Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 18, Number 15, December 2013 ISSN 1531-7714

Educational Research with Real-World Data: Reducing Selection Bias with Propensity Scores

Jill L. Adelson, *University of Louisville*

Often it is infeasible or unethical to use random assignment in educational settings to study important constructs and questions. Hence, educational research often uses observational data, such as large-scale secondary data sets and state and school district data, and quasi-experimental designs. One method of reducing selection bias in estimations of treatment effects is propensity score analysis. This method reduces a large number of pretreatment covariates to a single scalar function and allows researchers to compare subjects with similar probability to receive the treatment. This article provides an introduction to propensity score analysis and stratification, an example illustrating its use, and suggestions for using propensity score analysis in educational research.

To meet the needs of students, educational researchers have a responsibility to conduct sound research, particularly on interventions, programs, and policies aimed at effective teaching and learning. The results of such studies are significant to researchers, teachers, administrators, parents, and policymakers. The challenge educational researchers face is "to develop, test, and refine" interventions and theories in a methodologically rigorous manner that maintains the field's "scientific integrity" (Graesser, 2009, p. 259). However, the most ideal research designs are typically not feasible to employ in educational settings. For instance, school structures, student needs, or economic constraints may limit the possibility to assign students randomly to receive or to not receive a particular intervention (such as gifted programming or special education). Similarly, it frequently is not possible to assign schools randomly to adopt a particular policy (such as full-day kindergarten, ability grouping, self-contained gifted or special education classes, or employing a full-time school counselor). In fact, in many cases, a school administrator or parent decides on the treatment for particular students, classes, or schools on a nonrandom basis. Therefore, researchers studying important educational polices, issues, and programs that affect students' learning, emotional well-being, and social development frequently use observational data and quasi-experimental designs that rely on comparison groups that may or may not be similar to the treatment groups.

As Graesser (2009) noted, "educational settings are inherently complex, so there is a delicate balance between preserving the methodological rigor of our research designs and testing the students in ecologically valid learning environments" (p. 259). Compared to in randomized studies, the students, settings, and treatments in quasi-experiments may be more representative of the real-world condition that the researcher wishes to study (Shadish, Luellen, & Clark, 2006). However, when randomization is not used, treatment and comparison groups (whether they be students, classes, or schools) may differ in their background characteristics. That is, students whom the policy affects or who receive the "treatment" may be systematically different from those who do not. Similarly, classes or schools that choose to implement a particular program or curriculum or that adopt a certain policy typically differ from those that do not make the

same policy or program decisions[1]. Those pretreatment differences may cause a difference in outcomes, rather than the treatment itself causing the difference. This is particularly true in education, a field in which many covariates (e.g., prior achievement, motivation, socio-economic status, home support) affect outcomes like achievement.

Although the large number of covariates can be daunting and statistically challenging with traditional methods, propensity score analysis offers an alternative approach that can balance treatment and comparison groups on many covariates. Accordingly, this article first provides a conceptual basis of causal inference and then a rationale for and accessible introduction to propensity score analysis in general and propensity score stratification in particular. To illustrate the method, I then present an analysis of the restricted-use Early Childhood Longitudinal Study, Kindergarten Cohort Class of 1998-1999 (ECLS-K) data to investigate the effects of providing gifted programming in reading, an approach advocated by some for meeting the academic, social, and emotional needs of talented children (e.g., Delcourt, Cornell, & Goldberg, 2007; Marsh, Hau, & Craven, 2004; Rogers, 2007) but critiqued by others (e.g., de Vise, 2008; Grant, 2002; Sapon-Shevin, 1993). The article concludes with recommendations for using propensity score analysis to advance educational theory and research.

Randomized Studies and Causal Inference

In a randomized study, the gold standard in research, each student has the same probability to be in the treatment group. This ensures that, over the long run, the groups are comparable prior to treatment (i.e., that the background characteristics of the treatment and comparison groups are the same) so that a difference in outcomes reflects treatment effects. The key characteristic of randomization is that it ensures that the assignment to treatment or comparison group is unrelated to the background characteristics, allowing statistical tests to indicate if a treatment effect is demonstrated.

When researchers randomly assign students to a treatment or comparison group, assignment is unrelated to the students' background characteristics. Therefore, the researchers have met one component of the stable unit treatment value assumption (SUTVA; Rubin, 1978, 1986). This a priori assumption states that a student's outcome value (such as achievement score) when exposed to the treatment would be the same regardless of how the student was assigned to treatment and regardless of what treatments the other students received (Rubin, 1986). To determine the treatment effect, Δ_i, for student *i* under SUTVA, a researcher would calculate the student's potential outcome, $Y_i(Z)$, in the comparison condition (Z=0) and subtract that from the *same* student's potential outcome in the treatment condition (Z=1): $\Delta_i = Y_i(1) - Y_i(0)$. However, it typically is impossible to observe a student's outcome in both the treatment and comparison conditions, therefore making the effect of the treatment on the individual unobservable (Holland, 1986).

Under Rubin's Causal Model, researchers could estimate the average population treatment effect (δ) using the expected outcomes in treatment and comparison conditions ($E[\Delta]$). Similar to the above expected outcome for individuals, the expected outcome in the conditions is equal to the expected value of the difference in individual treatment effect over all individuals:

$$\delta = E[\Delta] = E[Y(1) - Y(0)] = E[Y(1)] - E[Y(0)].$$

The final step, $E[Y(1)] - E[Y(0)]$, indicates that the observed values on different units in the two conditions can be observed to estimate the average treatment effect. However, this is only true under the assumption of independence, or SUTVA. That is, causal inference can only be made if assignment to treatment or policy implementation is statistically independent of all other variables, as in randomization (Holland, 1986).

However, randomization oftentimes is not feasible or ethical in educational research, resulting in the use of observational data and quasi-experiments to research important questions regarding effects of and on psychological mechanisms that are crucial to students' learning and well-being, such as bilingual programs (Branum-Martin, Foorman, Francis, & Mehta, 2010), age mixing and age segregation (Allen, 1989), and role-playing pedagogy (Stroessner, Beckerman, & Whittaker,

[1] Educational researchers study not only students but also classes, schools, and districts. As noted, random assignment, "treatment" or policy/program implementation, and outcomes of interest can be at any of these levels. For simplicity, this article refers to "students" as the subject of the studies in the general discussion of propensity score analysis. However, these same principles and methods could be applied to any research subjects, such as in the example analysis provided.

2009). When randomization is not used, treatment typically is *not* independent of all other variables; that is, students are chosen for a particular treatment or are affected by policy based on their background characteristics. Therefore, SUTVA is not met. As a result, researchers need analytical tools to adjust for these systematic differences between treatment and comparison groups with respect to one or more pretreatment characteristics, which are referred to as selection bias (Braitman & Rosenbaum, 2002).

Introduction to Propensity Score Analysis

Researchers have a range of analytical adjustments they can make for selection bias. They may focus on the relationship between pretreatment variables and outcomes by modeling the response directly through methods such as regression. However, another option is to focus on the relationship between pretreatment variables and assignment to treatment in an effort to reconstruct a situation similar to random assignment after the fact (Braitman & Rosenbaum, 2002). The propensity score can be used to model this relationship between pretreatment variables and treatment assignment as it represents the conditional probability of assignment to treatment based on measured pretreatment characteristics (Rosenbaum & Rubin, 1983b). For example, given pretreatment variables such as prior achievement, socio-economic status, mother's education, motivation, and teacher's rating of academic skills, a researcher could determine the probability of students being in a gifted program (assignment to treatment), and this conditional probability would be represented by the propensity score. Similar to randomization in which the researcher knows the probability a student will be in the treatment group, the researcher now has an estimate of the student's probability to be in the treatment group, given their background characteristics. Thus, the process of adjusting for pretreatment variables, or confounders, through a propensity score could be viewed as "a means of obtaining quasi-randomization of treatment groups to minimize bias and to better estimate the true effects of treatment" (Newgard, Hedges, Arthur, & Mullins, 2004, p. 954).

Unlike conventional multivariable techniques that typically include all the pretreatment variables in the statitistical analysis of treatment efficacy, the propensity score controls for systematic differences in background characteristics between the treatment and comparison groups that would not occur in a randomized experiment by "reducing the entire collection of background charactersitics to a single composite characteristic that appropriately summarizes the collection" (Rubin, 1997, p. 757). By being able to reduce the number of variables by such a tremendous amount, researchers have more degrees of freedom for estimating treatment effects and are less likely to suffer from instable models, misleading results, or statistical inefficiency (Newgard et al., 2004). Additionally, although ANCOVA may be effective when the distributions of the covariates are similar among treatment and comparison students, propensity score methods are better when the groups are very different (Dehejia & Wahba, 1999; Rubin, 1997), as they often are in observational and quasi-experimental research. In fact, treatment and comparison groups differing greatly in the distribution of the pretreatment covariates may violate the fundamental assumption of regression models – that covariates and the outcome have a linear relationship – thus producing unreliable results (Newgard et al., 2004). On the other hand, using the propensity score to estimate treatment effects does not rely on any particular form, such as linearity, for the relationship between outcome and pretreatment covariates within each group (Rubin, 1997).

Statistically, the propensity score, $e(x)$, is the conditional probability of receiving the treatment given the observed pretreatment variables, x, denoted as

$$e(x) = \text{prob}(z = 1 \mid x),$$

where $z = 1$ for subjects in the treatment group and $z = 0$ for subjects in the comparison group (Rosenbaum & Rubin, 1983b). For subjects with the same propensity score (when the propensity score is held constant), the joint distribution of the observed covariates is balanced between the treatment and comparison groups. Similar to a randomized study, the propensity score adjusts for the observed covariates, and the researcher assumes conditional independence, or strongly ignorable treatment assignment given those observed covariates (Rosenbaum & Rubin, 1983b).

In a randomized study, all subjects typically have a probability of .5 to receive treatment. However, in quasi-experiments, bias is introduced when comparing groups because they do not have equal probability of receiving the treatment. For instance, using a secondary data set to examine the effect of kindergarten retention policy, Hong and Raudenbush (2005) found that non-Hispanic males from a family with a lower socio-economic status were more likely to be retained in kindergarten while children who did not

have a disability and who had parents who showed higher levels of commitment to parenting responsibility were less likely to be retained. The propensity score allows researchers to compare those with equal probability of receiving treatment but in different groups (treatment or comparison), thus allowing researchers to address the problem created by subjects not being randomly assigned (Rosenbaum & Rubin, 1983b). When subjects have an equal chance of receiving treatment based on their observed pretreatment variables, these variables will not help predict which of the subjects receive treatment. This means that the subjects with equal propensity for treatment should be balanced in the pretreatment variables (Rosenbaum, 2002). Because the propensity score is a probability, its values range from 0 to 1, with 0 indicating that the subject has no chance of receiving the treatment and 1 indicating that the subject will, without a doubt, receive the treatment.

Considerations in Developing Propensity Scores

Propensity score analysis works best for larger samples (Rubin, 2007). With larger samples, the researcher can include a more extensive set of variables and better estimate the propensity score model and, like in randomization, is more likely to balance the treatment and control groups. Additionally, if a method like one-to-one matching is used, there may be a substantial decrease in sample size due to a greatly unbalanced proportion of treatment and control subjects or because many observations do not have suitable matches and must be deleted. Unfortunately, research is needed to provide guidelines on how large is "large" in terms of propensity score sample size (Schafer & Kang, 2008).

An important (perhaps *the* most important) consideration for researchers is which covariates to include in the propensity score model (Shadish, Clark, & Steiner, 2008). As noted by Rubin (1997) and Newgard et al. (2004), researchers should include even weakly predictive pretreatment variables when constructing the propensity score as the biasing effects of omitting them may override the statistical efficiency gains of not including them. In fact, researchers should remember that the goal is to match treatment and comparison subjects on as many *theoretically relevant* pretreatment variables as possible, making the propensity score as rich and complete as possible. Parsimony is not necessary because in estimating the treatment effect the propensity score acts as a scalar function and summarizes the collection of pretreatment variables. By including a rich set of interrelated, diverse covariates, the researcher might avert the negative effects of not including a hidden covariate by including available covariates that are related to that unobserved treatment selection variable (Stone & Tang, 2013). Rosenbaum (2002) cautioned against only using covariates that are statistically significantly different between the treatment and comparison groups because (a) this does not take into account the relationship between the covariate and the outcome, (b) the statistical test relies heavily on sample size and does not indicate practical relevance, and (c) the covariates are considered in isolation rather than collectively. Based on Monte Carlo simulation studies, both Brookhart et al. (2006) and Adelson, McCoach, and Rogers (2009) recommended including variables related to both treatment and to outcome, even if only weakly related to one or both of those variables. Furthermore, for propensity score adjustment to adequately reduce bias, scores cannot be predicted only from predictors of convenience, such as age, sex, and ethnicity (Shadish, Clark, & Steiner, 2008). Rather, researchers must include substantively important variables like pretest scores. Additionally, only pretreatment variables and *not* the outcome variable or variables measured during or after treatment should be used to generate the propensity score to avoid introducing bias and to establish temporal precedence (Rosenbaum & Rubin, 1983b; Schneider, Carnoy, Kilpatrick, Schmidt, & Shavelson, 2007). Although a rich set of covariates is desired, the researcher "must be sensitive to the nature of the data at hand and the possibility of violations of assumptions" (Guo & Fraser, 2010, p. 137). For instance, the researcher must conduct routine diagnostic analyses, examining issues such as multicollinearity and tests of influential observations, thus assessing the fit of the final model.

Once a large number of pretreatment variables related to treatment or outcome are identified, the propensity score model often is estimated by entering all those covariates into a binary logistic regression with the dependent variable being the treatment or comparison group. After estimating propensity scores, researchers can use them in a number of ways, including matching (see Rudner & Peyton, 2006, for an example), inverse-propensity weighting, stratification (or subclassification), or in a dual-model strategy in which ANCOVA is applied in a propensity-matched sample or within propensity-defined strata. The estimation, uses, pros, and cons of each of these

various methods are beyond the scope of this article, and interested readers should consult Guo and Fraser (2010) and Schafer and Kang (2008). However, this article will briefly describe one method, stratification.

Propensity Score Stratification

An issue with one-to-one matching, particularly when the number of students receiving the treatment is small compared to the number of comparison students (or vice versa), is that many of the subjects are not used, resulting in an analysis sample size equal to the number of subjects in the treatment or comparison group (whichever is less). Stratification, or grouping subjects with similar propensity to be in the treatment group, is a commonly used method that includes the majority, if not all, of the subjects and controls for examining systematic differences. Rosenbaum and Rubin (1984) noted that using stratification on estimated propensity scores has several advantages, including (a) being easy to implement, (b) being easy to interpret , (c) often being convincing to nontechnical audiences, and (d) easily accommodating additional adjustments, such as controlling for during-treatment covariates. Although stratification on individual covariates can get unwieldy, as noted previously, the propensity score is a scalar function of covariates that "summarizes the information required to balance the distribution of the covariates" (Rosenbaum & Rubin, 1984, p. 516). That is, researchers can use only the propensity score, regardless of the number of observed pretreatment covariates used to construct it, to form strata that will balance the covariates between the treatment and comparison groups (Rosenbaum & Rubin, 1983b). To develop strata, after propensity scores are estimated, cases are stratified on the logit of the propensity score. Readers interested in further details about stratifying cases should consult D'Agostino (1998), Guo and Fraser (2010), or Rubin (1997). Once researchers stratify subjects, they can conduct comparisons of treatment and comparison subjects within the same strata, thus controlling for overt selection bias (Rosenbaum & Rubin, 1984).

A key step that the researcher must take after creating propensity score strata is to check balance in the strata on the propensity score and on each covariate. Recall that a key feature of propensity scores is that they balance the distribution of the covariates, making them useful for causal inferences (Rosenbaum & Rubin, 1983b). However, the distributional balance of the covariates is *expected,* not guaranteed, similar to in randomization. Substantial random imbalances of some covariates can happen in both random experiments and quasi-experiments, especially with small sample sizes. To check for balance in the covariates, the researcher can regress each covariate as well as the logit of the propensity score on the treatment assignment, controlling for S-1 dummy indicators for the S propensity strata and their interactions with treatment assignment. Alternatively, the researcher may conduct 2 x S (Conditions x Strata) ANOVAs, using both the propensity score and each predictor individually as dependent variables. If strata are balanced, then there should not be a difference between groups on the propensity score or the covariates. Less than 5% of the analyses (the percentage expected by chance, assuming a Type I error rate of $\alpha = .05$) should indicate statistically significant differences between treatment and control group for the stratum (statistically significant regression weights in the first method or statistically significant Condition x Strata interactions in the second method). Should the strata not balance the covariates, the researcher may choose to restratify or to add more terms, such as excluded predictors, nonlinear transformations of predictors, or interactions between predictors. This ability to directly test the distribution balance on the covariates for the treatment and comparison groups is a benefit of propensity score stratification over multivariable regression models, which assume the addition of observed covariates to the model meets the assumption of strongly ignorable treatment assignment but cannotly directly assess whether adding these variables truly balances the two groups (Newgard et al., 2004).

After strata are formed and the researcher has ensured that the strata have achieved balance, the treatment effect can be estimated. Estimating the treatment effect across each stratum would indicate if there is a Strata x Treatment interaction. For instance, it may be that the treatment is only effective for those most likely to receive the treatment. To find the average treatment effect, the researcher takes the weighted average of the estimated treatment effect across all strata. Rather than estimate the treatment effect across each stratum, the researcher also can use methods like regression or hierarchical linear modeling (HLM) to model the treatment effect, accounting for the propensity strata as a fixed effect.

Assessing the Effects of Hidden Bias

Regardless of how propensity scores are used to estimate treatment effects, if treatment effects are

found, the researcher must conduct sensitivity analyses. Although randomized studies balance both observed and unobserved pretreatment covariates (overt and hidden bias, respectively), one limitation to propensity score analysis, like any analytical adjustment for bias, is that the use of propensity scores only balances the groups with respect to the *observed* pretreatment covariates that were used to construct the propensity score. Although propensity scores remove overt biases from measured covariates, they cannot be expected to remove hidden biases from unobserved covariates, underscoring the need for careful theoretical selection of a great number of pretreatment variables. To determine the possible effect of hidden bias, or whether unmeasured covariates could explain the differences in treatment and comparison outcomes, researchers must conduct sensitivity analyses. Sensitivity analyses allow the researcher to examine the treatment effects for possible departures from SUTVA, indicating if the general conclusion would change with further adjustment for an unmeasured covariate (Lin, Psaty, & Kronmal, 1998; Rosenbaum, 2002). In general, to conduct sensitivity analysis, the researcher assumes that an unmeasured covariate of comparable magnitude to important measured covariates exists and then tests the null hypothesis to see if by adjusting for that confounder, the conclusion regarding the treatment effect would be altered. This would indicate that hidden bias *could* alter the conclusions but does not indicate if an unobserved covariate that explains the differences does exist or to what magnitude. Readers interested in more details on conducting sensitivity analyses should consult Lin, Psaty, and Kronmal (1998), Rosenbaum (1991), and Rosenbaum and Rubin (1983a).

A Propensity Score Stratification Example

So what difference does accounting for observed bias in a quasi-experiment make? To illustrate the use of propensity score stratification, I present an example, comparing findings that do and do not account for observed bias. The reader is cautioned that the analysis presented here is for illustrative purposes and is not the most complete analysis of the data (taking into account non-independence of observations and during-treatment covariates) so inferences should not be made based on the findings. However, the propensity score strata developed *do* allow for comparisons to be made about the type of schools more or less likely to provide the treatment (in this case, gifted programming in reading).

Using data from the restricted-use Early Childhood Longitudinal Study, Kindergarten Cohort (ECLS-K)[2], I examined the effects of a school's policy to have a gifted program in reading on average fifth-grade reading achievement in schools. Such programming might allow teachers to work with more homogeneous groups and target specific needs more effectively, and the culture of a school that chooses to provide gifted programming also may be oriented towards higher standards (i.e., a "rising tide lifts all ships;" Renzulli, 1998). However, some have argued that non-gifted students may be disadvantaged and their achievement may actually decrease when programming is provided to gifted students but not to other students. As Worrell & White (2010) pointed out, some critiques have suggested that "gifted education is responsible for maintaining the achievement gap" (p. 259). Accordingly, the research question here involves whether by providing gifted programming to meet talented readers' needs, schools are harming, benefitting, or having no effect on overall achievement in the school. Thus, I had two groups: schools with a gifted program in reading (about 480 schools; treatment) and those without (about 370 schools; comparison).

Doing a *t*-test to compare the means, the average achievement in schools with a gifted program was 2.58 points lower than in schools without a gifted program (t_{840} = 2.12, p = .03, d = 0.15). This was statistically significant and had a small effect. However, this could be due to pre-existing differences other than provision of a gifted program. That is, schools with a gifted program in reading and the students in those schools could be different from schools without a gifted program in reading, and those differences could explain the achievement difference, rather than the program or lack of program.

The ELCS-K data set is a rich collection of variables. I identified 82 school-level reading pretreatment covariates. These variables were measured prior to fifth grade, which established temporal precedence. The variables were chosen because they were related to the school's gifted reading programming policy (treatment) or to the outcome (reading achievement). They included such variables as average third-grade reading score, region, average

[2] Because the data are restricted-use, all sample sizes and degrees of freedom have been rounded to the nearest 10, as required by the Institute for Education Sciences.

student socio-economic status, sector, average number of achievement groups per third-grade class for reading activities, frequency of various evaluation and instruction techniques, school goals, and state gifted education mandates. Once I estimated the propensity scores using logistic regression, I stratified the schools into quintiles. To check whether the five strata had balanced distribution of the pretreatment covariates, I used regression analyses (i.e., logistic regression for binary covariates; otherwise, linear regression). I regressed the logit and each pretreatment covariate on a binary indicator of gifted program provision, four dummy codes for the strata (with the reference group being the middle stratum), and four treatment-by-policy interactions. When I checked the distribution balance, I found that schools with a gifted program and schools without a gifted program in the same stratum did not statistically significantly differ in any of the pretreatment variables or in propensity to have a program, suggesting that the propensity score strata did, indeed, balance the two groups on the covariates.

Table 1 displays the proportions and means of the full sample of schools and the sample of schools at each stratum for 11 of the 82 pretreatment covariates. Not surprisingly, the schools in the different strata were quite different in their pretreatment characteristics, and using propensity score stratification allowed those differences to be brought to the forefront. Schools that were more likely to have a gifted program were public schools and tended to be in the South and in large or mid-size suburbs and large towns. Teachers in those schools tended to report spending less time on teacher-directed individual

Table 1: *Comparison of Third-grade Demographics (Means/Proportions) for Schools of Varying Propensity to Have a Gifted Program in Reading*

Characteristic	Full Data	Stratum 1[a]	Stratum 2	Stratum 3	Stratum 4	Stratum 5
Proportion in the Midwest	.25	.20	.22	.25	.25	.31
Proportion in the South	.33	.56	.43	.31	.20	.15
Proportion in the West	.24	.21	.28	.29	.19	.23
Proportion in the Northeast	.18	.03	.07	.15	.36	.31
Proportion in a large or mid-size city	.41	.27	.36	.45	.45	.52
Proportion in a small town or rural	.22	.27	.24	.23	.20	.14
Proportion in a large or mid-size suburb or large town	.37	.46	.40	.32	.35	.34
Proportion private schools	.23	.00	.00	.00	.23	.90
Proportion public schools	.67	1.00	1.00	1.00	.77	.10
Mean student SES	3.19	3.20	2.99	2.86	3.08	3.85
Mean reading achievement	118.23	121.44	115.66	113.40	116.04	124.67
Percent free lunch	32.65	28.80	37.73	44.74	36.87	15.00
Percent minority	40.35	30.89	42.02	49.60	44.45	34.68
Average number of achievement groups for reading	2.62	2.98	2.91	2.69	2.68	1.85
Amount of emphasis placed on the goal of openness to new ideas and methods	1.81	1.85	1.85	1.83	1.81	1.72
Amount of emphasis placed on the goal of using curricula aligned with high standards	1.92	1.97	1.94	1.93	1.93	1.81
Amount of emphasis placed on the goal of providing challenging tasks for higher-achieving children	1.70	1.77	1.75	1.68	1.66	1.61

[a] Stratum 1 is most likely to adopt a gifted programming policy for reading while Stratum 5 is most likely to not provide gifted programming in reading

mathematics activities and using a greater number of achievement groups for mathematics and reading. They tended to report placing higher emphasis on using reading curricula aligned with high standards, on providing challenging tasks for higher-achieving students, and on being open to new ideas and methods. The classes in those schools tended to have higher average achievement scores.

Having stratified the schools and achieved balance across the strata, I next identified the schools with a gifted program in reading and those without in each stratum. Then, I compared the reading achievement in each stratum (Table 2). Although the *t*-test indicated statistically significant differences between treatment and comparison groups in reading achievement (not controlling for pretreatment differences), once those differences were taken into account, there were no statistically significant differences in any stratum between schools with different gifted reading program policies. This indicates that the background characteristics, such as region, average student socio-economic status, sector, and state gifted education mandates, resulted in a statistically significant difference in means rather than the gifted education policy (treatment). Using the propensity score strata to account for those 82 pretreatment covariates created comparable groups prior to treatment so that comparison of mean reading achievement in the two groups could be made based on the remaining observed difference: gifted reading program policy. Because no treatment effect was found, conducting a sensitivity analysis to determine if unmeasured covariates could explain the treatment effect was unnecessary.

A benefit of propensity score stratification is that dummy codes for the strata can be used in a model-based approach to account for pretreatment differences rather than including, in this instance, 82 pretreatment covariates. For example, Adelson, McCoach, and Gavin (2012) conducted further analyses of the above data using multilevel modeling with during-treatment covariates in the model as well. Similar to the analyses within each stratum, the differential for a school not having a gifted program was not statistically significant.

Possibilities for Application in Educational Research

The possibilities for applying propensity score analysis in educational research are endless. Researchers have access to many large-scale data sets that would be appropriate for researching important questions in the field. These include the Trends in Mathematics and Science Study (TIMSS), the National Assessment of Educational Progress (NAEP), data sets like the ECLS-K from the Institute of Education Sciences (IES) and the National Center for Education Statistics (NCES), and state and school-district data. Additionally, this method can be used with cross-sectional as well as longitudinal data.

Quasi-experimental comparison groups can be created to study a wide range of questions that are not feasible or ethical to address with random assignment. For example, researchers might be interested in accounting for a child's propensity to participate in a

Table 2: *Within-stratum Average Mean Fifth-grade Reading Achievement between Schools with a Gifted Program in Reading and Those Without*

Stratum[a]	Gifted program		No gifted program			
	N[b]	Mean (SD)	*N*[b]	Mean (SD)	Mean diff	*p*
1	160	141.72 (14.58)	10	143.68 (15.64)	-1.96	.72
2	140	136.49 (18.43)	30	138.19 (16.42)	-1.70	.71
3	120	135.82 (17.01)	50	133.25 (18.25)	2.57	.49
4	50	134.88 (18.99)	120	136.69 (20.74)	-1.81	.61
5	10	151.55 (8.07)	160	145.85 (14.93)	5.70	.40

[a] Stratum 1 has the highest probability to have a gifted program in reading, while Stratum 5 has the highest probability to not have a gifted program in reading.

[b] Sample sizes have been rounded to the nearest 10, as required by IES for analyses using restricted-use data.

Title I program or to take Algebra I in a particular grade, a school's propensity to have a policy allowing grade-skipping or to provide a particular social skills program, or a teacher's propensity to use ability grouping, to recommend students for further evaluation, or to assign homework. Other policy issues that do not lend themselves to experimental techniques but are of interest to educational researchers include the effects of inclusion on students with special needs,

the difference in various outcomes for students in public versus private schools, the effects of disciplinary referrals and suspensions, and how school choice affects student achievement and self-perceptions. Although we cannot assign students randomly to these conditions and cannot assign schools randomly to these policies, we can conduct propensity score analyses to balance students and schools on pretreatment variables and draw causal inferences about the effects of these and other important educational issues.

When considering using propensity score analysis, researchers need to keep in mind that no analytical procedure can make up for poor research design (Rubin, 2007). When designing studies that will use propensity scores to estimate causal effects, researchers must use theory to determine important pretreatment covariates to include in the propensity score model, must use measures that provide reliable scores and allow for valid inferences, and must ensure that groups are balanced with respect to those covariates. Causal inference always is challenging but is especially so in quasi-experiments and with observational data. However, with good design and appropriate application of the method, which includes using a large selection of theoretically-relevant pretreatment covariates, propensity score analysis allows for "relatively more comprehensive control of the pretreatment differences than previously possible" (Hong & Raudenbush, 2005, p. 220) and thus more confidence in causal inferences.

References

Adelson, J. L., McCoach, D. B., & Gavin, M. K. (2012). Examining the effects of gifted programming in mathematics and reading using the ECLS-K. *Gifted Child Quarterly, 56,* 25-39.

Adelson, J. L., McCoach, D. B., & Rogers, H. J. (2009, April). *Estimating treatment effects using stratification on the propensity score: Variable selection and stratification issues.* Paper presented at the 2009 American Educational Research Association Annual Meeting and Exhibition, San Diego, CA.

Allen, J.P. (1989). Social impact of age mixing and age segregation in school: A context-sensitive investigation. *Journal of Educational Psychology*, *81*, 408-416.

Braitman, L. E., & Rosenbaum, P. R. (2002). Rare outcomes, common treatments: Analytic strategies using propensity scores. *Annals of Internal Medicine, 137*, 693-695.

Branum-Martin, L., Foorman, B.R., Francis, D.J. & Mehta, P.D. (2010). Contextual effects of bilingual programs on beginning reading. *Journal of Educational Psychology*, *102*, 341-355. doi: 10.1037/a0019053

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*, 1149-1156.

D'Agostino, J. R. (1998). Tutorial in biostatistics: Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine, 17*, 2265-2281.

de Vise, D. (2008, December 16). Montgomery erasing gifted label: Implications concern some school parents. *Washington Post,* p. B01. Retrieved December 18, 2008, from http://www.washingtonpost.com/wp-dyn/content/article/2008/12/15/AR2008121503114.html?hpid_seceducation

Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association, 94,* 1053–1062.

Delcourt, M. A. B., Cornell, D. G., & Goldberg, M. D. (2007). Cognitive and affective learning outcomes of gifted elementary school students. *Gifted Child Quarterly, 51*, 359-381.

Grant, B. (2002). Justifying gifted education: A critique of needs claims and a proposal. *Journal for the Education of the Gifted*, *25*, 359–374.

Graesser, A.C. (2009). Inaugural editorial for Journal of Educational Psychology. *Journal of Educational Psychology*, *101*, 259-261. doi: 10.1037/a0014883

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

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*, 945-960.

Hong, G., & Raudenbush, S. W. (2005). Effects of kindergarten retention policy on children's cognitive growth in reading and mathematics. *Educational Evaluation and Policy Analysis, 27*, 205-224.

Lin, D. Y., Psaty, B. M., & Kronmal, R. A. (1998). Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics, 54*, 948-963.

Marsh, H. W., Hau, K.-T., & Craven, R. (2004). The big-fish-little-pond effect stands up to scrutiny. *American Psychologist, 59*, 269- 271.

Newgard, C. D., Hedges, J. R., Arthur, M., & Mullins, R. J. (2004). Advanced statistics: The propensity score - A method for estimating treatment effect in observational research. *Academic Emergency Medicine, 11*, 953-961.

Renzulli, J. S. (1998). A rising tide lifts all ships: Developing the gifts and talents of all students. *Phi Delta Kappan, 80*, 104-111.

Rogers, K. B. (2007). Lessons learned about educating the gifted and talented: A synthesis of the research on educational practice. *Gifted Child Quarterly, 51,* 382-396.

Rosenbaum, P. R. (1991). Discussing hidden bias in observational studies. *Annals of Internal Medicine, 115*, 901-905.

Rosenbaum, P. R. (2002). *Observational studies*. New York: Springer-Verlag.

Rosenbaum, P. R., & Rubin, D. B. (1983a). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society, Series B, 11*, 212-218.

Rosenbaum, P. R., & Rubin, D. B. (1983b). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 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*, 516-524.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics, 6*, 34-58.

Rubin, D. B. (1986). Comment: Which ifs have causal answers. *Journal of the American Statistical Association, 81*, 961-962.

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine, 127*, 757-763.

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine, 26*, 20-36.

Rudner, L. M., & Peyton, J. (2006). Consider propensity scores to compare treatments. *Practical Assessment, Research & Evaluation, 11*(9), 1-9.

Sapon-Shevin, M. (1993). Gifted education and the protection of privilege: Breaking the silence, opening the discourse. In L.Weis & M.Fine (Eds.), *Beyond silenced voices: Class, race and gender in the United States* (pp. 25–44). Albany, NY: SUNY Press.

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

Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H., & Shavelson, R. J. (2007). *Estimating causal effects: Using experimental and observational designs*. Wasington, D. C.: American Educational Research Association.

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,* 1334-1344.

Shadish, W. R., Luellen, J. K., & Clark, M. H. (2006). Propensity scores and quasi-experiments: A testimony to the practical side of Lee Sechrest. In R. R. Bootzin, & P. E. McKnight, *Strengthening research methodology: Psychological measurement and evaluation* (pp. 143–157). Washington, DC: American Psychological Association.

Stone, C. A., & Tang, Y. (2013). Comparing propensity score methods in balancing covariates and recovering impact in small sample educational program evaluations. *Practical Assessment, Research & Evaluation, 18*(13), pp. 1-12. Available online: http://pareonline.net/getvn.asp?v=18&n=13

Stroessner, S.J., Beckerman, L.S. & Whittaker, A. (2009). All the world's a stage? Consequences of a role-playing pedagogy on psychological factors and writing and rhetorical skill in college undergraduates. *Journal of Educational Psychology*, *101*, 605-620. doi: 10.1037/a0015055

Worrell, F., & White, L. (2009). Review of 'Critical issues and practices in gifted education: What the research says'. *Psychology of Aesthetics, Creativity, and the Arts*, *3*, 259-261. doi:10.1037/a0015410.

Author Note:

The author would like to thank Jesse Owen and Jason Osborne at the University of Louisville for feedback on this manuscript and D. Betsy McCoach at the University of Connecticut for feedback on a previous version on which it was based. A previous version of this paper was presented at the 2009 American Educational Research Association Annual Meeting and Exhibition in San Diego, CA. Data used in the example were from the author's dissertation research at the University of Connecticut, which was supported by a Fellowship from the American Psychological Foundation.

Citation:

Adelson, Jill L. (2013). Educational Research with Real-World Data: Reducing Selection Bias with Propensity Score Analysis. *Practical Assessment, Research & Evaluation*, 18(15). Available online: http://pareonline.net/getvn.asp?v=18&n=15

Corresponding Author:

Jill L. Adelson, Ph.D.
Educational & Counseling Psychology, Counseling, and College Student Personnel Department
College of Education and Human Development
University of Louisville
Louisville, KY 40292

Email: jill.adelson [at] louisville.edu