Using Big Data to Sharpen Design-Based Inference in A/B Tests

Adam C Sales
University of Texas at Austin
536C George I. Sánchez Building
Austin, TX 78705
asales@utexas.edu

Anthony Botelho
Worcester Polytechnic
Institute 100 Institute Rd
Worcester, MA 01609
abotelho@wpi.edu

Thanaporn Patikorn
Worcester Polytechnic
Institute 100 Institute Rd
Worcester, MA 01609
tpatikorn@wpi.edu

Neil T. Heffernan
Worcester Polytechnic
Institute 100 Institute Rd
Worcester, MA 01609
nth@wpi.edu

ABSTRACT
Randomized A/B tests in educational software are not run in a vacuum: often, reams of historical data are available alongside the data from a randomized trial. This paper proposes a method to use this historical data–often high-dimensional and longitudinal–to improve causal estimates from A/B tests. The method proceeds in two steps: first, fit a machine learning model to the historical data predicting students' outcomes as a function of their covariates. Then, use that model to predict the outcomes of the randomized students in the A/B test. Finally, use design-based methods to estimate the treatment effect in the A/B test, using prediction errors in place of outcomes. This method retains all of the advantages of design-based inference, while, under certain conditions, yielding more precise estimators. This paper will give a theoretical condition under which the method improves statistical precision, and demonstrates it using a deep learning algorithm to help estimate effects in a set of experiments run inside ASSISTments.

1. INTRODUCTION
Randomized A/B tests hold a lot of promise for the study of student learning within intelligent tutors. Not only do they allow for causal inference without fear of confounding, but they also allow for "design-based" effect and standard error estimates that are virtually guaranteed to be unbiased [13]. These strengths are to the fact that data analysts know exactly how, and with what probability, conditions were assigned to subjects.

The traditional tools for analyzing experiments estimate effects using *only* data from the experiments themselves, discarding data from (potential) subjects that were not randomized. For instance, data from past users or from concurrent users who, for whatever reason, were not included in the A/B test, are excluded from the analysis. We refer to users such as these, with similar covariate and outcome data as the participants in the A/B test, but who were not randomized, as the "remnant." Excluding the remnant makes good statistical sense: after all, the probabilities of assignment are known only for participants in the experiment, not for the remnant. However, data from the remnant may be quite useful—in particular, the extra sample size could improve the statistical precision, i.e. reduce the standard errors, of experimental effect estimates. This is especially the case for experiments run within intelligent tutors or other big data environments. Vast amounts of log data, collected prior to the experiment, in conjunction with powerful machine-learning methods, could help sharpen causal estimates considerably.

This paper introduces a method for using the remnant in analyzing experiments, without sacrificing any of the benefits of experimentation or making additional modeling assumptions. The core of the method is residualization—predicting experimental subjects' outcomes using a model fit to the remnant, and then estimating effects using prediction residuals instead of the outcomes themselves. We call the method "remnant-based residualization," or "rebar." Rebar builds on methods suggested in [9], [7], and [1]. Rebar was first introduced in [12] as a method to reduce confounding bias in observational studies. Here we show that rebar can also reduce standard errors in randomized A/B tests, particularly in educational data mining contexts.

Most other methods incorporating machine learning into analysis of experiments, either to estimate average effects (e.g. [16]), to estimate subgroup effects (e.g. [3]), or to optimally allocate treatment (e.g. [11], [19]) use machine learning to replace, rather than complement, design based methods. This is a very promising avenue of research, but lacks the statistical guarantees of well-worn design-based estimates.

The next section will formally introduce rebar, and sections 3, 4, and 5 will illustrate it using a deep-learning model to sharpen effect estimates in a set of 22 experiments run within the ASSISTments system [14]. Section 6 will conclude.

2. REBAR

2.1 Experiments and Modeling

To learn if an intervention worked, or to figure out which of two conditions (say, condition 0 and condition 1) produces better outcomes, statistical models can often be quite helpful. To take a common example, analysts might regress an outcome Y on an indicator for condition Z, along with a vector of covariates $\boldsymbol{x}$. Then, the estimated coefficient on Z is taken as the estimated effect of condition 1 versus 0, controlling for $\boldsymbol{x}$ [18].

The shortcomings of this approach are well-known: if the vector $\boldsymbol{x}$ is missing a confounder—a covariate that predicts both subjects' choice of condition, 0 or 1, and outcomes Y—then the regression estimate will be biased. Moreover, even if there are no unmeasured confounders, if the regression model is misspecified, for instance, modeling the relationship between Y and $\boldsymbol{x}$ as linear, then the estimate will also be biased. On the other hand, a regression model may be run on all available data, producing precise (if inaccurate) estimates.

Randomized experiments correct regression's faults. If subjects are randomly assigned to conditions 0 or 1, then the difference in mean outcomes between the two groups is an unbiased estimate of the average treatment effect. More precisely, following [15] and [10], let y_{1i} be the outcome a subject i *would* experience *if* assigned to condition 1, and let y_{0i} be the outcome i would experience under condition 0. A subject's observed outcome $Y_i = y_{1i}$ if i is assigned to 1, $Z_i = 1$; $Y_i = y_{0i}$ if i is assigned to 0. (Since observed outcomes Y are a function of Z, they are random; we may model potential outcomes y_0 and y_1 as fixed.)

Under this framework, we define causal effects based on potential outcomes y_0 and y_1, rather than observed outcomes Y. An individual i's treatment effect τ_i is the difference of those two: $\tau_i \equiv y_{1i} - y_{0i}$—the difference between i's outcome under treatment versus under control. Without strong assumptions, these individual effects are not identified by the data; instead, we estimate quantities such as the average treatment effect (ATE) over all subjects $\bar{\tau} = \sum_i \tau_i / n$, or the average effect of the treatment on the treated (TOT) $\bar{\tau}_{Z=1} = \sum_i Z_i \tau_i / n_1$, where n and n_1 are the total number of subjects and the number of treated subjects, respectively. In a simple randomized experiment, the ATE and TOT have the expectation, but their estimators may have different standard errors. For the sake of simplicity, we will focus on the TOT.

Observed outcomes may be used to estimate the ATE, TOT, or other causal parameters. In particular, an unbiased estimator of the TOT is:

$$\hat{\tau} = \bar{Y}^{Z=1} - \bar{Y}^{Z=0}$$

where $\bar{Y}^{Z=1}$ is the mean of Y for treated subjects, $\sum_i Z_i Y_i / n_1$, and $\bar{Y}^{Z=0}$ is the mean of Y in the control group.

An unbiased estimator of the squared standard error is:

$$SE_{TOT}^2 = n/(n_1 n_0) s^2(\boldsymbol{Y})_{Z=0}$$

where $s^2(\boldsymbol{Y})_{Z=0}$ is the sample variance of Y in the control group. See the Technical Appendix, and [4] for more details. Estimators $\hat{\tau}$ and SE_{TOT}, and their properties, are derived solely from the experimental design, via survey sampling theory; they do not depend on the (unknown) distributions of y_1 and y_0, or any other modeling assumptions. They are "design-based."

In a randomized experiment there are no confounders. Since the probability distribution of Z is known exactly, no statistical models, or modeling assumptions, are necessary—the analysis may be "design-based" instead of model-based. In particular, the estimate $\hat{\tau}$ and its standard error derive from survey sampling theory, not the distribution of y_0 or y_1. On the other hand, any data from the "remnant" of an experiment—the set of subjects outside the experiment, who were not randomized to either condition—play no role in this analysis. Since subjects in the remnant were not randomized, there is no telling how they may differ from the $Z = 0$ or $Z = 1$ groups, in ways measured or unmeasured, and there is no telling (exactly, statistically) how their data came to be, so design-based analysis is impossible and any model fit to the remnant is most likely misspecified. However, though dropping the remnant from the analysis brings unbiasedness, it also brings a loss of precision—all that sample size, thrown away.

2.2 A Role for the Remnant

Assume the following setup: a set of users, "the experimental set" were randomized to either condition 0 or condition 1, and their outcomes $\boldsymbol{Y}$ were measured at the end of the experiment. Conditions 0 or 1 could be two different treatments, or control and treatment condition; we will refer to condition 0, as "control" and 1 as "treatment." The goal of the experiment is to estimate the TOT, $\bar{\tau}_{Z=1}$, the average effect in the treatment group. Some more subjects, the remnant, were not randomized; instead, they all received condition 0, the default (this isn't strictly necessary—the theory also works if they received condition 1, a mix of conditions, or something else altogether—but it makes things simpler). Outcomes Y were also measured for members of the remnant. Finally, a set of covariates $\boldsymbol{x}$, possibly high-dimensional, of mixed-types, and/or longitudinal, were measured for everyone, in the experimental set and in the remnant.

Experimental estimates typically drop the remnant, and pay the price of lower precision. Instead, we suggest training a machine-learning model on the remnant, and using it to "residualize" the data from the experimental set—that is, estimate effects using prediction residuals. We call this algorithm "remnant-based residualization" or "rebar." The process is as follows:

1. Using data from the remnant, train a model $\hat{y}_0(\cdot)$ to predict y_0 as a function of $\boldsymbol{x}$.

2. Validate $\hat{y}_0(\cdot)$ (using cross-validation or other techniques). if it performs well, proceed; otherwise return to step 1, choosing a different model.

3. Use $\hat{y}_0(\cdot)$ and covariates $\boldsymbol{x}$ in the experimental set to generate predicted outcomes $\hat{y}_0(\boldsymbol{x})$ and residuals, $e = Y - \hat{y}_0(\boldsymbol{x})$.

4. Estimate the TOT as a difference in mean residuals,

$$\hat{\tau}_{rebar} = \bar{e}_{Z=1} - \bar{e}_{Z=0}$$

with estimated standard error

$$SE_{rebar} = \sqrt{n/(n_1 n_0)} s(\boldsymbol{e})_{Z=0}$$

Where $s(\boldsymbol{e})_{Z=0}$ is the sample standard deviation of e in the control group.

Just like the traditional estimator $\hat{\tau}$, the rebar estimator $\hat{\tau}_{rebar}$ is design-based—its logical basis is the designed experiment, not a model. On the other hand, it harvests information from the remnant to improve upon $\hat{\tau}$.

Rebar works because the predictions $\hat{y}_0(\boldsymbol{x})$ were generated from an external sample—the remnant—and pre-treatment covariates $\boldsymbol{x}$. Subject i's prediction $\hat{y}_0(\boldsymbol{x}_i)$ will be the same whether i is assigned to 0 or to 1. Since there's no treatment effect on $\hat{y}_0(\boldsymbol{x})$, subtracting $\hat{y}_0(\boldsymbol{x})$ from Y only removes noise—not part of the treatment effect. When treatment is randomized, Z is independent of $\hat{y}_0(\boldsymbol{x})$, so, in expectation, the mean of $\hat{y}_0(\boldsymbol{x})$ will be equal across the two treatment groups. In fact, the rebar estimator can be re-written as $\hat{\tau}_{rebar} = \bar{Y}_{Z=1} - \bar{Y}_{Z=0} - (\overline{\hat{y}_0(\boldsymbol{x})}_{Z=1} - \overline{\hat{y}_0(\boldsymbol{x})}_{Z=0})$. The first term is $\hat{\tau}$, which is unbiased for the TOT. The second term is the difference in means of $\hat{y}_0(\boldsymbol{x})$, which is zero in expectation—therefore, $\hat{\tau}_{rebar}$ is unbiased. This property holds not just for the difference-in-means estimator—rebar can sharpen any treatment effect estimator that is linear in Y and unbiased.

Rebar's main tool is the model $\hat{y}_0(\cdot)$, which predicts y_0 as a function of $\boldsymbol{x}$. In EDM settings, the dimension of available covariates is often very large, and sample sizes are often large as well—machine learning algorithms make strong candidates for $\hat{y}_0(\cdot)$. $\hat{y}_0(\cdot)$ is not a statistical model *per se*, estimating the parameters of a probability distribution, but as a tool for prediction. It need not be correct in any sense, and its estimates need not be unbiased or consistent. Since $\hat{y}_0(\cdot)$ is fit on a separate sample from the experimental subjects, the process of fitting it—steps 1 and 2 above—do not affect standard errors, and model misspecification does not lead to bias.

On the other hand, for rebar to be more precise than the usual difference in means, $\hat{y}_0(\cdot)$ must be able to generate decent predictions of y_0 in the experimental set. This will be the case if $\hat{y}_0(\boldsymbol{x})$ is a good prediction of y_0—by residualizing, we subtract out the component of Y's variance that is predicted by $\hat{y}_0(\cdot)$. The variance of the rebar estimator is proportional to the difference between the mean-squared prediction error of $\hat{y}_0(\cdot)$, $MSE = ||\boldsymbol{y_0} - \hat{y}_0(\boldsymbol{x})||^2/n$ and its squared bias. (Recall that both $\hat{\tau}$ and $\hat{\tau}_{rebar}$ are unbiased estimates of the TOT; the bias here refers to $\hat{y}_0(\cdot)$'s predictions of y_0, not to treatment effect estimates.) The extent to which it outperforms the standard estimate $\hat{\tau}$, measured as percent improvement $(SE_{TOT}^2 - SE_{rebar}^2)/SE_{TOT}^2$, is at least as large as $\hat{y}_0(\cdot)$'s prediction R^2 in the control group, $R^2 = 1 - ||Y - \hat{y}_0(\boldsymbol{x})||_{Z=0}^2/||Y - \bar{Y}||_{Z=0}^2$ (see the Technical

Appendix for derivations). If $\hat{y}_0(\cdot)$ performs poorly in the control group—so that $||Y - \hat{y}_0(\boldsymbol{x})|| > ||Y - \bar{Y}||$—then this R^2, as we have defined it, could be negative, and $\hat{\tau}_{rebar}$ will be less precise than $\hat{\tau}$; however, it will still be unbiased. The improvement $\hat{\tau}_{rebar}$ offers rests entirely on the performance of $\hat{y}_0(\cdot)$. The better we can predict how subjects would have performed in the control condition, the more precisely we can estimate treatment effects.

Since $\hat{y}_0(\cdot)$ is trained in the remnant, its performance in the experimental set (as measured by, e.g. prediction R^2) will be hard to gauge at the outset. If the distribution of Y, conditional on $\boldsymbol{x}$, differs widely from the between the two sets, $\hat{y}_0(\cdot)$'s performance may suffer in extrapolation. This problem is not fatal: the rebar estimate is unbiased regardless of $\hat{y}_0(\cdot)$'s properties. However, a model with poor predictive power in the experimental set will not reduce standard errors substantially, and may increase them. Of course an analyst may calculate $\hat{y}_0(\cdot)$'s R^2 in the experimental set, but choosing a model based on Y induces dependence between Y and $\hat{y}_0(\boldsymbol{x})$, and may cause bias. Models trained on a subset of the remnant that resembles the experimental set—or which weight such a remnant more heavily—may perform better than those trained on the entire remnant.

The previous discussion assumed simple randomization. However, rebar easily extends to more complex designs, including experiments with more than two treatment conditions. Further, as we will illustrate below, rebar can be extended to regression estimators of causal effects as well, modeling low-dimensional covariates within sample and high-dimensional covariates out of sample.

3. DATA: 22 EXPERIMENTS AND MORE

The 22 experiment dataset is a feature-rich dataset on students who participated in randomized controlled trials (RCTs) ran inside a free, online tutoring called ASSISTments [14]. This dataset consists of student-level data from 8,205 unique students participating in 22 A/B tests, 14,947 unique student-RCT pairs in total.

These RCTs were run within skill builders. Inside ASSISTments, a skill builder is a type of problem set that requires students to practice solving problems until they master the associated skill. Skill mastery is determined by the student's ability to answer a certain number of problems correctly, usually three, in a row.

This feature-rich dataset includes 30 features, including both categorical features, such as student grade levels, and numerical features, such as student performances prior to the experiment. This dataset also includes two dependent measures. The first dependent measure, "completion," is whether the student completed the assignment and achieved mastery. The other dependent measure is the number of problems attempted; for students who achieved mastery, this may be interpreted as mastery speed. The analysis in this paper will focus on the first dependent measure, completion.

4. DEEP LEARNING TO PREDICT COMPLETION

As described in Section 2.1, the model $\hat{y}_0(\cdot)$ is an integral part of the rebar methodology with the purpose of producing predicted outcome y_0 as a function of $\boldsymbol{x}$. The methodology does not rely on a specific type of model, nor any specific algorithm to be used so long as an estimate for the outcome variable of interest is generated by the model from included covariates. As also stated in that section, the accuracy of the model, however poor, does not lead to bias. That said, models that are more accurate at estimating the outcome variable of interest will likely lead to better estimates of treatment effects. Deep learning models have been previously applied in educational contexts with promising results, often reporting higher performance over existing methods [8][6][2]. While the application of such methods is not appropriate to all problem applications due to the size and complexity, the use of such models in this work is justified due to 1) the scale of data available for model training and 2) the inconsequence of producing an uninterpretable model (e.g. the significance and coefficients of individual variables in the model are not intended for study or knowledge discovery). What is needed, again, is simply a prediction model.

We develop and apply a type of deep learning model known as a long short term memory (LSTM) [5] network. This model is a variant of a recurrent neural network (RNN) [17] that is commonly applied to time series data to model temporal relationships within the sequences. The model produces its estimates for each time step by utilizing both covariates provided corresponding to the current time step as well as information from all previous time steps within the series. As such, the model is developed as a sequence-to-sequence method that observes a sequence of student data as input and produces a sequence of outcome estimates of equal length. The model structure utilizes a 3-layer design, with an input layer feeding into a recurrent hidden layer (represented as a layer also connected to itself in previous time steps), before then proceeding to an output layer.

Two separate datasets are used to train and apply the model. The application dataset, comprised of student data from the 22 experiment dataset combined with assignment-level information for all work each student started before beginning the respective experiment. In an attempt to reduce the complexity of the data from which the model must learn, the sequence length of student assignment history is limited such that no more than 10 prior assignments are included for each student. In other words, students who were in a single experiment have a sequence length of 10, with the last time step representing the most recent assignment prior to beginning the experiment. Conversely, students in multiple students may exhibit sequences longer than 10 if participation in the experiments was separated by fewer than 10 assignments. The dataset is comprised of data from 8,297 distinct students and a total of 130,935 student assignments.

The second dataset, used to train and validate the model, is comprised of student data exclusive to that comprising the 22 experiment dataset. Student data, again non-inclusive of students within the 22 experiment dataset, is collected from the non-experimental problem sets found in the application dataset. From these, assignments are randomly sampled, with which the dataset is constructed using the 10 most recent assignments before students begin the sampled

assignment. This step, helps to ensure a similar structure of the dataset to that of the application set. Again, for purposes of validity, it is important to stress that no students are found in both the training and application datasets. The dataset contains data from 134,141 distinct students and a total of 686,590 student assignments.

The model uses just 4 assignment-level covariates per time step to predict assignment-level performance on the subsequent assignment. These covariates include the simple measures of completion of the assignment, the number of problems attempted, the number of problems completed, and a measure of inverse mastery speed; this last measure is a transformation of mastery speed, using 1 divided by the number of problems when the assignment was complete, or 0 when the assignment was not completed. While simple in the number of covariates, again, the model also uses information from previous time steps, adding to its complexity (i.e. time step 2 is informed by time step 1, time step 3 is informed by time steps 2 and 1, etc.). The model produces two values per time step corresponding with the desired outcome variable of completion of the next assignment, and also an estimate of inverse mastery speed on the next assignment, using a combined cost of these two measures to update model parameters during training; this second measure was included as it is believed the model may better learn from the data by observing a continuous variable in addition to the binary value of completion and also acts as an example as to how future works may utilize the same methodology to observe beyond the measure of completion presented in this work.

The model is first evaluated using a 5-fold student-level cross-validation. The model is trained for multiple epochs, or training cycles through the data, using a 30% holdout set, sampled from the training set of each fold, to determine the stopping point of model training; this holdout set also helps stop the model training process before overfitting is detected. It is found that the model produces average AUC of 0.81 and an RMSE of 0.34 for next assignment completion over the 5 folds. Once completed, a final model is trained over the entirety of the training dataset and applied to the application dataset, which has acted as a holdout set during the training and validation process. The next assignment completion estimates, collected from the most recent assignment before students begin each experiment, is then used as the estimated value of completion that is used in subsequent steps of the rebar analyses.

5. RESULTS
We estimated treatment effects of interventions on skill-builder completion for the 22 experiments using both raw outcomes $\boldsymbol{Y}$, the usual approach, and using $\boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{y}}_0(\boldsymbol{x})$, the rebar estimator. We also estimated standard errors in both cases. We used difference in means estimators, so the effect estimates are in units of percentage points—how much more likely were students to finish skill builders under the treatment condition than under control.

Figure 1 shows improvement in precision of the rebar estimator over the usual estimator: the difference of the two standard errors, divided by the usual standard error, $(SE_{usual} - SE_{rebar})/SE_{usual}$. The x-axis shows the predic-
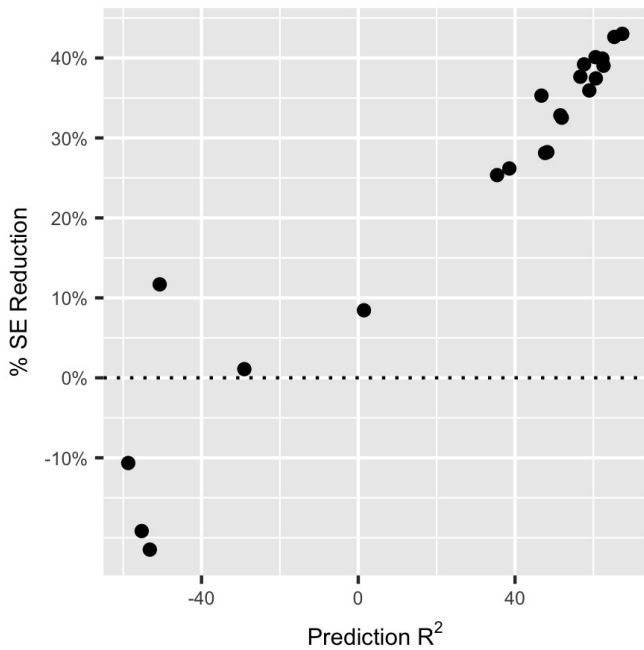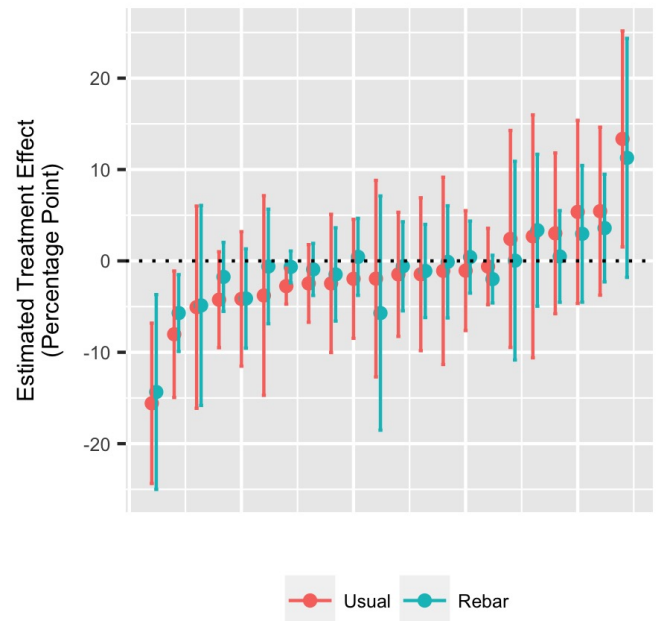
Figure 1: The improvement in precision of effect estimates as a percentage of the usual precision estimate, $[SE(\hat{\tau})-SE(\hat{\tau}_{rebar})]/SE(\hat{\tau})$, plotted as a function of $\hat{y}_0(\cdot)$'s prediction R^2 in the experimental set.

Figure 2: Effect estimates and 95% confidence intervals for the 22 experiments, using both the usual and rebar estimates. Experiments are ordered by their estimated effect.

tion R^2 of the deep learning model when extrapolated to each of the 22 experimental datasets. In 19 of the datasets, the rebar standard errors were lower than the usual standard errors. In 15 of those datasets, there was a greater than 25% improvement, and in four datasets the improvement was greater than 40%. The extent of the reduction in standard error corresponded closely to the prediction R^2 of $\hat{y}_0(\cdot)$, with the most dramatic improvements occurring when $R^2 \gtrsim 0.5$.

Figure 2 shows the estimated treatment effects and approximate 95% confidence intervals (two standard errors in each direction) for the two sets of estimators. In all but three cases, the rebar estimate was slightly closer to zero than the usual estimate. This is what we would expect if most of the true effects were null, so that reducing the noise of the treatment effect estimates would draw them closer to their true values. For that reason, although rebar reduced the standard errors in almost all of the experiments, it did not cause any of the non-significant results to become statistically significant. In fact, in two cases it had the opposite effect; though this may be disappointing for researchers, it is probably more accurate.

We also used linear regression to estimate treatment effects, regressing either indicators for completion or prediction errors on indicators for treatment assignment and two covariates: the proportions of students' prior skill builders completed and the proportions of prior skill builder problems students worked that they answered correctly. The results, available upon request, are nearly identical. Although the two covariates improved precision slightly, rebar continued

to dominate the usual estimate.

6. DISCUSSION
The rich, high-dimensional, fine-grained data that educational technology makes available should be a boon to causal inference. However, big data is subject to the same maladies as small data—confounding from unmeasured variables, and model misspecification. Classical randomized experiments remain relevant.

The same may not be true for classical design based estimators. Big data may not be able to correct unmeasured confounding and may exacerbate model misspecification, but we have shown that it can play a significant role reducing the standard errors of treatment effect estimates. The method we proposed here retains all of the statistical properties that recommend design-based estimators, while, in most cases, delivering substantially lower standard errors. We demonstrated the method's effectiveness using a cutting-edge deep learning algorithm trained to log data from ASSISTments which yielded impressive gains in precision when used to analyze a set of 22 experiments.

Rebar's most important tool in this exercise was the deep learning algorithm, which in 17 of the 22 experiments predicted completion better than the within-sample proportion. In general, designing prediction algorithms that perform well in the target dataset is the central challenge to effectively implementing rebar. Along the same lines, the most important open question is how to design diagnostics for prediction performance that do not rely on "peeking" at the experimental outcomes. One such diagnostic, termed "proximal

validation," was described in [12]—extending it to experimental studies and showing that it works is the next step in developing this method.

Wedding classical randomization-based causal inference with modern machine learning and big data can yield unbiased, robust, precise treatment effect estimates in technology-based educational datasets.

Technical Appendix

This discussion roughly follows [4], Section 1.1. The TOT $\sum_i Z_i(y_{1i} - y_{0i})/n_1$ may be re-written as $n_1^{-1}\left(\sum_i Y_i - \sum_i y_{0i}\right)$, the difference between the total of Y across both treatment groups, and the total that would have been observed had everyone received the control condition. The first sum is known exactly, but the second must be estimated using data from the control group. From elementary survey sampling, $n\bar{Y}^{Z=0}$ is unbiased for $\sum_i y_{0i}$. Further, the standard deviation of $n\bar{Y}^{Z=0}$ is $\sqrt{n^2(1 - n_0/n)s^2(\boldsymbol{y_0})/n_0} = \sqrt{n/(n_1n_0)}s^2(\boldsymbol{y_0})$, where $s^2(\boldsymbol{y_0})$ is the sample variance, over the whole sample, of y_0, and $1 - n_0/n = n_1/n$ is the finite population correction. Finally, due to random sampling, $s^2(\boldsymbol{Y})_{Z=0}$ is an unbiased estimator for $s^2(\boldsymbol{y_0})$. Substituting $\sigma_{Z=0}$ for σ_0 and dividing by n_1 gives the expression for SE_{TOT}.

Each individual treatment effect τ_i is the same whether the outcome (dependent variable) is Y or e:

$$e_1 - e_0 = (y_1 - \hat{y}_0(\boldsymbol{x})) - (y_1 - \hat{y}_0(\boldsymbol{x})) = y_1 - y_0$$

since $\hat{y}_0(\boldsymbol{x})$ is invariant to treatment. Therefore, the theory supporting standard estimates $\hat{\tau}$ and SE_{TOT} applies equally to $\hat{\tau}_{rebar}$ and SE_{rebar}. In particular, $\hat{\tau}_{rebar}$ is unbiased for the TOT with consistent standard error estimate SE_{rebar}, due to survey sampling theory.

The sample variance of e in the control group is

$$s^2(e)_{Z=0} = \frac{||\boldsymbol{e} - \bar{e}||^2_{Z=0}}{n_0 - 1}$$
$$= \frac{||\boldsymbol{Y} - \hat{\boldsymbol{y}}_0(\boldsymbol{x}) - (\bar{Y} - \overline{\hat{y}_0(\boldsymbol{x})})||^2_{Z=0}}{n_0 - 1}$$
$$= \frac{||\boldsymbol{Y} - \hat{\boldsymbol{y}}_0(\boldsymbol{x})||^2_{Z=0}}{n_0 - 1} - \frac{n_0}{n_0 - 1}\left(\bar{Y}_{Z=0} - \overline{\hat{y}_0(\boldsymbol{x})}_{Z=0}\right)^2$$

or the MSE of $\hat{y}_0(\cdot)$ in the control group, minus its squared bias. Since the squared bias is always positive,

$$s^2(e)_{Z=0} \leq \frac{||\boldsymbol{Y} - \hat{\boldsymbol{y}}_0(\boldsymbol{x})||^2_{Z=0}}{n_0 - 1}$$

Therefore, the ratio of the estimated rebar standard error to the usual TOT standard error is:

$$\left(\frac{SE_{rebar}}{SE_{TOT}}\right)^2 = \frac{s^2(e)_{Z=0}}{s^2(Y)_{Z=0}} \leq \frac{||\boldsymbol{Y} - \hat{\boldsymbol{y}}_0(\boldsymbol{x})||^2_{Z=0}}{||\boldsymbol{Y} - \bar{Y}||^2_{Z=0}} = 1 - R^2_{Z=0}$$

with equality if $\hat{y}_0(\cdot)$ is unbiased.

7. ACKNOWLEDGEMENTS

8. REFERENCES

[1] P. M. Aronow and J. A. Middleton. A class of unbiased estimators of the average treatment effect in randomized experiments. *Journal of Causal Inference*, 1(1):135–154, 2013.

[2] A. F. Botelho, R. S. Baker, and N. T. Heffernan. Improving sensor-free affect detection using deep learning. In *International Conference on Artificial Intelligence in Education*, pages 40–51. Springer, 2017.

[3] W. Duivesteijn, T. Farzami, T. Putman, E. Peer, H. J. Weerts, J. N. Adegeest, G. Foks, and M. Pechenizkiy. Have it both waysâĂŤfrom a/b testing to a&b testing with exceptional model mining. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 114–126. Springer, 2017.

[4] B. B. Hansen and J. Bowers. Attributing effects to a cluster-randomized get-out-the-vote campaign. *Journal of the American Statistical Association*, 104(487):873–885, 2009.

[5] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[6] M. Khajah, R. V. Lindsey, and M. C. Mozer. How deep is knowledge tracing? *arXiv preprint arXiv:1604.02416*, 2016.

[7] T. Patikorn, D. Selent, N. T. Heffernan, J. E. Beck, and J. Zou. Using a single model trained across multiple experiments to improve the detection of treatment effects. In *Proceedings of the 10th International Conference of Educational Data Mining*, 2017.

[8] C. Piech, J. Bassen, J. Huang, S. Ganguli, M. Sahami, L. J. Guibas, and J. Sohl-Dickstein. Deep knowledge tracing. In *Advances in Neural Information Processing Systems*, pages 505–513, 2015.

[9] P. R. Rosenbaum et al. Covariance adjustment in randomized experiments and observational studies. *Statistical Science*, 17(3):286–327, 2002.

[10] D. B. Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, 6(1):34–58, 1978.

[11] P. Rzepakowski and S. Jaroszewicz. Decision trees for uplift modeling with single and multiple treatments. *Knowledge and Information Systems*, 32(2):303–327, 2012.

[12] A. C. Sales, B. B. Hansen, and B. Rowan. Rebar: Reinforcing a matching estimator with predictions from high-dimensional covariates. *Journal of Educational and Behavioral Statistics*, 43(1):3–31, 2018.

[13] P. Z. Schochet. Statistical theory for the" rct-yes" software: Design-based causal inference for rcts. ncee 2015-4011. *National Center for Education Evaluation and Regional Assistance*, 2015.

[14] D. Selent, T. Patikorn, and N. Heffernan. Assistments dataset from multiple randomized controlled experiments. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale*, pages 181–184. ACM, 2016.

[15] J. Splawa-Neyman. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Roczniki Nauk Rolniczych*, 10:1–51, 1923. Translated and edited by D. M. Dabrowska and T. P. Speed for *Statistical Science*, 5(4):456–72, 1990.

[16] M. J. Van der Laan and S. Rose. *Targeted learning: causal inference for observational and experimental data.* Springer Science & Business Media, 2011.

[17] R. J. Williams and D. Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280, 1989.

[18] G. U. Yule. An investigation into the causes of changes in pauperism in england, chiefly during the last two intercensal decades (part i.). *Journal of the Royal Statistical Society*, 62(2):249–295, 1899.

[19] S. Zhao and N. Heffernan. Estimating individual treatment effect from educational studies with residual counterfactual networks. In *Proceedings of the 10th International Conference on Educational Data Mining, EDM*, 2017.