A Tutorial and Case Study in Propensity Score Analysis:
An Application to Education Research

A Paper Presentation for the 2016 AERA Annual Meeting, Washington D.C.

April 10, 2016

Rebecca S. Putman

Tarleton State University

College of Education

Assistant Professor

putman@tarleton.edu
Abstract

Randomized control trials are considered the gold standard for conducting research and estimating causal effects; however, educational research rarely lends itself to experimental design and true randomization. In recent years, there has been a growing interest in finding new approaches to estimate causal effects in nonrandomized studies in education. Propensity score matching allows researchers to estimate causal effects when randomized studies are not possible. In this paper, I present a primer on propensity score matching through a practical example of literacy research. In addition, I provide a broad theoretical framework and a step-by-step outline for using propensity scores in observational literacy research in the context of my own research.
A Tutorial and Case Study in Propensity Score Analysis: An Application to Education Research

Objective

Randomized control trials are considered the gold standard for conducting research and estimating causal effects (Shadish, Cook, & Campbell, 2002); however, educational research rarely lends itself to experimental design and true randomization. In recent years, there has been a growing interest in finding new approaches to estimate causal effects in nonrandomized studies in education (Austin, 2011). Many of these new approaches have suggested that observed bias can be removed from estimated treatment effects by incorporating covariates, based on sound theory and previous research, into the statistical models (Murnane & Willett, 2011). In theory, these covariates would help researchers account for differences in baseline characteristics between treated and untreated participants in a study, allowing them to estimate the true effects of treatment on the outcomes (Austin, 2011).

One of these new approaches that allows for the estimation of causal effects is propensity score matching. Propensity score matching is now viewed as one of the most robust methods for estimating causal inferences between treatments and outcomes when randomized studies are not possible. (Shadish, Cook, & Campbell, 2002). Despite the calls for more scientifically based methodologies in education, propensity score matching remains significantly underutilized in the educational literature (Lane & Henson, 2010).

In this paper, I present a primer on propensity score matching through a practical example of literacy research. For illustrative purposes, I report on how I used propensity score matching to estimate the causal effects of Istation on the early literacy achievement of children in twelve kindergarten classrooms. (Istation is an integrated learning system used by
approximately 4,000,000 students in the United States.) I also provide a broad theoretical framework and a step-by-step outline for using propensity scores in observational literacy research in the context of my own research on Istation.

**Propensity Score Matching**

*Theoretical framework*

Propensity score matching was first introduced as a methodology in the seminal work of Rosenbaum and Rubin (1983). Their goal was to present a mathematical solution that would account for group differences and address the limitations of observational research. Propensity score matching is widely used in the fields of medical and statistical research to estimate causal effects when a randomized control trial is not possible; however, the application of propensity scores to educational research is fairly new.

Rosenbaum and Rubin (1983) defined propensity scores as the conditional probability of treatment assignment based on certain observed baseline covariates. More simply, the propensity score is the predicted probability of treatment as predicted by relevant matching variables (Reutzel, Spichtig, & Petscher, 2012). One common analogy for propensity scores is flipping a coin (Lane & Henson, 2010). The probability of getting heads or tails is 50% and would result in a propensity score of \( p = 0.50 \). Similarly, participants in a randomized study have an equal chance of being in the control or treatment group, and their propensity score would be \( p = 0.50 \).

In a non-randomized study, this probability must be estimated using a set of relevant covariates. A propensity score \((e)\) for each individual \((i)\) is expressed as the conditional probability \((P)\) of assignment to a particular treatment or control group \((T)\) given a set of covariates \((X)\) (Rosenbaum & Rubin, 1983). The formula is \( P(T = 1|X) = P(e) \).
The goal or objective for a researcher using propensity scores is to select a sequence of variables that are considered important in matching participants (Reutzel, Spichtig, & Petscher, 2012). If the theory and history on which the researcher bases his/her selection of covariates is good, then the model is sound and causal inferences can be made (Reutzel, Spichtig, & Petscher, 2012; Thoemmes & Kim, 2011). As Thoemmes & Kim (2011) point out, “Under the assumption that all relevant covariates have been assessed, a propensity score analysis can yield unbiased causal effect estimates” (p. 92). Accounting for these differences in characteristics between the two groups in an observational study would mimic the random selection of participants in a randomized control trial and would recreate a population that would have been expected in a randomized experiment (Thoemmes & Kim, 2011). Once calculated, propensity scores can then be used to balance control and treatment groups. Furthermore, once cases are matched, the effects will be more reflective of the true treatment effect and similar to the interpretation of randomized designs.

**Method**

While the full range of options and practices in propensity score matching is beyond the scope of this paper, this section will provide a simplified step-by-step outline of the steps involved in propensity score matching. (See Randolph et al., 2014, for a more detailed guide to propensity score matching in R, or Thoemmes, F.J., 2012, for a more detailed guide to propensity score matching in SPSS.)

**Step 1—Select Covariates:**

The first step of propensity score matching is to select a set of relevant covariates for your study that theory and prior research indicate are important or that influence the outcome
measure(s). The goal is that no confounders are omitted from the set of covariates. There is no limit to the number of covariates that can be used.

**Step 2—Estimate Propensity Scores:**

Using logistic regression, the second step is to estimate the propensity score for each participant based on the included covariates and on whether the participant was in the control or treatment group. These scores can be estimated using most commercially available statistical software, including SPSS, SAS, STATA, and R.

**Step 3—Match Participants:**

Next, match the participants of your study based on their estimated propensity scores. There are several matching options, but the goal is to find the matching option that results in the lowest mean differences among the covariates between groups. The two most common matching approaches are nearest neighbor matching and optimal matching. Nearest neighbor, or greedy matching, uses the propensity scores to pair a control case with the treatment case whose propensity score is closest. Optimal matching calculates the distances between all of the treatments and controls and minimizes the distances between all possible matched sets (Beal & Kupzyk, 2014). Currently, no commercially available statistical programs are able to match participants through point and click methods; however, there are several macros and syntax programs that can be used to match control and treatment cases.

**Step 4—Estimate Treatment Effect:**

Using any statistical model or test, estimate the treatment effect in the matched subsamples.
**Objective**

Guided by Vygotsky’s social learning theory, the illustrative study reports a 24-week investigation on whether regular use of Istation® had an effect on the early literacy achievement of children in twelve kindergarten classrooms. A mixed-method, quasi-experimental design was constructed using propensity scores. Because randomly assigning children to use or not use Istation® was not possible, the propensity scores allowed me to
construct matched control and treatment groups in order to control potential variation (beyond the instructional format presented) at the participant level. The goal of the illustrative study was to determine if teachers or technology were more effective in instructing students in early literacy skills and what role technology should play in an early literacy classroom.

**Data Sources**

One hundred fifty students and 12 teachers returned the consent forms for the Istation study. The final analysis included 72 students matched through propensity score matching, with 36 students in each of the treatment and control groups. Data on participants’ literacy achievement were collected from two sources:

1. Developmental Reading Assessment 2 [DRA2]: The twelve participating teachers provided students’ middle of the year DRA2 scores to me. This teacher-administered individual assessment was given to all participants in January 2014. This measure was used to determine participants’ independent reading levels.

2. Clay’s Observation Survey: Two trained research assistants and I individually administered five subtests of the Observation Survey to the 150 students who returned the consent forms. Subtests include hearing and recording sounds, writing vocabulary, letter sound knowledge, concepts about print, and word reading. Each testing session averaged approximately 30 minutes.

**Illustrative Study Method**

Step 1—Select Covariates:

For the illustrative study, I used both theory and prior empirical research to identify seven predictor variables that influence young children’s early literacy skills. Participants in the study were matched on the following variables: (a) age on the first day of kindergarten, (b) gender, (c)
ethnicity, (d) free and reduced lunch status, (e) English language learner status, (f) beginning of year letter identification score, and (g) level of literacy support provided by the teacher (low, medium, high).

Step 2—Estimate Propensity Scores:

Using the covariates identified in step one as predictors of the dichotomous treatment variable, I used logistic regression to estimate propensity scores for each of the 150 participants in my study. While most commercially available statistical software will conduct logistic regression, I used a syntax in R to estimate propensity scores for each of the participants in the control and treatment groups.

Step 3—Match Participants:

Once propensity scores were estimated using logistic regression, treated and untreated participants were matched using a nearest neighbor with caliper matching algorithm (Austin, 2011). The caliper width used was equal to 0.2 of the standard deviation of the logit of the propensity score (Austin, 2011, 2014). This narrow caliper width reduced my matched sample size; however, it lead to more variance in the treatment effect (Austin, 2011). Research has confirmed that caliper matching leads to improved balance on baseline covariates and less bias in treatment effect estimates (Austin, 2014). When participants who used Istation® were matched with participants who did not use Istation® based on the propensity score algorithm, 36 matched pairs were formed, for a total sample of 72 participants. For this step, I used a matching syntax in R.

Step 4—Estimate Treatment Effect:

For my analysis, I used a descriptive discriminant analysis [DDA] (Huberty, 1994) to evaluate the effect of Istation® on the early literacy skills of kindergarteners and to determine which
variables contributed to any differences between the two groups, but any statistical analysis can be used.

**Findings**

Propensity score matching allowed for causal estimates of the effectiveness of Istation including:

1. Istation® did have a statistically significant effect on the early literacy skills of the kindergarten students studied and was able to explain almost 18% of the variance in group differences.
2. Differences in Hearing/Recording Sounds and Letter Sound Knowledge were the two main contributors to the variability between the two groups. Variability in Writing Vocabulary contributed minimally to the group differences while Concepts About Print was a suppressor variable in the model.

**Limitations of Propensity Scores**

The overarching assumption when estimating propensity scores is unconfoundedness (Murnane & Willett, 2011). That is, the researcher assumes that all variables that influence and affect treatment assignment have been accounted for in the statistical model (Austin, 2011). It should be noted that the assumption of unconfoundedness cannot be empirically tested; instead, researchers must attempt to provide theoretical and empirical evidence that all relevant covariates have been included in the model (Thoemmes & Kim, 2011). If researchers fail to include an important confounder, the propensity scores will lead to biased results (Beal & Kupzyk, 2014). Another limitation of propensity score matching is that it often produces smaller sample sizes than initially obtained in the data collection process.
**Significance**

This paper provides a practical illustration of how to use propensity scores in educational research. Propensity scores are one way in which educational researchers can infer causality when randomized control trials are not possible. In addition, propensity score matching is a method recommended by the U.S. Department of Education to improve the strength of quasi-experimental studies.

While the illustrative study does not provide an exhaustive survey on the use of propensity scores in education, it does provide readers with the basic information and context needed to understand the potential usefulness of propensity score matching or to evaluate other studies that incorporate propensity score matching.
References


