JSO JOURNAL of STEM OUTREACH

Vol. 8, Issue 1, April 2025

Testing the Efficacy of Educational Interventions on Matched Student Samples: A Primer for Propensity Score Matching in R

Nicholas D. Evans[1], Perla C. Perez[2], and Osvaldo F. Morera[2]

[1]Department of Psychology, University of Manitoba, Winnipeg, Manitoba, CANADA and [2]Department of Psychology, University of Texas at El Paso, USA

Keywords: Propensity Score Matching, R Programming, Education Interventions, Non-experimental Design, Evaluation

Publication Date: April 14, 2025

DOI: https://doi.org/10.15695/jstem/v8i1.05

ABSTRACT: In many educational intervention programs, it is not possible to randomly assign students to an experimental and control condition. For example, in our research we wanted to compare students who were enrolled in a biomedical pathway program to students who were not in such a program. However, students select their academic pathway program and a randomized controlled trial cannot be conducted. Propensity score matching (PSM) is a valuable statistical technique in areas of research when randomized control trials are not always possible. It can be widely used to mimic the process of randomization by creating comparable groups based on key covariates while increasing causal inference and reducing bias. The aim of this article is to provide guidance for science education researchers to make informed decisions about the selection of matching methods and implementation of PSM using the MatchIt package (Ho et al., 2011) in R. In this article, we 1) discuss the utility of using PSM for research involving educational interventions, 2) provide a comprehensive guide for conducting PSM with educational data and provide a detailed step-by-step guide on conducting PSM for nearest neighbor matching using R, and 3) apply it to a National Institutes of Health (NIH)-funded high school education program.

INTRODUCTION

When conducting research, achieving a randomized experimental design is often ideal, as it allows for robust causal inference while reducing bias. However, in real-world settings, random assignment is not always possible due to practical or ethical constraints. This is particularly the case in educational research, in which researchers and program evaluators implement educational programs and interventions. In such cases, researchers may turn to quasi-experimental designs. For example, when testing a new educational program, the students and their parents are given the freedom to decide for themselves whether to enroll in the program. The goal for researchers, then, is to compare the academic success of the students who enroll in the program to the students who do not. Although such designs offer flexibility, other challenges are introduced when establishing causal relationships. To address this, several alternatives exist.

One option is to add additional variables or comparison groups. However, this may not always be feasible if the data have already been collected. Another approach is to include covariates in Analysis of Covariance (ANCOVA) or Multiple Linear Regression (MLR) models, but this is only suitable for a small number of covariates. One statistical approach that addresses these limitations, especially when testing causal effects of an educational intervention, is the use of propensity score matching (PSM). PSM assumes that the only differences between participants in each group is their group membership and no other potentially confounding variable.

The overarching goal, therefore, is to demonstrate the utility of PSM for education research, present a primer for conducing PSM analyses in R, and demonstrate how this can be applied to a National Institutes of Health (NIH)-funded high school education program: Project ACtion for Equity (ACE). Specifically, we will assess whether students exposed to Project ACE differ from a comparable group of students on cumulative GPA, as cumulative GPA is used as a criterion for many scholarship programs.

What is Propensity Score Matching? Matthay et al. (2020) describe two types of techniques that researchers can use to establish causal inference when randomization is not possible. The first type of approach, instrumental based approaches, indicate that a third variable (an instrument) is exogeneous and is related to the treatment variable. The instrumental variable does not directly affect the outcome variable, but rather, affects the treatment variable, which in turn, affects the outcome variable. Statistical approaches that Matthay et al. (2020) classify as instrumental approaches include difference in differences designs and regression discontinuity designs. Interested readers are referred to Matthay et al. (2020) for a comparison of the two approaches.

For the purposes of this paper, we turn to the second type of approach that Matthay et al. (2020) highlight as a viable means to establish causal inference for non-randomized designs. These approaches, confounder-control techniques, statistically control for the effects of covariates (Matthay et al., 2020). In confounder-control approaches, a researcher must identify a "sufficient set" (Matthay et al., 2020) of control variables. For example, in determining whether selection to a formalized undergraduate training program causes improved academic outcomes, there may be a host of other variables that are related to selection like Socioeconomic Status (SES), participant gender, and high school grade point average.

Confounder-control approaches assume that the researcher has identified and measured all such control variables or has proxy variables for such variables. One key technique that falls under this umbrella of the confounder-control approach is PSM. By estimating the probability for each participant of being in one condition or another condition (e.g., a "propensity"), conditional on how each participant scores on the matching variables, PSM is an invaluable and often underutilized technique to account for the biases of non-randomized designs. PSM is a statistical technique that is often used in quasi-experimental designs where random assignment is not always possible and has gained popularity among education, social sciences, health, and other fields (Fan & Nowell, 2011).

At this stage, a researcher may be seeking guidance for determining variables to include for matching purposes. Clearly, the matching variables should not occur after the dependent variable has occurred as they may be influenced by the dependent variable. Matching variables may also be either continuous or categorical in nature, so there is no restriction on the type of variable to be included for matching purposes. Examples of matching variables may include participant sex, participant race or various indices of socioeconomic status. Finally, it is worth emphasizing that the researcher must determine which theoretically-relevant variables should be included for matching purposes and the researcher should explicitly state which variables were included when matching potential control participants to participants in an experimental condition.

As with all statistical techniques, there are assumptions that underlie such approaches. The primary assumption that underlies confounder-control approaches is called exchangeability (which is also known as ignorability or lack of confounding). In the example concerning Project ACE, that we have introduced and will discuss further, a violation of exchangeability would take place if individuals who would have done well academically without enrollment in a formalized training program were more likely to be selected into the training program. This is why it is important to identify a sufficient set of control variables in propensity score matching, as group differences between selected and matched control participants would lead to biased parameter estimates of group differences and violate the internal validity of the study without such matching on a sufficient set of control variables (Pearl, 2000; Shadish et al., 2002).

While exchangeability is an assumption of propensity score matching, the researcher still needs to assess assumptions underlying any statistical technique that they would use after control group participants have been matched to the experimental group participants. For example, if a researcher were to perform analyses involving general linear models (e.g., t-tests, ANOVAs, multiple linear regression), the researcher would still need to assess the assumptions of normality and homogeneity of variance (see for review Keppel & Wickens, 2004 and Cohen et al., 2003).

How is Propensity Score Matching Useful for Education Research? The benefit of employing a propensity score matching approach is at least three-fold. Since many NIH training programs do not allow for randomized control trials, propensity score matching allows for a way to compare individuals exposed to an intervention to a comparable sample of individuals who are similar to the intervention group on key demographic variables. Secondly, propensity score matching allows for reduced burden on partnering organizations. For example, the researcher would make a future data request from the participating high schools for only the data from students in the experimental condition and a smaller number of students from the general population of students. A principal at a school would then make a request to the

district office for academic data. Because the researcher would only make a request for the data from a fraction of the school's students, the work of the individuals at the district office would be substantially reduced. As a result, they may be more apt to fulfill the request, and the request may be completed in a more time efficient manner.

A third benefit of propensity score matching is that meaningful assessments between intervention students and control students can be made. Once the matching process is completed, researchers can explore a wide range of analyses to examine the impact of the treatment on outcomes of interest. Examples of statistical techniques that can be applied include independent sample t-tests to determine group differences on a continuous dependent measure. Nonparametric procedures can also be used to examine group differences. Chi-square tests of independence can be used to examine associations between group membership and categorical outcome measures.

Finally, other points to be considered include sample size considerations and distributional assumptions. In our example, we only have 29 students in an experimental condition and we are using a one-to-one matching procedure. We will be comparing 29 students enrolled in a high school educational pathway program designed to promote biomedical careers to 29 students who are not enrolled in such an educational pathway program. The power of any statistical test is the ability to detect an effect if that effect is there and will be influenced by sample size. Moreover, should the outcome variable be skewed or peaked, the researcher may want to transform the outcome variable (see Cohen et al., 2003) or use nonparametric approaches to ascertain group differences.

How is Propensity Score Matching Carried Out? Using logistic regression, propensity scores are estimated using the probability of an individual being assigned to a treatment condition based on key covariates that may influence equivalence between groups (Benedetto et al., 2011; Olmos & Govindasamy, 2015). The treatment observations are then matched to observations within the control group based on their probability of being in the treatment group and size of the treatment group (i.e., Kane et al., 2021). Once matched, unused control observations are removed, leaving the treatment and new control observations (see Figure 1 for a visual depiction). The treatment and control groups in this new matched sample can then adequately be compared when testing researchers' outcome variables of interest. Ultimately, by using PSM, researchers can enhance their causal inference validity of their studies in which they utilize non-experimental designs and observational studies (Fan & Nowell, 2011).

We aim to provide a comprehensive guide on conducting PSM in educational research. In our demonstration, we will present a step-by-step process of PSM using the MatchIt package in R with nearest neighbor matching (see Figure 2 for an overview of the process). While various statistical programs such as SPSS or Stata can be used to perform PSM, we have chosen R programming due to its flexibility, availability of specialized packages like MatchIt (Ho et al., 2011), and its widespread use in the research community. We will outline the necessary steps, importing & preparing data, installing the MatchIt (Ho et al., 2011) package on R, estimating propensity scores, and computing propensity scores on matched scores. The focus will be on the nearest neighbor matching technique, as it is commonly used and an intuitive approach in PSM. However, please note that the choice of matching method should be guided by specific research context, distribution of propensity scores, sample size, and the goals of the study. See the supplementary material for a brief description of the different methods to match individuals in the R program. Additional comprehensive information about the several matching methods that can be implemented using the MatchIt package can be found in the Comprehensive R Archive Network (https://cran.r-project.org/web/packages/MatchIt/vignettes/matching-methods.html).

It is also important to note that the focus of this paper will be propensity score matching as it relates to matching one treatment condition to a control. In cases of multiple treatment conditions, researchers can employ a generalized pro-

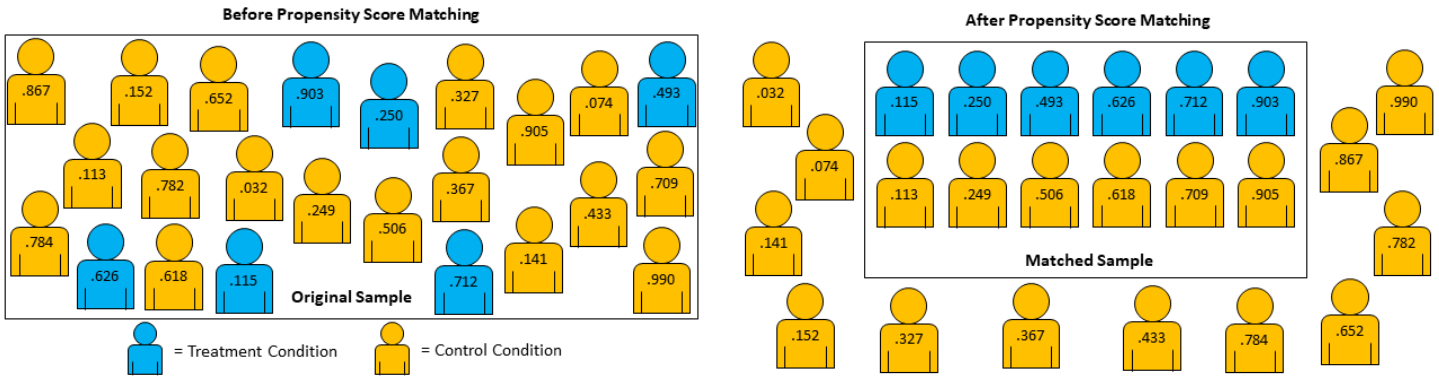

Figure 1. Propensity Score Matching Visual Depiction. Note: Values for each participant indicate their corresponding propensity scores.

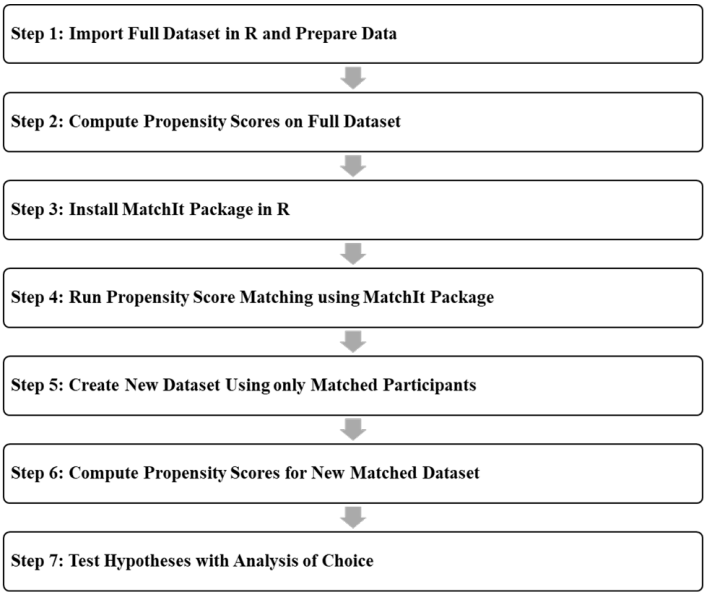

Figure 2. Propensity Score Matching Process Flow Chart.

pensity score approach (see Yoshida et al., 2017 and Zhao et al., 2021 for a description and tutorial).

METHODS

Demonstration: Project ACE High School Data. Data from a high school that participated in the Project ACE program was used to illustrate this technique in R Studio. Project ACE aims to expand the biomedical science pipeline by collaborating with teachers and students from three underrepresented and underprivileged high schools in West Texas and Southern New Mexico. To investigate if Project ACE membership predicted higher GPA in Project ACE students compared to non-Project ACE students, PSM was used. Project ACE students were matched with equivalent high school students based on (a) grade classification, (b) gender, (c) ethnicity, which was coded as Latinx or non-Latinx, (d) race, (e) English language learner (ELL), which was coded as ELL or non-ELL, (f) special education, which was coded as being in special education or not, (g) homelessness, which was coded as being homeless or not, and (h) free lunch status, coded as being eligible for free lunch or not. We selected these covariates in our study, as the literature suggests that demographic variables such as gender, ethnicity, and eligibility for free or reduced price lunch are associated with Latinx academic achievement (Taggart, 2018). Note that the selected covariates will differ based on the research question that is guiding each researcher. For this example, we focused on ninth graders at one high school, resulting in a sample of 29 ACE students and 325 non-ACE students ($N = 354$).

Step 1: Import and Prepare Data. R is a programming language developed by John Chambers and colleagues that can be used for statistical analysis and data visualization as it offers a wide variety of extension packages (R Core Team, 2023). The free software is available to be downloaded in various operating systems such as Windows, MacOS, and Linux (https://www.r-project.org/). Additionally, for ease of use, we recommend using the R Studio environment (https://posit.co/products/open-source/rstudio/) to run all of your analyses. After installing and downloading R and R Studio, you first must set your working directory and import the dataset. There are several ways to do so, and some methods are easier on certain systems than others (e.g., Windows vs. MacOS). For our demonstration, we first used the *setwd()* function to set the working directory (i.e., the file folder from which we would be importing our dataset). Once we set our working directory using our file path, we then imported our dataset using the *read.csv()* function, since the file we imported is a .csv file. Within this function, we first specified the name of the file in quotations and specified that we wanted to retain the headers from our data (see lines 1 and 3 of our available code on OSF).

Once the dataset has been imported, you may need to clean the data (e.g., indicator code categorical variables, rename variables, etc.) depending on your needs. For example, our dataset contained categorical variables with more than two levels, so we applied indicator coding accordingly. However, analyses are not constrained to binary variables as both categorical with more than two levels (e.g., racial or ethnic identity) and continuous variables (e.g., age) can be used doing this method. Additionally, based on the type of analyses that the researcher uses to best answer their research questions, data can be formatted accordingly to their needs (e.g., transforming data from long to wide format). For a more detailed guide on how to use R, you can explore various resources available at RStudio Education (https://education.rstudio.com/learn/beginner/) and the "R for Beginners" by Emmanuel Paradis (Paradis, 2005). Both resources offer a comprehensive tutorial for beginners with materials to help readers understand and be proficient in R.

Step 2: Compute Propensity Scores for Original Dataset. Once the dataset is cleaned and all key matching variables are prepared, we then need to compute propensity scores for all participants based on our matching variables of interest. To do so, we conduct a logistic regression model using the *glm* function, in which we regress the ACE membership binary variable on each of the matching variables of interest (see example below).

```
ps <- glm(ACE_Membership ~ Gender_dc + Ethnicity_dc + Black_dc + NativeAmer_dc + ELL_dc + SPED_dc + Homeless_dc + Migrant_dc + FreeLunch_dc, data = Dataset, family = "binomial)
```

Table 1. *Variables used in the Logistic Regression Model for PSM.*

Variable Name	Variable Symbol	Coding Scheme
ACE Membership	ACE_Membership	1 = ACE, 0 = non-ACE
Gender	Gender_dc	1 = Male, 0 = Female
Race/Ethnicity: Hispanic/Latinx	Ethnicity_dc	1 = Hispanic, 0 = not Hispanic
Race/Ethnicity: Black	Black_dc	1 = Black, 0 = not Black
Race/Ethnicity: Native American	NativeAmer_dc	1 = Native American, 0 = not Native American
English Language Learner Status	ELL_dc	1 = Yes, 0 = No
Special Education Status	SPED_dc	1 = Yes, 0 = No
Homelessness Status	Homeless_dc	1 = Yes, 0 = No
Migrant Status	Migrant_dc	1 = Yes, 0 = No
Free and Reduced Lunch Status	FreeLunch_dc	1 = Yes, 0 = No

In the command line above, we are regressing ACE membership on a number of explanatory categorical variables. These variables include participant gender, participant race/ethnicity (three indicator-coded variables with White being the referent: Hispanic/Latinx, Black, and Native American), English language learner status, special education status, homelessness status, migrant status, and free and reduced lunch status (see Table 1 for a detailed list of the variables and their corresponding coding schemes). Once we have assessed this model, we then need to extract our propensity scores for each participant and input them in the original dataset:

Dataset$psvalue <- predict(ps, type = "response")

The command above computes the propensity scores and inputs a new variable of propensity scores (denoted psvalue) in the original dataset. Doing this will allow us to understand the distribution of propensity scores between ACE and non-ACE students. As shown in Figure 3, the distribution of propensity scores between ACE and non-ACE students reflects an unequal distribution of propensity scores between the treatment and control condition, indicating that using propensity score matching is necessary. While this step demonstrates that the propensity scores differ across groups, we need to identify and match a propensity score from a participant in the experimental condition with a propensity score in the control condition, requiring the use of an R package like MatchIt.

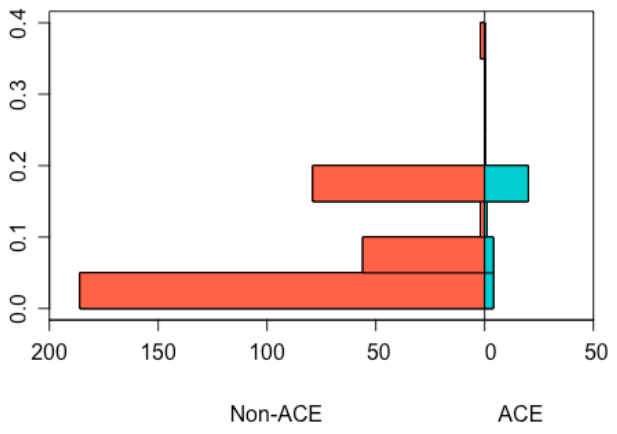

Figure 3. Distribution of Propensity Scores for ACE and non-ACE Students Before Matching.

Step 3: Install MatchIt Package in R. Estimating the propensity scores using the generalized linear model in Step 2 allows us to compute the propensity scores for each student. Using these scores, we now need to match students in the treatment condition with students in the control condition, such that we have a subset dataset with an equal number of students in the control condition who are similar to students in the treatment condition with regard to the key variables we considered when calculating propensity scores.

One of the most commonly used packages in R to match treatment and control participants based on propensity scores is the "MatchIt" package (Ho et al., 2011). This package implements a wide range of matching procedures including propensity score matching. To install the "MatchIt" package into your library, use the code:

install.packages("MatchIt")

In order to load the package from your library, use the code:

library(MatchIt)

Step 4: Run Propensity Score Matching using MatchIt Package. Several matching methods can be used to estimate propensity scores including exact, subclassification, and greedy/nearest neighbor (see supplementary material). The matching procedure selected will vary based on the researcher's question. Since greedy/nearest neighbor matching is the most frequently implemented matching procedure (Thoemmes and Kim, 2011; Zakrison et al., 2018), we will be using this method for our demonstration.

In greedy matching, a participant in the treatment condition is randomly selected and matched to a participant in the control condition who has a propensity score close to

the randomly selected participant in the experimental condition. Once the participant in the experimental condition has been matched, another participant in the treatment condition is randomly selected and is matched to an unselected participant in the control condition (i.e., matching without replacement; however, matching with replacement is possible, see Austin, 2011). This process proceeds until all randomly selected participants in the experimental condition are matched to the best remaining participants in the control condition, resulting in one participant from the experimental condition being uniquely matched to one participant in the control condition. While there are several matching methods (see Baser, 2006), optimal matching can be contrasted with greedy matching, where matches between participants' propensity scores in the experimental and control condition are selected to minimize the overall difference in propensity across matched pairs. On the surface it may seem that optimal matching is the ideal matching procedure. However, Gu and Rosenbaum (1993) found that optimal matching does not outperform greedy matching.

To perform the matching, we will use the *matchit* function, which will take the same logistic regression model we used to compute the propensity scores for the original dataset and use it to match one ACE student with one non-ACE student based on how close their propensity scores are to each other (see below). In the syntax below, the function specification to select nearest-neighbor or greedy matching is *method = "nearest."*

```
matched <- matchit(ACE_Membership ~ Gender_
dc + Ethnicity_dc + Black_dc + NativeAmer_dc +
ELL_dc + SPED_dc + Homeless_dc + Migrant_
dc + FreeLunch_dc, data = Dataset, method =
"nearest", ratio = 1)
```

Step 5: Create New Dataset of Only Matched (1-to-1) Students. Using the results of the PSM analyses, we can then create a subset of the original dataset that only contains the matched students (e.g., the 29 ACE students and 29 matched control counterparts). This dataset can then be used to test the researchers' hypotheses using the analyses most appropriate to address their research questions. Below, we provide the code to create the new dataset, "matched.data".

```
matched.data <- match.data(matched)
```

Step 6: Compute Propensity Scores for Matched (1-to-1) Dataset. To test whether the matching procedure resulted in a dataset of equally matched students, we will then compute propensity scores for this new dataset using the same *glm* function and logistic regression model (see below):

```
MatchedModel <- glm(ACE_Membership
~ Gender_dc + Ethnicity_dc + Black_dc +
NativeAmer_dc + ELL_dc + SPED_dc +
Homeless_dc + Migrant_dc + FreeLunch_dc, data
= matched.data, family = "binomial")
```

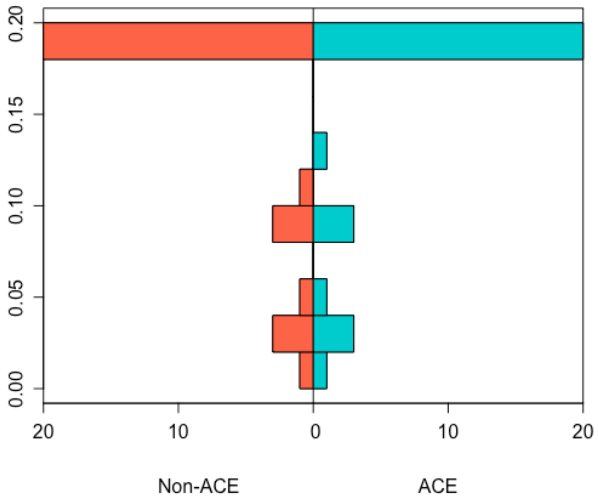

Figure 4. Distribution of Propensity Scores for ACE and non-ACE Students After Matching.

As shown in Figure 4, the distribution of propensity scores is even when comparing ACE and non-ACE students, suggesting that we now have an equally matched sample of students. In other words, when we examine the distribution of propensity scores in Figure 4 across the two conditions, we see that the distribution of propensity scores almost mirrors one another, indicating that we have identified 29 participants in the control condition that mirror the 29 participants in the treatment condition.

Step 7: Test Hypothesis with Analysis of Choice. Using the new matched dataset, the researcher(s) can now carry out the analyses that best address their research questions. For this demonstration, we will be using the 29 ACE students and 29 matched control counterparts from the example above. Students from the ACE cohort were primarily female (86.2%) and Hispanic (100%). The majority of ACE students were not identified as ELL (79.3%), in special education (93.1%), as migrant (100%), or homeless (96.6%). Similarly in terms of demographics, matched control students were primarily female (86.2%), Hispanic (100%), and white (100%). We also see the same pattern as in the ACE students where controlled matched students were not identified as ELL (82.8%), in special education (93.1%), migrant (100%), or homeless (100%). See Table 2 for demographic information on ACE and matched control students.

Using this data, we sought to examine whether ACE and matched non-ACE students differed in terms of cumulative GPA. To test this hypothesis, an independent sample t-test was conducted to assess if current GPA (i.e., *Current*.GPA) means differ between ACE and matched non-ACE students (i.e., *Tracking.Pathway*). Using the function "t.test", we ran the following code on R:

```
t.test (Current.GPA ~ Tracking.Pathway, var.
equal=TRUE, data = matched.data, na.rm=TRUE)
```

Table 2. *Demographics of ACE vs. Matched Control.*

Demographics	ACE (n = 29)		Matched Control (n = 29)	
Gender	Female	25 (86.2%)	Female	25 (86.2%)
	Male	4 (13.8%)	Male	4 (13.8%)
Ethnicity	Hispanic	29 (100%)	Hispanic	29 (100%)
	Non-Hispanic	0 (0%)	Non-Hispanic	0 (0%)
Race	White	29 (100%)	White	29 (100%)
English Language Learner	Yes	6 (20.7%)	Yes	5 (17.2%)
	No	23 (79.3%)	No	24 (82.8%)
Special Education	Yes	2 (6.9%)	Yes	2 (6.9%)
	No	27 (93.1%)	No	27 (93.1%)
Migrant Status	Yes	0 (0%)	Yes	0 (0%)
	No	29 (100%)	No	29 (100%)
Homeless Status	Yes	1 (3.4%)	Yes	0 (0%)
	No	28 (96.6%)	No	29 (100%)

The results suggest that current 9th grade GPA did not significantly differ between ACE and matched non-ACE students, $t(56) = -1.40$, $p = 0.1656$. That is, the average GPA of ACE students ($M = 2.37$, $SD = 0.73$) versus matched non-ACE students ($M = 2.02$, $SD = 1.10$) differed by 0.35 but this difference was not significant. The lack of significance between ACE and non-ACE students could be attributed to the fact that these students were just beginning both their high school journey and involvement in the program. At that point, students might not have had enough time to fully engage in the program. However, we are conducting a longitudinal study for future analysis that is tracking matched 9th grade students over time. This will enable us to assess if group membership (i.e., ACE versus non-ACE) influences students' GPA throughout their high school career.

DISCUSSION

In this demonstration, we illustrated the utility of propensity score matching for evaluating science education programs. We used data from a program funded by a NIH Science Education Partnership Award to show how propensity score matching can be done in R. In this example, we matched 29 non-pathway students from a total of 325 non-pathway students to 29 pathway students. We provided the R code and described step by step instructions to aid evaluators and researchers in the use and application of propensity score matching.

Using the dataset we created, we can make requests to the school district for data on the 29 pathway students and the 29 matched non-pathway students. Therefore, partnering organizations will not need to find data from all 325 non-pathway students and 29 pathway students. Rather, follow-up assessments can consist of 29 pathway students and 29 non-pathway students. Often, individuals working in schools must work with their district offices to obtain and clean data. The reduction of personal hours involved in the extraction of this data cannot be understated. Teachers, principals and school district personnel have additional demands placed on them when an educational program is introduced in a school. In the example above, we have been able to reduce efforts in obtaining data from non-pathway students by over 90%.

If the two groups of students are being followed over time, growth models and latent growth models can be used to model longitudinal change and to determine whether group membership explains variability in intercept and slope variability. Readers are referred to Snijders and Bosker (2012), Heck, Thomas and Tabata (2022), and Little (2024) for further information on handling nested data, where participants may be nested with classroom or repeated measures may be nested within person.

In this longitudinal example, matching would ideally take place on participants' characteristics at baseline. These analyses allow for a rigorous examination of the treatment effect while accounting for potential confounding factors. The goal of this article is to provide a guide for researchers who are interested in using PSM in education data by providing a detailed demonstration and explanation using R (see Appendix). Moreover, we aim to empower researchers to make informed decisions about the selection and implementation of appropriate matching methods with PSM.

This approach to assessing the efficacy of educational interventions can also be applied to other interventions in which randomization of participants or students may not be possible. For example, research investigating interventions to improve mental health outcomes of college and university students would benefit from utilizing a propensity score matching approach to reduce bias between treatment and intervention groups. While prior work has employed the use of propensity score matching (e.g., Victor et al., 2017), this method of reducing selection bias is still not widely used. Lattie et al. (2019) demonstrated that the studies included in their meta-analysis that used nonrandomized designs—most of which did not use propensity score matching—exhibited a high bias risk. We argue that this bias would have been mitigated by using propensity score matching prior to data analysis.

While PSM is a valuable tool in casual inference, it is important to discuss limitations of this statistical technique. Propensity score matching can be used to reduce bias in estimating treatment effects; however, hidden bias can still be present due to unobserved confounders, such as students' motivation or parents' involvement (Baser, 2006). Another limitation of PSM is the assumption of overlap between groups. A lack of overlap introduces error, as it might be difficult to find good matches. Although PSM aims to identify control condition participants based on theoretically-important variables, the results of matching are sensitive to the choice of selection, which can lead to different matching

results. Despite these limitations, PSM remains a valuable method for reducing bias and improving validity when randomization is not possible.

With this demonstration, we have provided researchers with an introduction to propensity score matching. We have guided researchers in using the R programming language to perform one-to-one matching of comparable control participants to intervention participants. The code for these analyses and the data set are available so that users can practice using R for matching control participants to intervention participants. Going forward, we hope that this paper helps recipients of funding from mechanisms like the NIH Science Education Partnership Award to evaluate their science educational programs through the use of propensity score matching.

ASSOCIATED CONTENT

Supplemental material mentioned in this manuscript can be found uploaded to the same webpage as this manuscript.

AUTHOR INFORMATION

Corresponding Author

Osvaldo F. Morera. Department of Psychology, University of Texas at El Paso. 500 W. University Ave., El Paso, TX, 79968. Phone: (915) 747-5417. omorera@utep.edu

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

ACKNOWLEDGMENTS

The authors would like to acknowledge Dr. Thomas Boland for assistance in obtaining data for Study 1. We would also like to dedicate this paper to the late Dr. Tony Beck, former Program Officer for the NIH Science Education Partnership Award (SEPA) funding mechanism. Dr. Beck was a tireless advocate for the SEPA program and he will be missed.

FUNDING SOURCES

This research was supported by grant 1R25GM132959-04. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

ABBREVIATIONS

ACE: Project ACtion for Equity; ANCOVA: Analysis of Covariance; ELL: English Language Learner; MLR: Multiple Linear Regression; NIH: National Institutes of Health; OSF: Open Science Framework; PSM: Propensity Score Matching; SEPA: Science Education Partnership Award; SES: Socioeconomic Status

REFERENCES

Austin, P.C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies, *Multivariate Behavioral Research*, 46, 399-424.

Baser, O. (2006). Too much ado about propensity score models? Comparing methods of propensity score matching. *Value in Health*, 9(6), 377-385.

Benedetto, U., Head, S. J., Angelini, G. D., & Blackstone, E. H. (2018). Statistical primer: propensity score matching and its alternatives. *European Journal of Cardio-Thoracic Surgery*, 53(6), 1112-1117. https://doi.org/10.1093/ejcts/ezy167

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Lawrence Erlbaum Associates Publishers.

Fan, X., & Nowell, D. L. (2011). Using propensity score matching in educational research. *Gifted Child Quarterly*, 55(1), 74-79. https://doi.org/10.1177/0016986210390635

Gu, X.S. & Rosenbaum, P.R. (1993). Comparison of multivariate matching methods: Structure, distances and algorithms. *Journal of Computational and Graphic Statistics*, 2, 405-420.

Heck, R.H., Thomas, S.L. & Tabata, L.N. (2022). *Multilevel and Longitudinal Modeling with IBM SPSS* (3rd edition). New York, NY: Routledge

Ho, D., Imai, K., King, G., Stuart, E. (2011). "MatchIt: Nonparametric Preprocessing for Parametric Causal Inference." *Journal of Statistical Software*, 42(8), 1-28. doi:10.18637/jss.v042.i08.

Kane, L. T., Fang, T., Galetta, M. S., Goyal, D. K., Nicholson, K. J., Kepler, C. K., Vaccaro, A. R., & Schroeder, G. D. (2020). Propensity score matching: a statistical method. *Clinical Spine Surgery*, 33(3), 120-122. https://doi.org/10.1097/BSD.0000000000000932

Keppel, G. & Wickens, T.D. (2004). *Design and analysis: A researcher's handbook* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

Lattie, E. G., Adkins, E. C., Windquist, N., Stiles-Shields, C., Wafford, Q. E., & Graham, A. K. (2019). Digital mental health interventions for depression, anxiety, and enhancement of psychological well-being among college students: Systematic review. *Journal of Medical Internet Research*, 21(7), e12869. https://doi.org/10.2196/12869

Little, T.D. (2024). *Longitudinal Structural Equation Modeling* (2nd edition). New York, NY: Guilford Press.

Matthay, E.C., Hagan, E., Gottlieb, L.M. Tan, M. L., Vlahov, D., Adler, N. E., & Glymour, M. M., Alternative causal inference methods in population health research: Evaluating tradeoffs and triangulating evidence. *Population Health*, 10, 100525

Olmos, A., & Govindasamy, P. (2015). A practical guide for using propensity score weighting in R. *Practical Assessment, Research, and Evaluation*, 20(13), 1-8. https://doi.org/10.7275/jjtm-r398

Paradis, E. (2005). *R for Beginners* (pp. 1-71). Institut des Sciences de l'Evolution. Université Montpellier II.

Pearl, J. (2000). Causality: Models, reasoning and inference applications. New York:Cambridge University press.

R Core Team (2023). R: A language and environment for statistical computing (4.3.0). [Computer software]. Foundation for Statistical Computing. http://www.R-project.org/.

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

Snijders, T.A.B. & Bosker, R.J. (2012). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling* (2nd edition). London, England: Sage

Taggart, A. (2018). Latina/o students in K-12 schools: A synthesis of empirical research on factors influencing academic achievement. *Hispanic Journal of Behavioral Sciences*, 40(4), 448-471.

Thoemmes, F. J., & Kim, E. S. (2011). A systematic review of propensity score methods in the social sciences. *Multivariate Behavioral Research*, 46(1), 90-118. https://doi.org/10.1080/00273171.2011.540475

Victor, P. P., Teismann, T., & Willutzki, U. (2017). A pilot evaluation of a strengths-based CBT intervention module with college students. *Behavioural and Cognitive Psychotherapy*, 45(4), 427-431. https://doi.org/10.1017/S1352465816000552

Yoshida, K. Hernández-Díaz, S., Solomon, D.H., et al., (2017). Matching Weights to Simultaneously Compare Three Treatment Groups: Comparison to Three-way Matching. *Epidemiology*, 28, 387-395.

Zakrison, T. L., Austin, P. C., & McCredie, V. A. (2018). A systematic review of propensity score methods in the acute care surgery literature: avoiding the pitfalls and proposing a set of reporting guidelines. *European Journal of Trauma and Emergency Surgery*, 44, 385-395. https://doi.org/10.1007/s00068-017-0786-6

Zhao, Q., Luo, J., Su, Y., Zhang, Y., Tu, G., & Luo, Z. (2021). Propensity score matching with R: Conventional methods and new features. *Annals of Translational Medicine*, 9(9), 1-39. https://doi.org/10.21037/atm-20-3998