Alberto Guzman-Alvarez; alg223@pitt.edu

Introduction

Over the last two decades, causal inference has played a central role in education research by providing tools and analytic strategies to generate evidence on the effectiveness of educational interventions and policies (Cook et al., 2002; Murnane & Willett, 2010; Rosenbaum, 2010, 2020). The push for causal work has been driven by policymakers and funding agencies emphasizing rigorous evaluations, which can answer questions regarding cause and effect, and exploring the mechanisms through which interventions may operate. Simultaneously, there has been a rapid emergence of increasingly sophisticated quantitative methods and access to an ever-increasing amount of high-dimensional administrative data (Daniel, 2019; Gibson & Ifenthaler, 2016; Williamson, 2017). However, there has been a lack of analytical tools that can deal with high-dimensional data in causal inference, especially in applied educational research.

High-dimensional data complicates the issue of causal inference because, generally, as you increase the number of variables, traditional propensity score based causal methods degrade in performance (D'Amour et al., 2020). Additionally, 'big' data creates computational issues using traditional statistical tools, where due to the sheer volume of data models may fail to run (Daniel, 2019; Prinsloo & Slade, 2016). However, computer science has generated a wealth of research on a potential solution: Deep Neural Networks (DNN) (Deng & Yu, 2014; Hernández-Blanco et al., 2019; LeCun et al., 2015; Perrotta & Selwyn, 2019). DNN's are algorithms based on our brain's neural architecture. An example of a simple DNN is found in Figure 1. They are commonly used in industry to model complex prediction and classification tasks and have already proven to excel in modeling high-dimensional data (LeCun et al., 2015). However, no work to date has applied DNN's to causal inference questions, particularly in educational research.

Questions of causality have traditionally been addressed using Randomized Control Trials (RCT), whereby students are randomly assigned to different treatment groups (e.g., Treatment|Control) (Murnane & Willett, 2010). The act of randomization assures, on average, that both groups will be balanced on all observable *and* unobservable characteristics, ruling out that an intervention impact is due to the differences in the group compositions (Rosenbaum, 2010, 2020). However, ethical and cost concerns can sometimes stand in the way of experimental studies in

education (Murnane & Willett, 2010). Instead, we often depend on quasi-experimental methods and observational data where students either self-select or are placed into an intervention without randomization.

In education, propensity score analysis is one of the most widely used quasi-experimental methods (Fan & Nowell, 2011; Powell et al., 2019; Rosenbaum, 2020; Stuart, 2007; Thoemmes & Kim, 2011). Propensity score analysis is used in non-experimental studies to balance the observed characteristics between treated and un-treated students, just as randomization to treatment and control conditions creates balance on observable variables. The propensity score represents the probability that a student *would have* been exposed to treatment, conditional on observable variables (Rosenbaum & Rubin, 1983, 1984). For example, recently, propensity score analysis has been applied in evaluating the impact of a math curriculum on high school students.

However, to estimate unbiased treatment effects correctly using propensity scores, certain strict assumptions must be met; *ignorability assumption* (Rosenbaum & Rubin, 1981, 1984):

(1) The *propensity score model* must capture *all* of the characteristics related to the selection into treatment, leaving out no potential confounders (Rosenbaum, 2010, 2020). Confounders are variables that influence both the treatment and the outcome.

(2) The *propensity score model* must correctly model the association between the student's characteristics and treatment selection, meaning that all proper interactions and non-linear terms should be specified (Rosenbaum, 2010, 2020).

The literature suggests a "kitchen sink" approach for variable selection to guard against the first condition since the penalties are high if a potential confounder is *not* included in the *propensity score model* (Pirracchio et al., 2015; Shortreed & Ertefaie, 2017). For the second condition, the literature suggests iterating through various model specifications, with each iteration adding interactions or non-linear terms until a balance is achieved amongst the observed characteristics between treated and un-treated students (Murnane & Willett, 2010). The consequences are steep if these assumptions are not met, as it will lead to biased treatment effect estimates.

These conditions may be difficult to satisfy in the social science literature, especially education, since behavioral data has many more intricacies than data used in the propensity score simulation

literature (Thoemmes & Kim, 2011). However, the literature provides few methods for the critical step of estimating the propensity score (Stuart, 2007; Thoemmes & Kim, 2011). The available research base suggests that the most widely used models for estimating the propensity score in education are logistic regression models, which are inadequate and lead to biased treatment effects when incorrectly specified (Lee et al., 2010; Pan & Bai, 2015; Setoguchi et al., 2008; Westreich et al., 2010). For example, logistic models tend to degrade when modeling high-dimensional data. Therefore, a key question is whether DNN's represent a viable and worthy alternative to estimating the propensity score.

DNN's have multiple properties that make them worthy of consideration for propensity score estimation:

(1) DNN's are highly flexible and capable of capturing complex interactions and non-linearities, and

(2) DNN's allow for automatic variable selection (Hernández-Blanco et al., 2019; LeCun et al., 2015; Zou et al., 2019), which aides in having to iterate through various models adding interactions or non-linear terms.

DNN's have already been shown to outperform logistic models and machine learning algorithms in simulations outside of propensity score analyses (Dahl et al., 2013; Hu et al., 2015; Mousavi et al., 2016). These methods could represent an essential tool for researchers in estimating the correct *propensity score model*, in general, and mainly when dealing with large datasets that include hundreds of variables on students, and when those variables have complex associations. Therefore, the fundamental question that motivates my dissertation is:

How can we improve propensity score estimations in large observational datasets across various education contexts with modern DNN techniques?

My dissertation includes three related studies: two methodological and a substantive application. *In the first study*, I will develop a new method for estimating the propensity score based on recent advancements in information and computer science on DNN's. *In the second study*, I will extend this approach into a causal mediation framework, which uncovers the mechanisms of

intervention effects. *In the final study*, I will apply my method to a large-scale evaluation of a college access nudge intervention implemented during the COVID-19 pandemic.

As with most novel methodological advances, a new method may be challenging to implement, requiring the researcher to write extensive code. Therefore, in addition to academic papers, this dissertation's key product will be an open-source R package, which is accompanied by a users' manual to make my proposed method publicly available and easily implemented by other researchers.

In the remainder of this narrative, I provide a brief overview of the literature on deep neural networks and causal mediation analysis, describe my research plan and progress for all three studies, and conclude with a brief summary.

Deep Neural Networks for Propensity Score Estimation

In my first study, I aim to overcome the limitation of existing methods by proposing a novel DNN-based approach to propensity score estimations. A precursor to my method – *single-layer feed-forward neural network (NN)* – was initially proposed by Setoguchi et al. (2008) and Keller et al. (2015) for propensity score estimation. Their simulation work showed that a NN could outperform the traditional logistic model, leading to less bias in the estimated treatment effect and improved the covariate balance. NN's only allow for the modeling of relatively simple relationships from the data unlike DNN's, which due to its architecture, can calculate exponentially complicated relationships between variables, making it suitable for high-dimensional data (LeCun et al., 2015).

DNN's are based on how our brains process information through the flow of data through interconnected neurons. DNN's can uncover complex relationships in high-dimensional data with better precision than traditional statistical methods and machine learning methods, which were novel just a decade ago (Hernández-Blanco et al., 2019; LeCun et al., 2015; Schmidhuber, 2014; Zou et al., 2019). A DNNs ability to identify the intricate relationships withing large-volume, high-dimensional data makes these algorithms applicable to causal methods, like propensity score analysis.

The architecture of DNN's comprises of multiple layers of processing units (i.e., neurons) that apply mathematical transformations to inputs fed-forward through various layers (See Figure 1) (Hernández-Blanco et al., 2019; Zou et al., 2019). In each layer, additional transformations are applied based on the previous layers (i.e., learning), creating a network that can learn increasingly

complex associations between characteristics. The last layer of the DNN is an output layer, which is adaptable for several uses, like prediction and classification (LeCun et al., 2015). The "deep" in deep neural networks comes from the number of layers between the input and output layers; the deeper the DNN, the more complex relationship the network can estimate (Hernández-Blanco et al., 2019).

DNN's have gained significant attention in the industry and academic research, making considerable advances in image recognition and natural language processing (Krittanawong et al., 2019; LeCun et al., 2015; Zou et al., 2019). Recent investments by Google in software development now makes it easy to develop DNN models using user-friendly programming languages (Doleck et al., 2020; Pang et al., 2019). Though to date, DNN has not been applied to the causal inference.

At first glance, high-dimensional data should be an asset to propensity score analysis since conventional practices stress the importance of including all available variables in the *propensity score model*, given that it increases the likelihood of capturing all relevant confounders. However, this is not the case. The traditional logistic model for propensity score estimation tends to degrade in performance as more variables are included in the estimation step (Hill et al., 2011). This can manifest in one of several ways when dealing with high-dimensional data, for example:

(1) The addition of a large number of variables may lead the logistic model to simply fail to run and not produce the required propensity score.

(2) If the model does run successfully, the model will likely overfit the data, meaning that the logistic model begins to model the random error in the data rather than the relationship between variables. This leads to propensity scores that have little to no variability, and are therefore not suitable for propensity score analysis.

(3) Lastly, in a high-dimensional setting, it becomes increasingly difficult to iteratively model all of the suitable interactions and non-linearities with a logistic model (D'Amour et al., 2020; Dorie et al., 2019; Hill et al., 2011).

Propensity score estimation is a crucial step in a propensity score analysis. Given that if the propensity score model is misspecified, there may be inadequate overlap between the treatment and control groups' estimated propensity scores. The insufficient overlap makes it challenging to balance the groups based on their pre-treatment characteristics, leading to biased treatment effects (D'Amour

et al., 2020; Murnane & Willett, 2010; Rosenbaum, 2010, 2020). In this simulation study, I aim to introduce DNNs into the field of education research and show that their unique properties, such as flexibility and lax model assumptions, are suitable and appropriate for propensity score estimation.

Contributions

In this first study, I aim to make the following methodological contributions.

First, I will incorporate a novel framework for simulating educational data modeled on a large-scale college access intervention study focused on providing outreach on a college-going task to nearly 170k high school students. This data set includes any information the students provided on their college application via the CommonApp. This database is indicative of the administrative data that is becoming more common in education research. Therefore, basing my simulated data on the CommonApp data, I can capture representative relationships and data types, which educational researchers experience in their studies. I will use this data to create simulated data, where I varied the sample size and number of variables and induce associations and treatment effects to test my method's performance in extracting unbiased treatment impacts. Using this framework, I will add to the propensity score simulation literature by simulating real educational data, which moves away from conventional and current propensity score simulation studies in epidemiology that simulate relatively small datasets (Cannas & Arpino, 2019; Karim et al., 2018; Lee et al., 2010; Pirracchio et al., 2015; Setoguchi et al., 2008).

Second, I will extend the simulation literature by testing complexities in both the *propensity score* and *outcome models*. This is relatively novel in propensity score literature (Cannas & Arpino, 2019), but more realistic for the education context where complicated relationships between variables, treatment assignment, and the outcome could exist. DNNs have been shown to work well in modeling these complexities by creating a parsimonious model that estimates and evaluates all possible interactions and non-linearities among *all* variables (Hernández-Blanco et al., 2019; Perrotta & Selwyn, 2019).

Third, I will provide the first evaluation of DNN's applied to propensity score analysis. After simulating the data, I will test the DNN's performance against the logistic model and other well-documented estimation techniques used in propensity score literature (e.g., SVM, CART, BART). I

evaluate the methods' performance when implemented in both propensity score-based weighting and propensity score-based matching, by assessing how well they achieve a covariate balance, and the bias in treatment effect estimates will be evaluated. Additionally, I will compare how robust these methods are to an unobserved confounder.

Progress

I conducted Monte Carlo simulations to assess this new analytical method's performance, following the guidance of recently published best-practices for simulation studies (Morris et al., 2019). My preliminary simulation results show that deep neural networks outperform the traditional logistic model, and a range of currently available machine learning approaches when interactions and non-linearities are introduced in any data generation models (e.g., *propensity score model*, *outcome model*).

Integrating Deep Neural Networks into Causal Mediation Analysis

Policymakers and researchers in education are increasingly interested in the overall impacts of a given intervention, and the processes and mechanisms through which the intervention operates (Nguyen et al., 2020). In education research, the standard approach to mediation analysis has been path analysis and structural equation modeling (SEM) (Bauer et al., 2006; Kenny et al., 2003). However, the traditional regression-based approach to causal mediation relies heavily on correct model specifications. Encouragingly, recent developments in *causal* mediation analysis remedy this concern, based on a flexible propensity score-based weighting technique, known as the ratio-of-mediator-probability weighting (RMPW), which does not require strong assumptions about the functional form of the outcome model (Hong, 2010).

In a basic mediation framework, a treatment affects a focal mediator, which affects an outcome. The total treatment effect can be decomposed into a *natural indirect effect* transmitted through a focal mediator, and a *natural direct effect* that represents the contribution of all the other unmodeled pathways by which the treatment affects the outcome. The identification of the natural indirect and direct effects relies on the *sequential ignorability assumption* (Imai, Keele, & Tingley, 2010; Imai, Keele, & Yamamoto, 2010)i.e., within levels of pre-treatment variables.

(1) The treatment is as if randomized, and

(2) The mediator is as if randomized within each treatment group.

Under this assumption, selection bias associated with non-random treatment and mediator assignment can be removed via weights constructed based on the treatment and the mediator's propensity scores.

Although a weighting-based causal mediation analysis is robust to outcome model misspecifications, possible misspecifications of the propensity score models may still lead to bias in the natural indirect and direct effect estimates. Logistic regressions have always been fitted for estimating propensity scores in the literature of weighting-based causal mediation analysis. Given the flexibility of DNN's, it is essential to assess if DNN's outperform logistic regressions in estimating the natural indirect and direct effects, especially when there are many observed variables. Therefore, in my second study, I will extend the proposed DNN-based method to a novel propensity score weighting method initially (RMPW) developed by Hong (2010) using the simulated data from my first study. The preliminary evidence of my first study naturally extends to these propensity-score based methods for causal mediation.

Contributions

I will attempt to make the following methodological contributions in this second study.

First, I will evaluate the alternative methods for estimating propensity scores in weighting-based causal mediation analyses using a DNN approach developed in my first study.

Second, I will examine whether alternative methods (e.g., DNN, SVM, CART, BART) may outperform logistic regressions for estimating the natural indirect and direct effects, under various scenarios where increasing complexity levels are generated in the treatment and mediator models. These scenarios have not been empirically tested in the causal mediation literature.

Estimating Impacts of a Large-Scale College Access Program During COVID-19

In my final study, I aim to illustrate how my DNN-based method can be applied to non-randomized large-scale education evaluations. At the start of the COVID-19 pandemic, a massive endeavor was undertaken by AdmitHub, CommonApp, and the College Advising Corps to help low-income high school students transition to college. The intervention provided outreach and guidance regarding college-going tasks to nearly 170k students across the nation.

The intervention was quickly deployed to help thousands of students and, as a result, was not implemented in the context of an experimental study. Nevertheless, understanding the program's impact will benefit students and policymakers interested in lowering college access barriers while providing evidence of the pandemic's effect on college access. This study is part of a more extensive collaborative research project. I will evaluate this intervention's impact by using my DNN-based propensity score estimation method developed in my first study.

Data

The research team collected a large amount of rich information on students. Essentially, anything that students provided on their college application will be available to us. This includes demographic information, standardized test scores, and data around college choices. Our *outcome* of interest will be a binary indicator of whether a student submitted their application. With potentially hundreds of observed variables, this dataset provides a textbook case to empirically test my DNN-based approach.

Sample

Students targeted for the intervention are from the class of 2020 and had to be both first-generation college students and low-income, as defined by qualifying for the CommonApp fee waiver. Our sample includes more than 170k students who were targeted for the intervention. We will refer to this group as the *intervention students*. CommonApp will also provide us with information from students who were not targeted for the intervention yet completed a CommonApp application. We will refer to this group as the *matched-intervention students;* this is discussed further below. In this final study, I will attempt to answer the following research question.

What is the impact of being targeted for outreach on college application submission?

Analysis

Given that students were not randomized into the intervention, we need to take a non-experimental approach to assess the program's impact. I will utilize my DNN-based propensity score method to match *intervention students* to students from the same high school who meet one but not both eligibility criteria and who are otherwise observationally similar on observable variables available through CommonApp data. Matching to students within the same high school is beneficial,

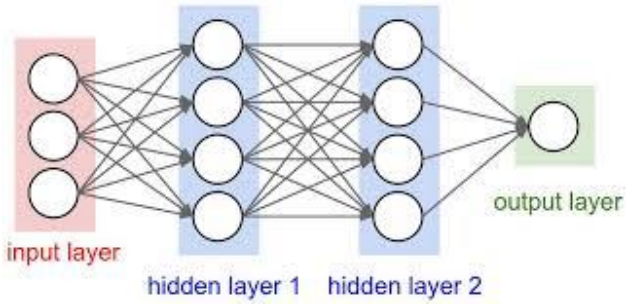
given that college opportunities vary geographically, and many college-goers attend college in relative proximity to their home (Hillman, 2016). We will refer to this group of same-cohort matched counterparts as *intervention-matched students*. Using a regression approach, we will compare the college application submission rate between *intervention students* and *intervention-matched students* to estimate the causal effect of being targeted for the outreach intervention on submitting a college application. The data for this study is currently being compiled but will become available before the start of the fellowship.

Conclusion

The growth of large data sets and deep learning methods presents opportunities to expand quantitative research in education. In my dissertation, I aim to show the benefits of deep neural networks in causal inference research. My methods will be beneficial for applied educational researchers examining program impacts with observational data. My methodological contributions will overcome significant limitations of existing propensity score estimation approaches, under high-dimensional data, and complex treatment selection. Finally, I will provide a substantive application of my method to a large-scale college-access intervention. This set of tools will inform not only program development but educational policy at-large.

Alberto Guzman-Alvarez; alg223@pitt.edu

Figure 1. Simple Deep Neural Network



Works Cited

Bauer, D. J., Preacher, K. J., & Gil, K. M. (2006). Conceptualizing and Testing Random Indirect Effects and Moderated Mediation in Multilevel Models: New Procedures and Recommendations. *Psychological Methods*, *11*(2), 142–163. https://doi.org/10.1037/1082-989x.11.2.142

Cannas, M., & Arpino, B. (2019). A comparison of machine learning algorithms and covariate balance measures for propensity score matching and weighting. *Biometrical Journal*, *61*(4), 1049–1072. https://doi.org/10.1002/bimj.201800132

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

Dahl, G. E., Stokes, J. W., Deng, L., & Yu, D. (2013). Large-Scale Malware Classification using Random Projections and Neural Networks. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 3422–3426. https://doi.org/10.1109/icassp.2013.6638293

D'Amour, A., Ding, P., Feller, A., Lei, L., & Sekhon, J. (2020). Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*. https://doi.org/10.1016/j.jeconom.2019.10.014

Daniel, B. K. (2019). Big Data and data science: A critical review of issues for educational research. *British Journal of Educational Technology*, *50*(1), 101–113. https://doi.org/10.1111/bjet.12595

Deng, L., & Yu, D. (2014). Deep Learning: Methods and Applications. *Foundations and Trends® in Signal Processing*, *7*(3–4), 197–387. https://doi.org/10.1561/2000000039

Doleck, T., Lemay, D. J., Basnet, R. B., & Bazelais, P. (2020). Predictive analytics in education: a comparison of deep learning frameworks. *Education and Information Technologies*, *25*(3), 1951–1963. https://doi.org/10.1007/s10639-019-10068-4

Dorie, V., Hill, J., Shalit, U., Scott, M., & Cervone, D. (2019). Automated versus Do-It-Yourself Methods for Causal Inference: Lessons Learned from a Data Analysis Competition. *Statistical Science*, *34*(1), 43--68. https://projecteuclid.org/euclid.ss/1555056030

Fan, X., & Nowell, D. L. (2011). Using propensity score matching in educational research. *Gifted Child Quarterly*, *55*(1), 74--79.

Gibson, D. C., & Ifenthaler, D. (2016). *Big Data and Learning Analytics in Higher Education, Current Theory and Practice*. 29–42. https://doi.org/10.1007/978-3-319-06520-5_4

Hernández-Blanco, A., Herrera-Flores, B., Tomás, D., & Navarro-Colorado, B. (2019). A Systematic Review of Deep Learning Approaches to Educational Data Mining. *Complexity*, *2019*, 1–22. https://doi.org/10.1155/2019/1306039

Hill, J., Weiss, C., & Zhai, F. (2011). Challenges With Propensity Score Strategies in a High-Dimensional Setting and a Potential Alternative. *Multivariate Behavioral Research*, *46*(3), 477--513. https://doi.org/10.1080/00273171.2011.570161

Hillman, N. W. (2016). Geography of College Opportunity. *American Educational Research Journal*, *53*(4), 987–1021. https://doi.org/10.3102/0002831216653204

Hong, G. (2010). *Ratio of Mediator Probability Weighting for Estimating Natural Direct and Indirect Effects*. American Statistical Association.

Hu, W., Qian, Y., Soong, F. K., & Wang, Y. (2015). Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers. *Speech Communication*, *67*, 154–166. https://doi.org/10.1016/j.specom.2014.12.008

Imai, K., Keele, L., & Tingley, D. (2010). A General Approach to Causal Mediation Analysis. *Psychological Methods*, *15*(4), 309–334. https://doi.org/10.1037/a0020761

Imai, K., Keele, L., & Yamamoto, T. (2010). Identification, Inference and Sensitivity Analysis for Causal Mediation Effects. *Statistical Science*, *25*(1), 51–71. https://doi.org/10.1214/10-sts321

Karim, M. E., Pang, M., & Platt, R. W. (2018). Can We Train Machine Learning Methods to Outperform the High-dimensional Propensity Score Algorithm? *Epidemiology*, *29*(2), 191–198. https://doi.org/10.1097/ede.0000000000000787

Kenny, D. A., Korchmaros, J. D., & Bolger, N. (2003). Lower Level Mediation in Multilevel Models. *Psychological Methods*, *8*(2), 115–128. https://doi.org/10.1037/1082-989x.8.2.115

Krittanawong, C., Johnson, K. W., Rosenson, R. S., Wang, Z., Aydar, M., Baber, U., Min, J. K., Tang, W. H. W., Halperin, J. L., & Narayan, S. M. (2019). Deep learning for cardiovascular

medicine: a practical primer. *European Heart Journal*, *40*(25), 2058–2073. https://doi.org/10.1093/eurheartj/ehz056
LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436--444. 10.1038/nature14539
Lee, B. K., Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine*, *29*(3), 337--346.
Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074–2102. https://doi.org/10.1002/sim.8086
Mousavi, S. M., Horton, S. P., Langston, C. A., & Samei, B. (2016). Seismic features and automatic discrimination of deep and shallow induced-microearthquakes using neural network and logistic regression. *Geophysical Journal International*, *207*(1), 29–46. https://doi.org/10.1093/gji/ggw258
Murnane, R. J., & Willett, J. B. (2010). *Methods matter: Improving causal inference in educational and social science research*. Oxford University Press.
Nguyen, T. Q., Schmid, I., & Stuart, E. A. (2020). Clarifying causal mediation analysis for the applied researcher: Defining effects based on what we want to learn. *Psychological Methods*. https://doi.org/10.1037/met0000299
Pan, W., & Bai, H. (2015). *Propensity Score Analysis: Fundamentals and Developments*. https://doi.org/https://doi.org/10.1007/s41237-018-0058-8
Pang, B., Nijkamp, E., & Wu, Y. N. (2019). Deep Learning With TensorFlow: A Review. *Journal of Educational and Behavioral Statistics*, *45*(2), 227–248. https://doi.org/10.3102/1076998619872761
Perrotta, C., & Selwyn, N. (2019). Deep learning goes to school: toward a relational understanding of AI in education. *Learning, Media and Technology*, *45*(3), 1–19. https://doi.org/10.1080/17439884.2020.1686017
Pirracchio, R., Petersen, M. L., & Laan, M. van der. (2015). Improving Propensity Score Estimators' Robustness to Model Misspecification Using Super Learner. *American Journal of Epidemiology*, *181*(2), 108–119. https://doi.org/10.1093/aje/kwu253
Powell, M. G., Hull, D. M., & Beaujean, A. A. (2019). Propensity Score Matching for Education Data: Worked Examples. *The Journal of Experimental Education*, *88*(1), 1–20. https://doi.org/10.1080/00220973.2018.1541850
Prinsloo, P., & Slade, S. (2016). *Big Data and Learning Analytics in Higher Education, Current Theory and Practice*. 109–124. https://doi.org/10.1007/978-3-319-06520-5_8
Rosenbaum, P. R. (2010). *Design of Observational Studies*. Springer New York. https://doi.org/10.1007/978-1-4419-1213-8
Rosenbaum, P. R. (2020). *Design of Observational Studies*. https://doi.org/10.1007/978-3-030-46405-9
Rosenbaum, P. R., & Rubin, D. B. (1981). *The Central Role of the Propensity Score in Observational Studies for Causal Effects.* https://doi.org/10.21236/ada114514
Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41
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*(387), 516. https://doi.org/10.2307/2288398
Schmidhuber, J. (2014). Deep Learning in Neural Networks: An Overview. *ArXiv*. https://doi.org/10.1016/j.neunet.2014.09.003
Setoguchi, S., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., & Cook, E. F. (2008). Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiology and Drug Safety*, *17*(6), 546–555. https://doi.org/10.1002/pds.1555
Shortreed, S. M., & Ertefaie, A. (2017). Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics*, *73*(4), 1111--1122. https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.12679
Stuart, E. A. (2007). Estimating Causal Effects Using School-Level Data Sets. *Educational Researcher*, *36*(4), 187–198. https://doi.org/10.3102/0013189x07303396

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

Westreich, D., Lessler, J., & Funk, M. J. (2010). Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology*, *63*(8), 826--833. http://www.sciencedirect.com/science/article/pii/S0895435610001022

Williamson, B. (2017). *Big Data in Education: The digital future of learning, policy and practice*. https://doi.org/10.4135/9781529714920

Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., & Telenti, A. (2019). A primer on deep learning in genomics. *Nature Genetics*, *51*(1), 12–18. https://doi.org/10.1038/s41588-018-0295-5