Practical Assessment, Research, and Evaluation

Volume 27 *(Volume 27, 2022)* Article 3

February 2022

Explained: Artificial Intelligence for Propensity Score Estimation in Multilevel Educational Settings

Zachary K. Collier
University of Delaware

Haobai Zhang
University of Delaware

Liu Liu
University of Delaware

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 27 Number 3, February 2022 ISSN 1531-7714

Explained: Artificial Intelligence for Propensity Score Estimation in Multilevel Educational Settings

Zachary K. Collier, *University of Delaware*
Haobai Zhang, *University of Delaware*
Liu Liu (*University of Delaware*

Although educational research and evaluation generally occur in multilevel settings, many analyses ignore cluster effects. Neglecting the nature of data from educational settings, especially in non-randomized experiments, can result in biased estimates with long-term consequences. Our manuscript improves the availability and understanding of artificial neural networks, an underutilized method trending in other disciplines. This method also shows promise for dealing with challenges faced by educational researchers, such as analyzing clustered data. Therefore, we simulated data to generalize the potential benefits of artificial neural networks to different data types. We also compared artificial neural networks to more familiar methods and investigated the time it demanded to perform each technique. Hence, readers can decide when it may be more appropriate to use one method instead of another.

Introduction

Education research and evaluation are dynamic, often due to their multilevel and observational nature. The methodological challenges associated with multilevel, observational data include potential selection bias, non-negligible clustered effects, and omitted variable bias (Barnard et al., 2013; Bellara, 2013; Yang et al., 2017). Theoretically, researchers address each of these challenges to estimate unbiased causal effects of educational interventions and policies. But, in practice, attempts to account for complexity, such as students nested within schools and the many possible combinations of interactions between student-level and school-level confounders, may lead to low convergence rates and inaccurate estimates. Furthermore, "big data" makes model specifications challenging to align with theory-based relationships due to many confounders.

This study aims to improve the availability and understanding of artificial neural networks (NN), a long-existing method underutilized in education research. NN is robust to assumptions of conventional statistical models and does not require manual specification of complex relationships (Collier & Leite, 2020). In recent simulation and empirical studies related to public health and drug safety, NN produced more accurate estimates than conventional methods while estimating propensity scores for single-level treatments (Setoguchi et al., 2008). The current paper demonstrates how to apply NN to estimate propensity scores for scenarios more likely to appear in educational research.

Literature Review

Rosenbaum and Rubin (1983) first introduced propensity score methods to deal with selection bias in single-level observational experiments. Later, several researchers extended propensity score methods to multilevel settings (Arpino & Mealli, 2011; Eckardt,

2012; McCormick et al., 2013; Xiang & Tarasawa, 2015). However, as mentioned above, these researchers focused exclusively on binary (treatment/control) treatments. Zhu et al. (2015) pointed out an important gap in the literature, i.e., no propensity score approach to cover continuous, multilevel treatments in public health.

Continuous, multilevel treatments also are common in education research and evaluation. A highly relevant example due to Covid-19 is online learning. A continuous treatment could be students' time spent using online test-prep platforms before taking an end-of-course examination (e.g., Aeite et al., 2019; Mitten et al., 2021). If students exposed to the online treatment live in different places (each place with its own educational standards), a multilevel propensity score analysis with continuous treatment exposure could reduce selection bias. This technique would require several special considerations, including deciding how to model the impact of confounding variables at the student and home environment levels. Failure to solve covariates' hierarchy may lead to biased estimated effects of the online test-prep exposure (Thoemmes, 2009). With the growing appeal of propensity score methods and the surging need to study causal relationships in educational environments, researchers need to be informed and apply the most optimal estimation methods.

Propensity Scores in Multilevel Settings

When estimating generalized propensity scores (GPS) for multilevel settings, the traditional approaches include single-level (SL), fixed-effect (FE), random-intercepts (RI), and random-slopes (RS) models. We discuss these approaches in the following section.

Generalized Linear Models. A SL model takes no notice of the hierarchical nature of the data, while both FE, RI, and RS models include a cluster-specific intercept for each j cluster to explain the unobserved heterogeneity among clusters (Schuler et al., 2016).

The individual level covariates X and cluster level covariates W are included in a SL model (Thoemmes & West, 2011):

$$e(x,w) = \beta_0 + \sum_{p=1}^{P}\beta_p X_i + \sum_{q=1}^{Q}\beta_q W_j + \sum_{i=1}^{I}\beta_i W_j X_i \quad (1)$$

where $e(x,w)$ is the estimated GPS, β_0 is an intercept, $\sum_{p=1}^{P}\beta_p X_i$ is a vector of regression coefficients and individual-level covariates, $\sum_{q=1}^{Q}\beta_q W_j$ is a vector of regression coefficients and cluster-level covariates, and $\sum_{i=1}^{I}\beta_i W_j X_i$ denotes all possible interactions between individual- and cluster-level covariates.

A FE model includes a dummy coded indicator C for each cluster:

$$e(x,w) = \sum_{p=1}^{P}\beta_p X_i + \sum_{c=1}^{C}\beta_c C_i + \sum_{i=1}^{I}\beta_i C_i X_i \quad (2)$$

where $\sum_{i=1}^{I}\beta_i C_i X_i$ represents all possible interactions between individual-level covariates and indicators of each cluster. In practice, researchers should not include cluster level covariates in FE models to avoid perfect collinearity.

Random effects models may include random intercepts, slopes, or both. A full random effects model estimates generalized propensity scores based on both fixed and random effects:

$$e(x,w) = \gamma_{00} + \sum_{p=1}^{P}\gamma_{p0}X_{ij} + \sum_{q=1}^{Q}\gamma_{0q}W_j + \sum_{i=1}^{I}\gamma_{1i}W_j X_{ij} + u_{0j} + \sum_{p=1}^{P}X_{ij}u_{1j} \quad (3)$$

where $\sum_{p=1}^{P}\gamma_{p0}X_{ij}$ represents the regression coefficients and individual-level covariates, $\sum_{q=1}^{Q}\gamma_{0q}W_j$ is a vector of regression coefficients and cluster level covariates, $\sum_{i=1}^{I}\gamma_{1i}W_j X_{ij}$ is the vector of all interactions between individual-level and cluster-level covariates, u_{0j} is the random effects influencing the intercept of each cluster *j*, and $\sum_{p=1}^{P}X_{ij}u_{1j}$ is the random effects influencing each of the regression slopes of individual-level covariates. The random effects, *u* are assumed to the normal distribution with a mean of zero and an estimated variance of τ (Thoemmes & West, 2011).

Previous literature on GPS estimation for multilevel settings is mixed. For example, Arpino and Mealli (2011) favored the FE model, while Kim and Seltzer (2007) favored random effects models. Both approaches allow the treatment assignment to differ across clusters. However, FE models are limited in that they 1) may not properly estimate propensity scores when sample sizes within clusters are small and 2) may have convergence issues if the number of clusters is large (Thoemmes & West, 2011). Nevertheless, an advantage of FE models is that they remove all

confounding due to cluster-level covariates without specifying said covariates in the propensity score model. An RI model is the simplest case of a random effects model based on the FE model with a randomly varying cluster effect (Schunck, 2013). A full random effects model (RS) is often perceived as the most realistic model because it allows heterogeneity in both intercepts and slopes (Li et al., 2013).

Generalized linear models are attractive for propensity score analysis because they are relatively familiar to educational researchers and are easy to carry out in statistical software (e.g., R, SAS). However, there are several limitations for most familiar methods, including insufficient attention to crucial assumptions (e.g., the correct concretization of complex relationships), small sample bias with maximum likelihood, and unstable estimates with large numbers of covariates (Keller et al., 2013; Maroco et al., 2011; Schumacher et al., 1996; Setoguchi et al., 2008; Weitzen et al., 2004). Any of these limitations could lead to failing to achieve covariate balance, resulting in biased estimates of treatment effects (Rubin, 2010). This article will discuss covariate balance further in a later section as it is an essential outcome for our study. Next, we discuss artificial intelligence (AI), an academic discipline started in 1956 (Crevier, 1993). AI is gaining popularity in the context of propensity score estimation.

Artificial Intelligence. Today, artificial intelligence or "AI" is a trending topic. AI is "a system's ability to interpret external data correctly, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation" (Kaplan & Haenlein, 2019). AI can be helpful in propensity score estimation because it can identify underlying patterns between treatments and confounding variables using machine learning without being explicitly programmed.

Many classification algorithms in machine learning can outperform the classical methods for propensity score estimation, mainly when processing the data with many covariates, including neural networks (NN), linear classifiers, decision trees, particularly classification and regression tree (CART), and boosting (Glynn et al., 2006; McCaffrey, 2004; Setoguchi, 2008, Weitzen et al., 2010). For example, Breiman (2001) pointed out that NN performs well with small sample sizes (i.e., when seven or fewer observations exist per confounder).

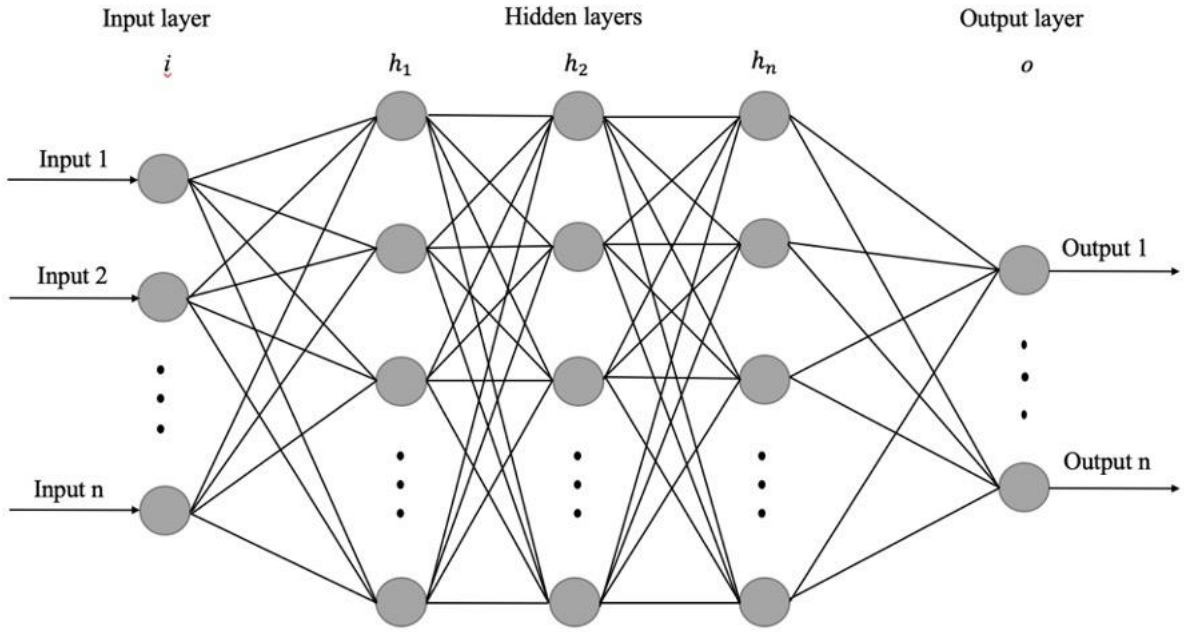
Artificial Neural Networks. McCulloch and Pitts (1943) introduced NN, a series of algorithms motivated by the formation of the nervous system. NN uses layers of nodes, and each node exchanges information similarly to neurons in the brain. Information is transferred based on calculations specified by the researcher. Various types of NN have been developed for different purposes. For example, Apple's "Hey Siri" uses NN to discover voice patterns. Nevertheless, all NN include one input layer (i), a user-specified number of "hidden" layers (h_n), and an output layer (o). All layers contain a number of nodes (also specified by the researcher) that connect to other nodes in the next layer by weights (Duda & Hart, 2006). Deep learning, a subset of machine learning, focuses solely on NN with multiple hidden layers. Figure 1 is an example of a deep learning NN because it includes multiple hidden layers.

The inner workings of a NN (for this study) can be divided into several steps: (1) the information is input to the input layer, which transfers the information to the hidden layer; (2) the interconnection between the two layers randomly assigns weights to each input; (3) a bias is added to each input after the weights are multiplied by each input separately; (4) the weighted sum is transferred to the activation function; (5) the activation function selects nodes for feature extraction; (6) the model applies an activation function to the final layer to provide the output; and, (7) the weight is adjusted, and the output is backpropagated to minimize the error (Chen, 1995; Chen et al., 2019; Zhang et al., 2018).

Training neural networks. Collier and his colleagues (2021) trained NN to estimate GPS for continuous treatments using data on food and nutrition by 1) splitting the data for training and testing (80% and 20% of the entire data set), 2) selecting hyperparameters (e.g., the number of nodes), 3) checking the training data set with the continuous treatment values while holding the treatment and covariates fixed, and 4) reweighting iteratively until the mean squared error was low.

Training NN improves the accuracy of propensity scores, which is critical before proceeding to

Figure 1. Neural Network Design



subsequent steps in any propensity score analyses. Too little training results in a misspecified propensity score model. Training too much will yield a well-performing NN on the training dataset but not on the test data (Brownlee, 2018; Zhang et al., 2019). The former model is an underfit model with high bias and low variance, while the latter is an overfit model with low bias and high variance. In both cases, the model is not generalized. A generalized model will detect patterns in similar data, such as new data from the same population of students. Researchers can improve an underfit model by training with more data, whereas addressing an overfitting model may be achieved by tailoring its hyperparameters and complexity (Lawrence & Giles, 2000; Srivastava et al., 2014; Tetko et al., 1995).

Computational Budget. Readers new to NN may consider the effort that goes into training computationally expensive. NN relies on several hyper-parameters with varying degrees of complexity; thus, these methods are onerous for educational researchers to apply to their unique research studies. Therefore, the means to automatically solve some problems in these design choices have been identified to reduce computational cost.

The research area that seeks to reduce the human effort in training is called "automated machine learning" (AutoML). There are many definitions of AutoML relative to its use and its benefits to educational researchers. For example, AutoML reduces the demand for intricate data science, enabling both educational content experts (often non-machine learning experts) to automatically build machine learning models without requiring too much machine learning knowledge (Zöller & Huber, 2021).

Several studies suggest that NN trained with AutoML has better results than NN trained by machine learning experts (Bergstra et al., 2011, 2013; Thornton et al., 2013). For example, Google AutoML (https://cloud.google.com/automl/) enables researchers to train high-end models with little effort and machine learning expertise. AutoML may be an effective strategy for reducing the computational resources needed to train NN in educational settings.

Covariate Balance. Covariate balance refers to the similarity of the distribution of covariates for different treatment exposure levels or groups (Austin, 2019). If covariate balance is not achieved, researchers should change the GPS model (e.g., include interactions) (Rubin, 2010).

Stratification is historically an approach to control the confounding from covariates. It is trendy for removing the confounding from a small number of covariates. As researchers add covariates, the number of strata increases and the sample size within the strata become scarce. Numerical methods and graphical visualizations can assess the balance of covariates when studied participants are stratified on the estimated GPS (Rai et al., 2018). The current study illustrates stratification, but readers who need to control for many covariates are encouraged to review the literature on alternative methods (McCaffrey et al., 2013).

Methodological investigations of alternative approaches to estimating propensity scores are on the rise (Ferri-García & Rueda, 2020). In the last decade, NN, particularly deep learning, gained popularity in various fields; however, no research applies deep learning models to estimate GPS for continuous multilevel treatments. To date, Collier and Leite (2021) is the only study to estimate GPS with NN for continuous treatments in the single-level setting. Still, they only focused on comparing different machine learning algorithms and did not assess covariate balance, computational time, or bias in average treatment effects (ATEs). Hence, this article adds to the existing literature by 1) estimating GPS with NN for continuous treatments in multilevel settings and 2) comparing GPS estimated with NN and traditional methods based on covariate balance, computational time, and bias in ATEs.

Methods

Monte Carlo Simualtion Study

A Monte Carlo simulation study allowed us to measure the performance of GPS estimation methods on across hypothetical scenarios relevant to educational settings. We generated the clustered data (e.g., students nested within schools) with three individual-level covariates ($X_1 - X_3$) and one group-level covariate W. The continuous treatment Z_{ij} mimicked student exposure to an online test-prep program using the linear regression model:

$$Z_{ij} = \beta_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \beta_3 X_{3ij} + \pi W_j + s_{0j} + s_{1j} X_{1ij} + s_{2j} X_{2ij} + s_{3j} X_{3ij} + r_{ij},$$

where Z_{ij} referred to the continuous treatment of the i^{th} individual in the j^{th} cluster, $\beta_1, \beta_2, \beta_3$ were the effects of level-1 (i.e., student level) covariates on treatment which were specified as .4, -.3, and .4 respectively. The individual level covariates ($X_1 - X_3$) were generated randomly from normal distributions. X_{1ij} and X_{2ij} were independent of cluster (cluster-level) membership, whereas X_{3ij} was generated depending on cluster membership. Also, $\pi = -.4$ was the regression coefficient of the cluster-level covariate on the continuous treatment which was defined as $\pi_j \sim N(0, \sigma_j^2)$. s_{0j} was the cluster intercepts drawn from a normal distribution. The random slopes of individual-level covariates s_{1j}, s_{2j}, s_{3j} were set to zero which assumed the same effects of individual-level covariates on treatment across clusters. The term r_{ij} represented the student-level residuals and were drawn from a logistic distribution with mean of zero and variance of $\pi^2/3$. For a half of our total iterations (i.e., 500), we omitted one individual level covariate in the GPS model to introduce omitted variable bias. Such bias occurs in most quantitative analyses in educational research and can bias estimated ATEs.

The continuous outcome was generated with the following model:

$$Y_{ij} = \gamma_0 + \gamma_1 Z_{ij} + \eta_1 X_{1ij} + \eta_2 X_{2ij} + \eta_3 X_{3ij} + \kappa W_j + u_{0j} + u_{1j} T_{ij} + \varepsilon_{ij},$$

where Y_{ij} were the continuous outcomes of the i^{th} individual in cluster j. η_1, η_2, η_3 are the fixed effects of three individual-level covariates on the outcomes which were set as 0.4, -0.3, and 0.4 respectively, and κ is the fixed effect of the cluster-level covariate which was set as -0.4. We simulated our effects of covariates on the outcomes based on the range of effects from existing propensity score simulation studies (e.g., Abdia et al., 2017; Collier et al., 2021). The intercept γ_0 was specified as 1 and γ_1 was 0.5, the effect of the treatment on the outcome. We selected a small effect of the treatment because educational interventions frequently have small effects (Kraft, 2018). The variances of both random intercept u_{0j} and the random slopes of treatment u_{1j} were set to 1 and the values of u_{0j} and u_{1j} were drawn from normal distributions. The variance of ε_{ij}, which represented

the within-cluster variance, was also drawn from normal distributions.

Also, we manipulated the sample sizes (n) by using different cluster sizes and different equally sized clusters of n_j: (J, n_j) = (50, 500), (50, 1000), (50, 1500); (100, 1000), (100, 2000), (100, 3000); (200, 2000), (200, 4000), and (200, 6000). We performed all simulations in R (R core, 2020). The entire experiment was performed on a Linux operating system. However, the packages we used to train models to estimate GPS were not operating system-dependent.

Machine Learning Training Procedures. We used 80% of each simulated dataset for training. We used the remaining 20% to test the accuracy of propensity score approaches with mean squared error (MSE) as a performance measure. According to Dobbin and Simon (2011), optimal splitting typically ranges from 40% to 80% of the full data. The MSE is

$$MSE = \frac{1}{N}\sum_{n=1}^{N}(\hat{R}_n - R_n)^2, \quad (4)$$

where $\hat{R}$ is a vector of N predicted GPS and R is a vector of actual GPS. Collier et al. (2021) also used MSE to train NN to estimate GPS for single-level continuous treatments.

Propensity Score Approaches. Five different models were used to estimate GPS: a single-level model (SL) with cluster-level confounders, a random intercept model (RI), a random slope model (RS), a single level neural network (NN) with cluster-level confounders, and a multilevel neural network (HLM.NN) model with cluster indicators.

We estimated regression-based models using the *lm* function and the *lme4* package in R (Bates et al., 2015). The deep learning models were implemented with the *automl* package (Boulangé, 2020). Unlike existing Monte Carlo simulations that required researchers to manually test the different combinations of NN model parameters, the automl tool that we used allowed for autotuning of hyperparameters (e.g., number of nodes, number of hidden layers) with the algorithm called metaheuristic PSO (Particle Swarm Optimization). This optimization algorithm started with a random set of hyperparameters (considered as a random particle in the space) and discovered the optimal solution for estimating GPS in the converging process. We used two functions in the *automl* package to do the automatic training (i.e., automl_train) and make predictions based on the trained model (i.e., automl_predict). The output predictions were the estimated GPS.

Propensity Score Stratification. After obtaining the GPS from each of the five models, we followed the steps proposed by Hirano and Imbens (2004) and Leite (2017) to evaluate these approaches. We stratified subjects into five equal-size mutually exclusive subsets based on the obtained GPS. Within each stratum where subjects were assigned roughly similar GPS, the effects of continuous treatment were estimated by fitting regression models.

Average Treatment Effect Estimation. Then, the overall ATEs were estimated by pooling the strata-specific estimates across five strata (Austin, 2011; Rosenbaum & Rubin, 1984). We modeled the outcomes as a function of the continuous treatment and estimated generalized propensity scores.

Analysis. Our study focused on three outcomes that are of interest to educational researchers using propensity score analyses: 1) percentage of covariates that achieved balance, 2) bias of ATEs, and 3) computational budget. We used descriptive statistics and graphs, split-plot ANOVA, and classification and regression trees (CART) to capture the differences in outcome variables under different manipulated conditions. The descriptive statistics and graphs helped us explain the bias and variance of ATEs. Generalized eta squared (GES) η^2 effect sizes from the split-plot ANOVAs were used to examine the contribution of manipulated factors where cluster size, number of clusters, and omitted variable bias were the between-dataset conditions, and five GPS estimation methods were the within-dataset factor. We fit the CART model using the *rpart* package (Therneau & Atkinson, 2019). CART helped us better capture interactions among manipulated conditions and visualized the simulation results.

Results

Evaluation of Covariate Balance

Our first outcome, the percentage of covariates that achieved balance, was based on covariates with an absolute value of standardized mean difference less than 0.1 (Leite, 2017). Each of our manipulated conditions had a significant effect on covariate balance, as expressed by η^2 shown in Table 1. The highest-

ranking order effect was the three-way interaction between the number of clusters, cluster size, and the GPS method. A regression tree in Figure 2 provides more insight into the interaction.

At the top, the root node shows that 94% of covariates achieved balance on average. The number below indicates the proportion of the simulations in this node (here at the top level, it is all simulations, 100%). Next, traveling down the tree branches to the following nodes, Figure 2 shows if the cluster size was 10, move left, and if 20 or 30, move right. If the cluster size was 10, only 88% of covariates achieved balance, while the other bucket shows on average 97% of covariates achieved balance.

Further down the left side of the regression tree, when HLM.NN, RS, and NN estimated GPS, 83% of the covariate achieved balance on average. On the other hand, when the other methods (SL and RI) were implemented, 96% of the covariates achieved balance on average. When the SL and RI models were implemented, and the number of clusters was 50, 88% of the covariates achieved balance on average. The final node on the right side, "the number of clusters = 50," shows that 100% of simulations achieved balance for other numbers of clusters (i.e., 100 and 200).

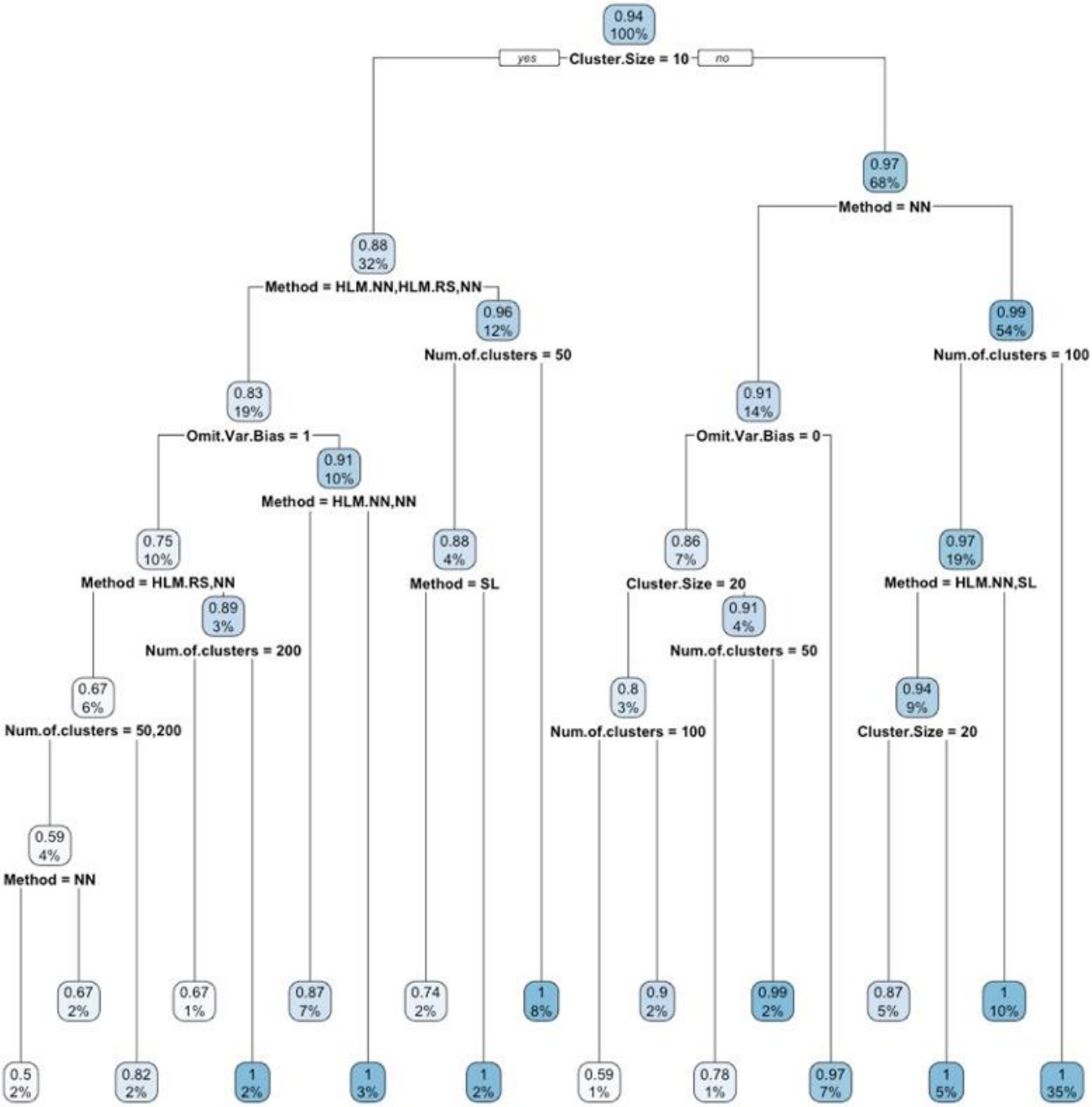
The right side of the regression tree shows that when cluster size was not ten and NN estimated GPS, 91% of covariates achieved balance. And when other methods estimated GPS and the number of clusters was not set at 100, all simulations performed 100% covariate balance. More complexity was shown when the number of clusters was 100. For example, when the number of clusters was 100 and GPS was estimated with HLM.NN and SL, with the cluster size of 20, 87% of the covariates achieved balance.

Table 1. Effects of Manipulated Conditions on Covariate Balance

Condition	GES
Num.of.clusters:Cluster.Size:Method	0.45
Cluster.Size	0.42
Method	0.41
Cluster.Size:Omit.Var.Bias:Method	0.34
Num.of.clusters:Cluster.Size:Omit.Var.Bias:Method	0.29
Num.of.clusters:Cluster.Size	0.28
Cluster.Size:Omit.Var.Bias	0.24
Cluster.Size:Method	0.24
Num.of.clusters:Omit.Var.Bias:Method	0.23
Omit.Var.Bias:Method	0.20
Num.of.clusters:Method	0.19
Num.of.clusters:Cluster.Size:Omit.Var.Bias	0.11
Num.of.clusters:Omit.Var.Bias	0.08
Omit.Var.Bias	0.03
Num.of.clusters	0.01

Note: GES denotes generalized eta squared, the effect sizes from split-plot ANOVAs. Red values highlight significant conditions.

Figure 2. CART Diagram for Effects of Manipulated Conditions on Covariate Balance



Bias of Average Treatment Effects

We kept the population's ATE equal to 0.5 across simulations. Figure 3 shows box plots of the ATEs across all conditions. The box plots indicate that ATE using GPS from the HLM.NN had the most minor variance and bias. Whereas the ATEs calculated from GPS estimated with SL and the NN performed most similar in terms of variance and bias. Outcome models based on the RI and RS models had the largest bias and variance.

Table 2 depicts the mean bias of ATEs. Across conditions, the RI and RS models yielded higher bias compared with deep learning models and the SL model. The HLM.NN performed very similarly for

both with and without omitted variable bias. All other methods, including NN, yielded comparably higher bias when covariates were omitted from the GPS model. To better understand the bias and variance of ATE in the presence of omitted bias, we ran three additional simulations and plotted the continuous treatment and outcomes and the regression line.

Plots in Figure 4 were created using three additional simulated datasets. Ideally, the linear regression line would fit mid-way through the actual data points. The lines estimated with GPS from the HLM.NN best fit the data. The plots align with the mean bias shown in Table 2 because the regression lines based on the NN and the SL models fit the data better than the RI and RS model, but not as well as the HLM.NN. Visually, there appears to be slight variance across the plots. Little variance across methods also aligns with box-plots in Figure 3, because the plots represent a single condition, and the box plots show all conditions. We would expect more variance across conditions rather than within a condition.

Figure 3. Box Plot of Average Treatment Effects

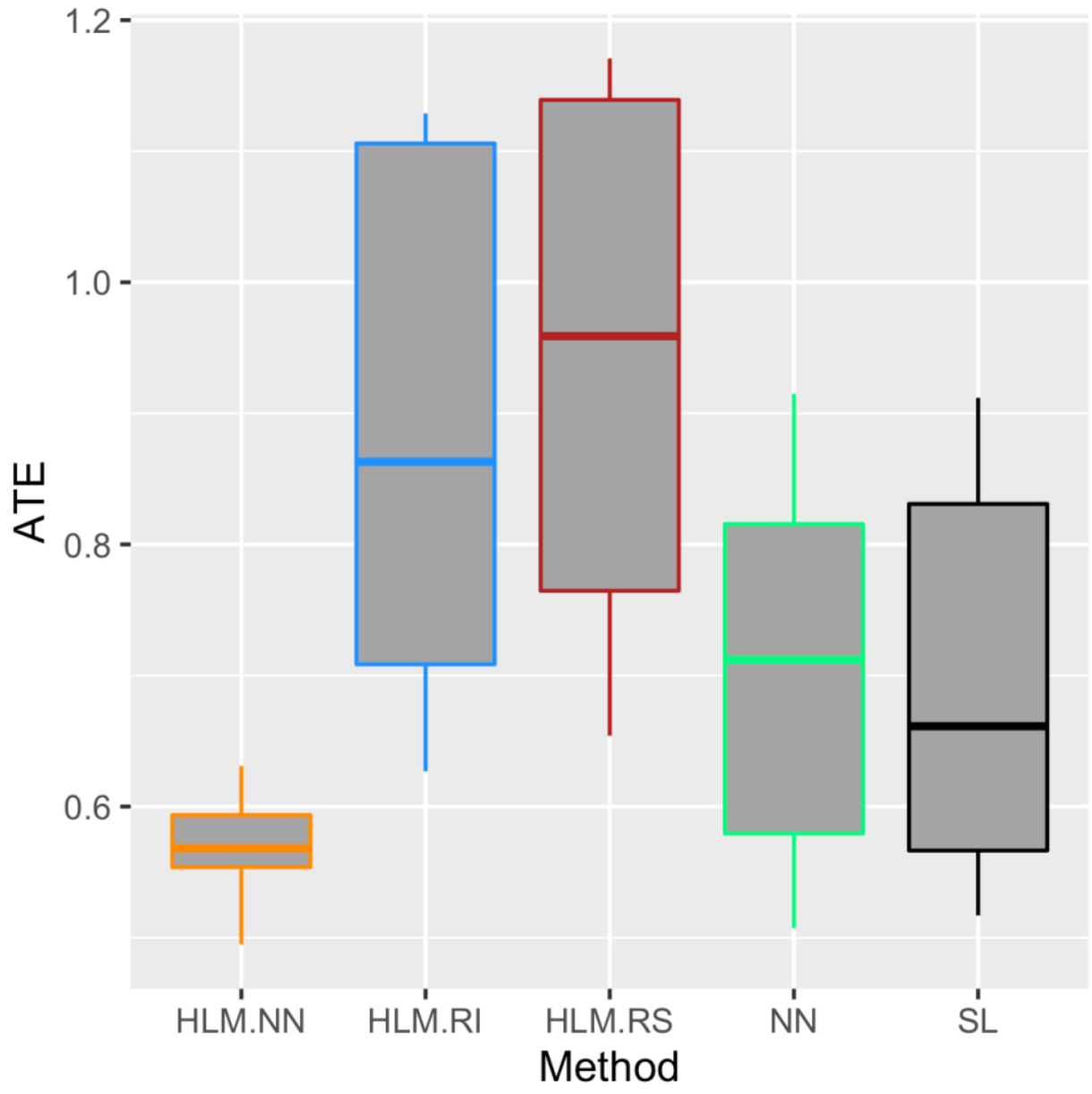


Table 2. Summary of Mean Bias of Average Treatment Effects

Omitted Variable Bias	Number of Clusters	Cluster Size	HLM.NN	NN	RI	RS	SL
0	50	10	0.04	0.06	0.2	0.27	0.05
	100		0.08	0.05	0.15	0.19	0.06
	200		0.09	0.09	0.2	0.26	0.08
	50	20	0.09	0.08	0.23	0.26	0.09
	100		0.06	0.12	0.21	0.24	0.08
	200		0.06	0.07	0.23	0.28	0.07
	50	30	0.09	0.09	0.24	0.27	0.06
	100		0.06	0.07	0.2	0.22	0.07
	200		0.07	0.08	0.2	0.26	0.07
1	50	10	0.01	0.27	0.48	0.59	0.22
	100		0.05	0.32	0.59	0.66	0.35
	200		0.09	0.32	0.56	0.6	0.34
	50	20	0.09	0.3	0.61	0.67	0.33
	100		0.06	0.4	0.63	0.66	0.4
	200		0.06	0.3	0.6	0.65	0.29
	50	30	0.1	0.36	0.62	0.64	0.33
	100		0.05	0.32	0.63	0.62	0.32
	200		0.07	0.31	0.61	0.63	0.31

Computational Budget

Table 3 shows the effects of manipulated conditions on the amount of time to train the GPS models. As expected, η^2 indicated substantial differences in training time based on the GPS method ($\eta^2 = .04$), the number of clusters ($\eta^2 = .01$), and cluster size ($\eta^2 = .01$). We also found significant effects of the following two-way interactions: 1) the number of clusters and GPS method, and 2) omitted variable bias and method. The following three-way interactions impacted the training time: 1) number of clusters, omitted variable bias, and GPS method; 2) the number of clusters, cluster size, and GPS method. To better explain these differences, we provided a table of the average training times (seconds) for each condition in Table 4.

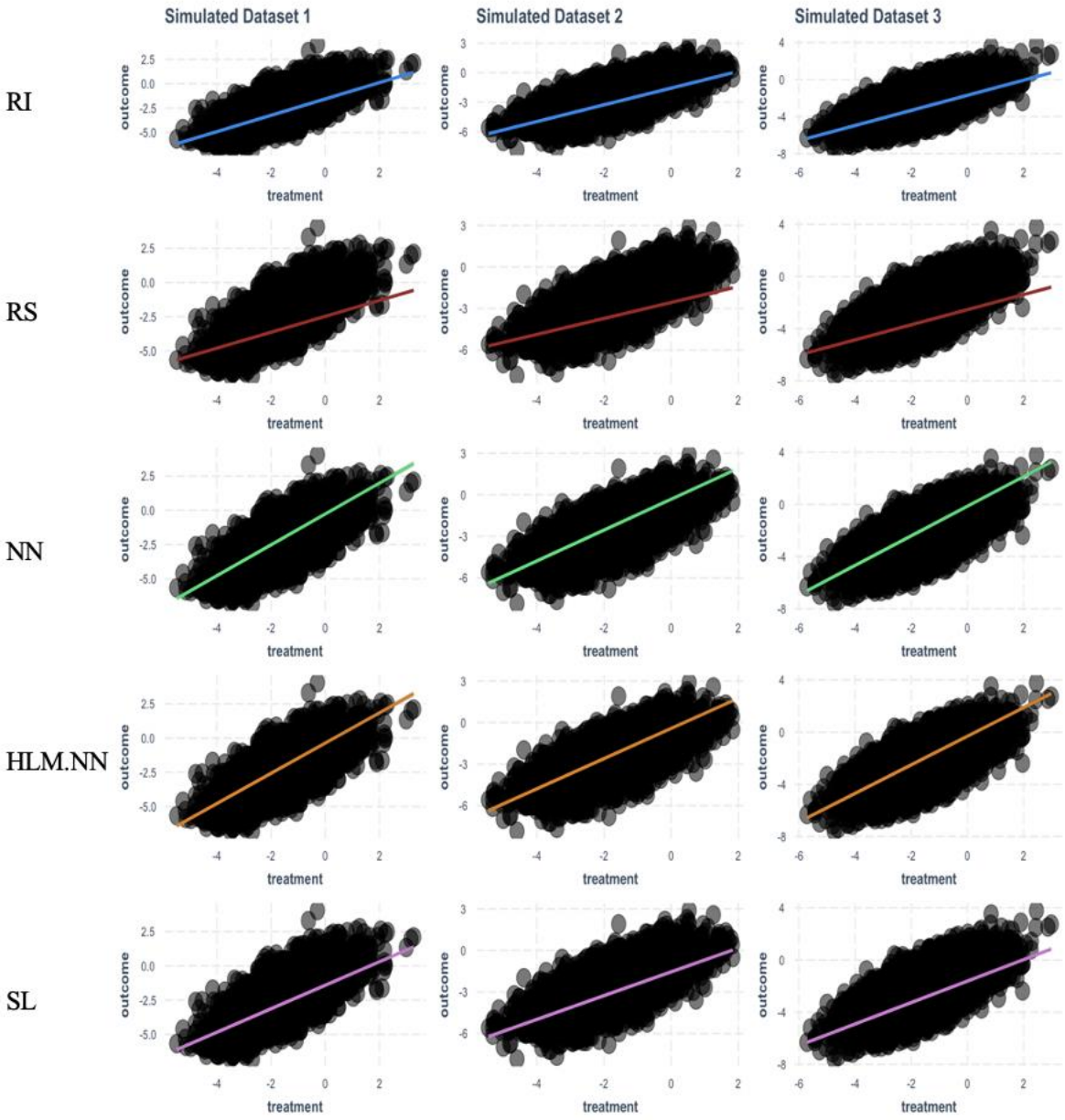
Table 4 shows how much longer it took to train the deep learning models than generalized linear models. There was also a two-way interaction between the number of clusters and the GPS method. In most cases, training time increases with increases in the number of clusters for deep learning techniques, but training time does not vary much in similar cases with other methods. Readers are encouraged to review Seger (2018) for more information on one-hot encoding and how it may improve computability with machine learning models. We did not one-hot encode our clusters, which may have reduced the efficiency of the deep learning techniques.

Average training times did not vary at all for the SL model, yet training time was considerably different at each cluster size with deep learning models. In most cases, training times were reduced when variables were omitted from the GPS models. And when variables were omitted and there were fewer clusters, the average training times were reduced for deep learning models.

Discussion

Educational research and evaluation rely on multilevel data, making accounting for both individual-level and cluster-level confounders vital for accurate findings. Austin (2011) explained how GPS methods offer a less parametric alternative to traditional

Figure 4. Regression Line Fit Across Simulated Data



regression adjustment when accounting for multilevel confounders. While applying propensity score techniques to educational data has become more frequent in recent years (Harris & Horst, 2016), the literature on GPS for continuous multilevel exposure is limited to regression-based estimation. Collier et al. (2021) first introduced deep learning as a more robust estimator to sample size and treatment distribution in a Monte Carlo simulation for single-level settings. The present study extends that methodological research to the multilevel setting and deepens the discussion around the practicality of machine learning in educational studies. For example, existing propensity score simulation studies (e.g., Chen, 2014; Leite et al., 2015) mainly focus on covariate balance and ATEs, but this study asks researchers to consider the computational budget and the overall efficiency of machine learning for propensity score estimation.

Table 3. Effects of Manipulated Conditions on Computational Budget

Condition	GES
Method	0.04
Num.of.clusters:Method	0.02
Cluster.Size:Method	0.01
Num.of.clusters:Omit.Var.Bias:Method	0.01
Num.of.clusters:Cluster.Size:Method	0.01
Num.of.clusters	0.01
Omit.Var.Bias:Method	0.01
Cluster.Size	0.01
Num.of.clusters:Cluster.Size:Omit.Var.Bias:Method	0.00
Omit.Var.Bias	0.00
Cluster.Size:Omit.Var.Bias:Method	0.00
Num.of.clusters:Cluster.Size:Omit.Var.Bias	0.00
Num.of.clusters:Cluster.Size	0.00
Cluster.Size:Omit.Var.Bias	0.00
Num.of.clusters:Omit.Var.Bias	0.00

Note: GES denotes generalized eta squared, the effect sizes from the split-plot ANOVAs. Red values highlight significant conditions.

Table 4. Computational Budget in Seconds for Deep Learning and Generalized Linear Models

Omit Bias	Num.of.clusters	Cluster.Size	HLM.NN	NN	RI	RS	SL
0	50	10	12.32	12.61	0.01	0.02	0.00
	100		25.14	26.19	0.02	0.02	0.00
	200		44.91	613.10	0.02	0.04	0.00
	50	20	236.20	25.43	0.02	0.03	0.00
	100		502.55	50.55	0.02	0.04	0.00
	200		85.75	580.89	0.03	0.06	0.00
	50	30	543.54	36.14	0.02	0.03	0.00
	100		223.11	268.03	0.03	0.05	0.00
	200		706.75	629.99	0.04	0.08	0.00
1	50	10	12.24	11.38	0.01	0.02	0.00
	100		26.21	24.72	0.02	0.02	0.00
	200		45.10	44.43	0.02	0.03	0.00
	50	20	26.31	26.02	0.02	0.02	0.00
	100		49.53	51.62	0.02	0.03	0.00
	200		316.58	314.01	0.03	0.06	0.00
	50	30	32.70	33.61	0.02	0.03	0.00
	100		72.85	77.02	0.02	0.04	0.00
	200		627.04	128.89	0.04	0.08	0.00

Note: Red values highlight the longest training times for each model.

Summary of Findings

Overall, the HLM.NN performed the best across all simulated conditions. The propensity scores estimated with HLM.NN consistently yielded the most accurate ATEs. To date, no previous studies investigated continuous multilevel settings with deep learning. However, our findings using HLM.NN were expected based on recent Monte Carlo simulations on single-level continuous treatments and proofs from seminal works (Collier et al., 2021; Rosenbaum & Rubin, 1983; 1984). Collier et al. (2021) found that the deep learning model correctly estimated GPS. And Rosenbaum and Rubin's (1983, 1984) proofs provide evidence that a correctly specified propensity score model will balance covariates and result in an unbiased estimate of the treatment effect.

Aligned with Shuler et al. (2016), cluster size and the number of clusters were significant predictors of covariate balance in our study. When the cluster size was 10, and the number of clusters was 50, the RI model averaged a 96% covariate balance. The single-level NN averaged 97% covariate balance for larger cluster sizes (>10). Covariate balance was least optimal using the single-level NN model when dealing with omitted variable bias, cluster size was 10, and the number of clusters was equal to 50 and 200.

Existing literature on cluster heterogeneity in propensity score estimation almost exclusively focuses on the omitted variable bias at the cluster level. In said cases, multilevel models (e.g., RI and RS models) can achieve a good balance (Arpino & Mealli, 2011; Fan, 2020: Li et al., 2013; Schuler et al., 2016). Instead, our simulation focused on omission at the individual level, a case where this robustness did not hold for conventional multilevel models. While models with random intercepts and slopes are not robust to the omission of individual-level confounders, our findings suggest that deep learning models provide more protection against omitted individual-level confounders. In addition, deep learning may be particularly advantageous when individual-level characteristics have not been measured or are not available to the researcher.

Implications for Practice in Education

Educational research and evaluation generally occur in a multilevel setting (Raudenbush & Schwartz, 2020). Our findings confirm earlier methodological results and reveal some new implications for educational researchers who conduct propensity score analysis with continuous treatments in multilevel settings. Therefore, we recommend the following based on this paper's results:

1. Sample GPS methods to see if they lead to similar covariate balance and ATEs.
2. Researchers should select variables with caution. Adding even a single confounder to a deep learning model can yield drastically different performance.
3. Deep learning may be helpful to confirm a theory about the treatment assignment. For example, better covariate balance using a NN may indicate an interaction in the treatment assignment (e.g., cross-level interaction).
4. NN architectures are not created equally. Researchers using the same data but different training methods (e.g., the 80/20 rule) and different hyperparameters can yield various performances.
5. Novice users should consider packages that automate the process of tuning hyperparameters, but it can take a while to run when they have a large number of covariates.

Recent criticism of machine learning for propensity score estimation presumes extreme time and computer memory requirements (Alam et al., 2019). The present Monte Carlo experiment provided evidence that training time for deep learning models far exceeds conventional approaches. However, more running time yielded better performance in terms of achieved covariate balance and less ATE bias thanks to automation. In addition, automation can save educational researchers time by not having to manually specify hyperparameters.

Using NN for propensity score estimation is available in many computer programming languages (i.e., Python, C++) and statistical packages, including R (Nagy, 2009) and SAS. For example, NeuroLab is an open-source NN library for Python, which contains training algorithms and a flexible framework to create and explore NN (García Roselló, 2003). Sknn package in Scikit-Neural Network library for Python can easily and quickly train deep neural networks for continuous

and categorical treatments/interventions (Maryasin & Lukashov, 2020). In R, the automated machine learning (automl) package is a quick tool to automate machine learning algorithms to real-world problems.

Limitations of the Study

Findings from this study provide applied researchers with an easy-to-use method for GPS estimation with deep learning NN and demonstrate how NN can achieve more accurate ATEs when individual-level cofounders are missing. However, this study is not exhaustive of all possible conditions when dealing with real-world data. In particular, we did not test how deep learning handles multiple types of missing data- such as omitted variable bias at the cluster level and data missing not at random (MNAR). Investigating missing data with deep learning is critical because analyses may result in biased ATEs if missing data are not appropriately addressed using propensity score methods (Malla et al., 2018).

Stratification potentially reduces bias due to the misspecification of treatment assignment. We found robustness to omitted variable bias (i.e., misspecification) using stratification with the HLM.NN. However, propensity score methods such as weighting and hybrid procedures are avenues for future research.

It could be indicated that we did not compare GPS estimation and outcome model combinations. The challenge is that NN does not have slopes like a regression. Since the slope is typically used to measure the ATE in propensity score analysis, research is needed to convert NN's weights into slopes before testing estimation and outcome model combinations.

Machine learning is still new to most educational researchers, but the procedures we demonstrated can reduce the anxiety around training and selecting hyperparameters. While it may take time for automated methods such as AutoML to train the deep learning model to estimate propensity scores, most applied researchers only work with one dataset. Therefore, applied researchers do not have to worry much about the computational burden of NN compared to methodologists running thousands of simulated datasets. Methodologists who wish to use NN in simulations and other users who simulate data for power analyses should consider methods like parallel computing to make the code run faster.

A recent review of graduate training in educational statistics and research methods programs in the U.S. shows that there was little to no mention of AI and machine learning methods (Randall et al., 2021). We believe this is why NN is understudied in educational research today. AI is a moving target that always seems advanced until people use it and get familiar with it. Hopefully, our paper moves the target further by keeping educational researchers up-to-date on the effectiveness of NN- a trending topic in propensity score analysis.

References

Alam, S., Moodie, E. E., & Stephens, D. A. (2019). Should a propensity score model be super? the utility of ensemble procedures for causal adjustment. *Statistics in Medicine, 38*(9), 1690-1702. https://doi.org/10.1002/sim.8075

Arpino, B., & Mealli, F. (2011). The specification of the propensity score in multilevel observational studies. *Computational Statistics & Data Analysis, 55*, 1770-1780. https://doi.org/10.1016/j.csda.2010.11.008

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research, 46*(3), 399-424. https://doi.org/10.1080/00273171.2011.568786

Austin, P. C. (2019). Assessing covariate balance when using the generalized propensity score with quantitative or continuous exposures. *Statistical Methods in Medical Research, 28*(5), 1365-1377. https://doi.org/10.1177/0962280218756159

Barnard, J., Frangakis, C. E., Hill, J. L., & Rubin, D. B. (2003). Principal stratification approach to broken randomized experiments: A case study of school choice vouchers in New York City. *Journal of the American Statistical Association*, *98*(462), 299-323. https://doi.org/10.1198/016214503000071

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*.

Bellara, A. P. (2013). Effectiveness of propensity score methods in a multilevel framework: A

Monte Carlo study. https://scholarcommons.usf.edu/etd/4635

Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011, December). Algorithms for hyper-parameter optimization. In *25th annual conference on neural information processing systems (NIPS 2011)* (Vol. 24). Neural Information Processing Systems Foundation.

Bergstra, J., Yamins, D., & Cox, D. D. (2013). Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms. In *Proceedings of the 12th Python in science conference* (Vol. 13, p. 20). Citeseer.

Boulange, A. (2020). *automl: Deep Learning with Metaheuristic*. URL: https://CRAN.R-project.org/package=automl r package version 1.3.2.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32.

Breiman, L., & Shang, N. (1996). Born again trees. *University of California, Berkeley, Berkeley, CA, Technical Report, 1*(2), 4.

Brownlee, J. (2018). *Better Deep Learning: Train Faster, Reduce Overfitting, and Make Better Predictions*. Machine Learning Mastery.

Chen, H. (1995). Machine learning for information retrieval: Neural networks, symbolic learning, and genetic algorithms. *Journal of the American Society for Information Science*, *46*(3), 194-216. https://doi.org/10.1002/(SICI)1097-4571(199504)46:3<194::AID-ASI4>3.0.CO;2-S

Chen, J. (2014). *A Bayesian propensity score approach for multilevel observational studies*. (Unpublished doctoral dissertation). University of Wisconsin Madison, Madison, WI.

Chen, M., Challita, U., Saad, W., Yin, C., & Debbah, M. (2019). Artificial neural networks-based machine learning for wireless networks: A tutorial. *IEEE Communications Surveys & Tutorials*, *21*(4), 3039-3071.

Collier, Z. K., & Leite, W. L. (2020). A Tutorial on Artificial Neural Networks in Propensity Score Analysis. *The Journal of Experimental Education*, 1-18. https://doi.org/10.1080/00220973.2020.1854158

Collier, Z. K., Leite, W. L., & Karpyn, A. (2021). Neural Networks to Estimate Generalized Propensity Scores for Continuous Treatment Doses. *Evaluation Review*. https://doi.org/10.1177/0193841X21992199

Collier, Z. K., Leite, W. L., & Zhang, H. (2021). Estimating propensity scores using neural networks and traditional methods: a comparative simulation study. *Communications in Statistics-Simulation and Computation*, 1-16.

Dobbin, K. K., & Simon, R. M. (2011). Optimally splitting cases for training and testing high dimensional classifiers. *BMC Medical Genomics*, *4*(1), 31. https://doi.org/10.1186/1755-8794-4-31

Duda, R. O., & Hart, P. E. (2006). *Pattern classification*. John Wiley & Sons.

Eckardt, P. (2012). Propensity score estimates in multilevel models for causal inference. *Nursing Research, 61*(3), 213-223. http://doi.org/10.1097/NNR.0b013e318253a1c4

Fan, M. (2020). *Performance of Parametric Vs. Data Mining Methods for Estimating Propensity Scores with Multilevel Data: A Monte Carlo Study* (Doctoral dissertation, University of Delaware).

Ferri-García, R., & Rueda, M. D. M. (2020). Propensity score adjustment using machine learning classification algorithms to control selection bias in online surveys. *PloS one*, *15*(4), e0231500. https://doi.org/10.1371/journal.pone.0231500

Friedman, J. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics, 29*(5), 1189-1232. http://www.jstor.org/stable/2699986

García Roselló, E., García Pérez-Schofield, J. B., González Dacosta, J., & Pérez-Cota, M. (2003). Neuro-Lab: A highly reusable software-based environment to teach artificial neural networks. *Computer Applications in Engineering Education, 11*(2), 93-102. https://doi.org/10.1002/cae.10042

Glynn, R. J., Schneeweiss, S., & Stürmer, T. (2006). Indications for propensity scores and review of their use in pharmacoepidemiology. *Basic & Clinical Pharmacology & Toxicology, 98*(3), 253-259. https://doi.org/10.1111/j.1742-7843.2006.pto_293.x

Harder, V. S., Morral, A. R., & Arkes, J. (2006). Marijuana use and depression among adults: Testing for causal associations. *Addiction*, *101*(10), 1463-1472. https://doi.org/10.1111/j.1360-0443.2006.01545.x

Harder, V. S., Stuart, E. A., & Anthony, J. C. (2008). Adolescent cannabis problems and young adult

depression: male-female stratified propensity score analyses. *American Journal of Epidemiology*, *168*(6), 592-601. https://doi.org/10.1093/aje/kwn184

Harder, V. S., Stuart, E. A., & Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods*, *15*(3), 234. https://doi.org/10.1037/a0019623

Harris, H., & Horst, S. J. (2016). A brief guide to decisions at each step of the propensity score matching process. *Practical Assessment, Research, and Evaluation*, *21*(1), 4. https://doi.org/10.7275/yq7r-4820

He, X., Zhao, K., & Chu, X. (2021). AutoML: A Survey of the State-of-the-Art. *Knowledge-Based Systems*, *212*, 106622. https://doi.org/10.1016/j.knosys.2020.106622

Hecht-Nielsen, R. (1992). Theory of the backpropagation neural network. In *Neural networks for perception* (pp. 65-93). Academic Press. https://doi.org/10.1016/B978-0-12-741252-8.50010-8

Hirano, K., & Imbens, G. W. (2004). The propensity score with continuous treatments. *Applied Bayesian modeling and causal inference from incomplete-data perspectives, 226164*, 73-84.

Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons, 62*(1), 15-25. https://doi.org/10.1016/j.bushor.2018.08.004

Keller, B., Kim, J., & Steiner, P. (2013). Abstract: Data mining alternatives to logistic regression for propensity score estimation: Neural networks and support vector machines. *Multivariate Behavioral Research, 48*(1), 164-164. https://doi.org/10.1080/00273171.2013.752263

Kim, J., & Seltzer, M. (2007). Causal Inference in Multilevel Settings in Which Selection Processes Vary across Schools. CSE Technical Report 708. *National Center for Research on Evaluation, Standards, and Student Testing (CRESST).*

Kraft, M. A. (2018). *Interpreting effect sizes of education interventions* (Brown University Working Papers). Providence. Retrieved from https://scholar.harvard.edu/files/mkraft/files/kraft_2018_interpreting_effect_s zes.pdf.

Lawrence, S., & Giles, C. L. (2000). Overfitting and neural networks: conjugate gradient and backpropagation. In *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium* (Vol. 1, pp. 114-119). IEEE. https://doi.org/10.1109/IJCNN.2000.857823

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436-444. https://doi.org/10.1038/nature14539

LeDell, E., & Poirier, S. (2020, July). H2o automl: Scalable automatic machine learning. In *Proceedings of the AutoML Workshop at ICML* (Vol. 2020).

Leite, W. (2016). *Practical propensity score methods using R.* Sage Publications.

Leite, W. L., Cetin-Berber, D. D., Huggins-Manley, A. C., Collier, Z. K., & Beal, C. R. (2019). The relationship between Algebra Nation usage and high-stakes test performance for struggling students. *Journal of Computer Assisted Learning, 35*(5), 569-581. https://doi.org/10.1111/jcal.12360

Leite, W. L., Jimenez, F., Kaya, Y., Stapleton, L. M., MacInnes, J. W., & Sandbach, R. (2015). An evaluation of weighting methods based on propensity scores to reduce selection bias in multilevel observational studies. *Multivariate Behavioral Research, 50*, 265-284.

Li, F., Zaslavsky, A. M., & Landrum, M. B. (2013). Propensity score weighting with multilevel data. *Statistics in Medicine*, *32*(19), 3373-3387. https://doi.org/10.1002/sim.5786

Malla, L., Perera-Salazar, R., McFadden, E., Ogero, M., Stepniewska, K., & English, M. (2018). Handling missing data in propensity score estimation in comparative effectiveness evaluations: a systematic review. *Journal of Comparative Effectiveness Research*, *7*(3), 271-279. https://doi.org/10.2217/cer-2017-0071

Maroco, J., Silva, D., Rodrigues, A., Guerreiro, M., Santana, I., & de Mendonça, A. (2011). Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. *BMC Research Notes, 4*(1), 1-14. https://doi.org/10.1186/1756-0500-4-299

Maryasin, O. Y., & Lukashov, A. I. (2020, September). A Python Application for Hourly Electricity Prices Forecasting Using Neural Networks. In *2020 International Russian Automation Conference (RusAutoCon)* (pp. 138-143). IEEE. https://doi.org/10.1109/RusAutoCon49822.2020.9208035

McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., & Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in Medicine*, *32*(19), 3388-3414. https://doi.org/10.1002/sim.5753

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity Score Estimation with Boosted Regression for Evaluating Causal Effects in Observational Studies. *Psychological Methods, 9*(4), 403–425. https://doi.org/10.1037/1082-989X.9.4.403

McCormick, M. P., O'Connor, E. E., Cappella, E., & McClowry, S. G. (2013). Teacher–child relationships and academic achievement: A multilevel propensity score model approach. Journal of school psychology, 51(5), 611-624.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, *5*(4), 115-133. https://doi.org/10.1007/BF02478259

Mitten, C., Collier, Z. K., & Leite, W. L. (2021). Online Resources for Mathematics: Exploring the Relationship between Teacher Use and Student Performance. *Investigations in Mathematics Learning*, 1-18. https://doi.org/10.1080/19477503.2021.1906041

Rai, S. N., Wu, X., Srivastava, D. K., Craycroft, J. A., Rai, J. P., Srivastava, S., ... & Baumgartner, R. (2018). Review: propensity score methods with application to the HELP clinic clinical study. *Open Access Medical Statistics, 8*, 11-23. https://doi.org/10.2147/OAMS.S156704

Randall, J., Rios, J. A., & Jung, H. J. (2021). Graduate training in educational measurement and psychometrics: A curriculum review of graduate programs in the US. *Practical Assessment, Research, and Evaluation*, *26*(1), 2.

Raudenbush, S. W., & Schwartz, D. (2020). Randomized experiments in education, with implications for multilevel causal inference. *Annual Review of Statistics and Its Application, 7,* 177-208. https://doi.org/10.1146/annurev-statistics-031219-041205

Reed, R., & MarksII, R. J. (1999). Neural smithing: supervised learning in feedforward artificial neural networks. MIT Press.

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*, 516-524. https://doi.org/10.1080/01621459.1984.10478078

Rubin, D. B. (2010). Propensity score methods. *American Journal of Ophthalmology*, *149*(1), 7-9. https://doi.org/10.1016/j.ajo.2009.08.024

Ruck, D. W., Rogers, S. K., Kabrisky, M., Oxley, M. E., & Suter, B. W. (1990). The multilayer perceptron as an approximation to a Bayes optimal discriminant function. *IEEE Transactions on Neural Networks*, *1*(4), 296-298. https://doi.org/ 10.1109/72.80266

Schuler, M. S., Chu, W., & Coffman, D. (2016). Propensity score weighting for a continuous exposure with multilevel data. *Health Services and Outcomes Research Methodology, 16*(4), 271-292. https://doi.org/10.1007/s10742-016-0157-5

Schumacher, M., Rossner, R., & Vach, W. (1996). Neural networks and logistic regression: Part 1. *Computational Statistics & Data Analysis, 21*(6), 661-682. https://doi.org/10.1016/0167-9473(95)00032-1

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. http://doi.org/10.1002/pds.1555

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. *arXiv preprint arXiv:1206.2944*.

Specht, D. F. (1991). A general regression neural network. *IEEE transactions on neural networks*, *2*(6), 568-576.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014).

Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*(1), 1929-1958.

Tenti, P. (2017). Forecasting foreign exchange rates using recurrent neural networks. In *Artificial Intelligence Applications on Wall Street* (pp. 567-580). Routledge.

Tetko, I. V., Livingstone, D. J., & Luik, A. I. (1995). Neural network studies. 1. Comparison of overfitting and overtraining. *Journal of Chemical Information and Computer Sciences*, *35*(5), 826-833. https://doi.org/10.1021/ci00027a006

Therneau, T. M., & Atkinson, E. J. (2018). *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-13. https://CRAN.R-project.org/package=rpart

Thoemmes, F. J. (2009). *The use of propensity scores with clustered data: A simulation study*. Arizona State University.

Thoemmes, F. J., & West, S. G. (2011). The use of propensity scores for nonrandomized designs with clustered data. *Multivariate Behavioral Research, 46*(3), 514-543.

Thornton, C., Hutter, F., Hoos, H. H., & Leyton-Brown, K. (2013). Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 847-855). https://doi.org/10.1145/2487575.2487629

Watkins, S., Jonsson-Funk, M., Brookhart, M. A., Rosenberg, S. A., O'Shea, T. M., & Daniels, J. (2013). An Empirical Comparison of Tree-Based Methods for Propensity Score Estimation. *Health Services Research*, *48*(5), 1798-1817. https://doi.org/10.1111/1475-6773.12068

Weitzen, S., Lapane, K. L., Toledano, A. Y., Hume, A. L., & Mor, V. (2004). Principles for modeling propensity scores in medical research: a systematic literature review. *Pharmacoepidemiology and Drug Safety, 13*(12), 841-853. https://doi.org/10.1002/pds.969

Westreich, D., Lessler, J., & Funk, M. J. (2010). Propensity score estimation: machine learning and classification methods as alternatives to logistic regression. *Journal of Clinical Epidemiology*, *63*(8), 826–833. http://doi.org/10.1016/j.jclinepi.2009.11.020

Wyss, R., Ellis, A. R., Brookhart, M. A., Girman, C. J., Jonsson Funk, M., LoCasale, R., & Stürmer, T. (2014). The role of prediction modeling in propensity score estimation: an evaluation of logistic regression, bCART, and the covariate-balancing propensity score. *American Journal of Epidemiology*, *180*(6), 645-655. https://doi.org/10.1093/aje/ kwu181

Xiang, Y., & Tarasawa, B. (2015). Propensity score stratification using multilevel models to examine charter school achievement effects. *Journal of School Choice, 9*(2), 179-196. https://doi.org/10.1080/15582159.2015.1028862

Yadav, A., & Sahu, K. (2017). Wind Forecasting using Artificial Neural Networks: A Survey and Taxonomy. *International Journal of Research in Science & Engineering*, *3*.

Yang, R., Carter, B. L., Gums, T. H., Gryzlak, B. M., Xu, Y., & Levy, B. T. (2017). Selection bias and subject refusal in a cluster-randomized controlled trial. *BMC Medical Research Methodology*, *17*(1), 1-10. https://doi.org/10.1186/s12874-017-0368-7

Yao, Q., Wang, M., Chen, Y., Dai, W., Li, Y. F., Tu, W. W., ... & Yu, Y. (2018). Taking human out of learning applications: A survey on automated machine learning. *arXiv preprint arXiv:1810.13306*.

Zhang, P., Shen, H., & Zhai, H. (2018). Machine learning topological invariants with neural networks. *Physical Review Letters*, *120*(6), https://doi.org/10.1103/PhysRevLett.120.066401

Zhang, H., Zhang, L., & Jiang, Y. (2019). Overfitting and underfitting analysis for deep learning based end-to-end communication systems. In *2019 11th International Conference on Wireless Communications and Signal Processing (WCSP)* (pp. 1-6). IEEE. http://doi.org/10.1109/WCSP.2019.8927876

Zhao, P., Su, X., Ge, T., & Fan, J. (2016). Propensity score and proximity matching using random forest. *Contemporary Clinical Trials*, *47*, 85-92. https://doi.org/10.1016/j.cct.2015.12.012

Zhu, Y., Coffman, D. L., & Ghosh, D. (2015). A boosting algorithm for estimating generalized propensity scores with continuous treatments. *Journal of Causal Inference*, *3*(1), 25-40. https://doi.org/ 10.1515/jci-2014-0022

Zöller, M. A., & Huber, M. F. (2021). Benchmark and survey of automated machine learning frameworks. *Journal of Artificial Intelligence Research, 70*, 409-472. https://doi.org/10.1613/jair.1.11854

Citation:
Collier, Z.K., Zhang, H., & Liu, L. (2022). Explained: Artificial Intelligence for Propensity Score Estimation in Multilevel Educational Settings. *Practical Assessment, Research & Evaluation*, 27(3). Available online: https://scholarworks.umass.edu/pare/vol27/iss1/3/

Corresponding Author:
Zachary, K. Collier
School of Education
University of Delaware

Email: collierz [at] udel.edu