Education*, 89, 271–287.

Finch, A. J., & Karakos, H. (2014/this issue). Substance abuse and recovery and schooling: The role of recovery high schools and collegiate recovery communities. *Peabody Journal of Education*, 89, 159–164.

Guo, S., & Fraser, M. W. (2010). *Propensity score analysis: Statistical methods and applications*. Thousand Oaks, CA: Sage.

Holland, P. (1986). Statistics and causal inference (with discussion). *Journal of the American Statistical Association*, 81, 945–970.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86, 4–29.

Johnston, L. D., O'Malley, P. M., Bachman, J. G., & Schulenberg, J. E. (2011). *Monitoring the Future national survey results on drug use, 1975–2010*. Ann Arbor: Institute for Social Research, The University of Michigan.

Latimer, W. W., Winters, K. C., D'Zurilla, T., & Nichols, M. (2003). Integrated family and cognitive-behavioral therapy for adolescent substance abusers: A Stage I efficacy study. *Drug and Alcohol Dependence*, 71, 303–317.

Moberg, D. P., & Finch, A. J. (2008). Recovery high schools: A descriptive study of school programs and students. *Journal of Groups in Addiction & Recovery*, 2, 128–161.

Moberg, D. P., & Finch, A. J. (2014/this issue). Recovery high schools: Student and responsive academic and therapeutic services. *Peabody Journal of Education*, 89, 165–182.

Neyman, J. S. (1923). Statistical problems in agricultural experiments. *Journal of the Royal Statistical Society, Series B*, 2, 107–180.

Piper, D. L., Moberg, D. P., & King, M. J. (2000). The healthy for life project: Behavioral outcomes. *Journal of Primary Prevention*, 21, 47–73.

Reynolds, C. R., & Kamphaus, R. W. (1992). *BASC: Behavior Assessment System for Children: Manual*. Circle Pines, MN: American Guidance Service.

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

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66, 688–701.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services Outcome Research Methodology*, 2, 169–188.

Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random to nonrandom assignment. *Journal of the American Statistical Association*, 103, 1334–1343.

Sheehan, D. V., Shytle, D., & Milo, K. (2006). *MINI International Neuropsychiatric Interview for Children and Adolescents English Version 5.0*. Tampa, FL: University of South Florida.

Shelton, K. K., Frick, P. J., & Wootton, J. (1996). Assessment of parenting practices in families of elementary school-age children. *Journal of Clinical Child Psychology*, 25, 317–329.

Sobell, L. C., & Sobell, M. B. (1995). *Alcohol Timeline Followback Users' Manual*. Toronto, Canada: Addiction Research Foundation.

Steiner, P. M., & Cook, D. (2013). Matching and propensity scores. ln T. D. Little (Ed.), *The Oxford handbook of quantitative methods: Volume 1 foundations* (pp. 237–259). Oxford, UK: Oxford University Press.

Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods*, 15, 250–267.

Tanner-Smith, E. E., Wilson, S. J., & Lipsey, M. W. (2013). Risk factors and crime. In F. T. Cullen & P. Wilcox (Eds.), *The Oxford handbook of criminological theory* (pp. 89–111). New York, NY: Oxford University Press.

Winters, K. C., & Henly, G. A. (1989). *Personal Experience Inventory (PEI): Manual*. Torrance, CA: Western Psychological Services.

Winters, K. C., & Henly, G. A. (1993). *Adolescent Diagnostic Interview (ADI): Manual*. Torrance, CA: Western Psychological Services.