Efficacy Analysis Methodology:

Zearn's approach to Coarsened Exact Matching

September 2022

Alisa Szatrowski, Ph.D.
Maria Riolo, Ph.D.
Zearn

Introduction

Zearn is the 501(c)(3) nonprofit educational organization behind Zearn Math, a top-rated math learning platform used by 1 in 4 elementary-school students and by more than 1 million middle-school students nationwide. Zearn Math has users in all 50 states including in some of the largest districts in the country.

Zearn regularly partners with districts and states to conduct efficacy analyses. Each efficacy study aims to isolate the impact of Zearn Math on student achievement with observational data. Zearn is committed to ensuring effectiveness for all students. Thus, these studies examine Zearn Math's impact on student achievement overall, and by subgroup, disaggregated for prior achievement levels and demographic factors.

In the absence of experimental research conditions, Zearn uses quasi-experimental matching methods, designed to facilitate causal inference. These methods approximate experimental design by creating a comparison group of low- or non-users that is as similar as possible, on observable factors, to the group of students that consistently uses Zearn Math.

Matching methodology

Drawing causal inference from observational data is challenging because factors that impact a person's likelihood to receive an intervention may also impact their outcomes. Therefore the differences in outcomes observed between individuals may not be caused by the intervention itself, but by other confounding covariates that imbalance the treatment and control groups.

Matching methods attempt to balance the composition of confounding covariates between individuals who received an intervention and a comparison group of individuals who did not receive the intervention. This is done to isolate the difference in outcomes from the intervention itself, separate from any impact due to potentially confounding covariates (Stuart, 2008; Iacus et al., 2011). The effectiveness of matching is conditional on the ability of observable factors to capture the selection process that sorted individuals into treatment and control groups. Models that do not capture major covariates may produce biased estimates.

Because an individual cannot both receive and not receive the intervention during the same year, matching methods approximate this scenario by finding a separate individual who is as similar as possible to the individual that received the intervention (Ling et al., 2020).

Historically, propensity score matching (PSM) has been the dominant matching method in

quasi-experimental evaluation (Pearl, 2010; King & Nielsen, 2019).[1] PSM approaches this challenge by modeling the probability of placement in the treatment vs. control group based on the composition of an individual's covariate values.[2] The propensity score is an amalgamated probability of treatment based on covariate combinations, rather than on the covariates themselves. Therefore, two individuals can be matched on propensity scores even if their actual covariate compositions differ substantially. The method balances out the propensity scores such that both treatment and control groups have equal probability of assignment to treatment (Rosenbaum & Rubin, 1983).

However, in 2019, King and Nielsen illustrated that PSM has a number of flaws that are minimized by other matching methods that match directly on covariates and maximize locally, rather than just globally, close matches. In particular, two major flaws are present with PSM. These flaws become exacerbated as researchers set calipers to constrain matching with the intention of eliminating members of the sample that are too dissimilar between groups.

The first flaw is that pruning cases results in increased variance in model estimates of the causal effect, such that estimates change substantially depending on which model is selected between two or more equally fitting models. This introduces researcher bias in model selection, whether intentional or not. The second flaw with PSM is what King and Nielsen term the "PSM Paradox." Despite its intention of creating closer matches, pruning observations through tightening calipers has the tendency to increase imbalance in the covariates. This occurs because the matches required to balance the groups overall do not necessarily match closely on their actual composition of confounding characteristics (King & Nielsen, 2019).

Instead of PSM, Zearn's efficacy analyses use a two-step Coarsened Exact Matching (CEM) method, with optimal matching, to create a control group that is as similar as possible to the treatment group of consistent Zearn Math users. CEM is a technique that simulates block sampling by matching students on covariates rather than propensity scores (Blackwell et al., 2009; Iacus et al., 2011). While propensity score matching simulates a process of random sampling, it is less effective than blocked or stratified

[1] In 2019, King & Nielsen counted over 141,000 journal articles that referenced or used Propensity Score Matching.

[2] For example, in education research, an individual's school or district likely impacts whether or not they receive a curriculum or classroom intervention. Prior performance might also be relevant if intervention is aimed specifically at low-performing or high-performing students. Language ability might be relevant if it is an intervention primarily provided to fluent English speakers. Demographic and socioeconomic factors are also frequently associated with whether or not an individual participates in a particular program or intervention.

In addition to shaping whether or not an individual receives an intervention, these covariate values are associated with differences in academic performance. Therefore in order to isolate an intervention as the "cause" of differences in academic achievement, quasi-experimental methods balance these factors so their impact is similar between treatment and control.

sampling at creating balanced samples because random assignment does not necessarily equal balance in groups (Imai et al., 2009; King et al., 2011; King & Nielsen, 2019). Similar to any matching methodology, CEM requires a robust set of observable covariates that effectively model the selection process to reduce bias in the estimates.

In order to see maximum benefit from Zearn Math, students are advised to complete three or more digital lessons per week during the school year. Therefore, the treatment group in an efficacy analysis is composed of consistent Zearn Math users, operationalized as students who completed an average of three or more digital lessons per week; 90 or more digital lessons per year. The control group is selected from other students in the district or state with little to no Zearn Math usage, operationalized as an average of less than one digital lesson per week; fewer than 30 digital lessons per year.

This definition of treatment and control does not use an intention-to-treat (ITT) framework that would include in the treatment all students that had been offered Zearn Math (McCoy, 2017). While the ITT approach is the most efficacious for identifying the impact of a program under real-world implementation constraints, the parameters of district and state partnerships and data availability preclude a robust implementation of this approach (the implications of Zearn's treatment framing are discussed further in the limitations section).[3]

Zearn Math users are matched with non-users on starting math and English Language Arts (ELA) achievement scores[4] along with demographic variables using a two-step CEM process. The goal of matching is to create 1:1 pairings between similar students, differing primarily on Zearn Math usage during the 2021-2022 school year. Controls are selected to match individuals in the treatment group, so analysis focuses on Zearn Math's average treatment effect.

Using CEM, treatment students are put into matching strata with control students that are in the same grade and have proximate scores. Scores may be in the form of scale scores or national percentiles and are selected based on the metric used by the district for internal goals and reports. Generally score distance is set at a maximum of 5-10 points on math and ELA. Score distance is selected to achieve closeness without losing a significant portion of the sample, thereby potentially making findings more

[3] Districts, many of which have purchased full district licenses, partner with Zearn to understand the impact of fidelity usage in the hopes of increasing fidelity usage of the platform across schools. In this sense, efficacy analyses examine the impact of Zearn Math at the student level, implemented with fidelity, vs. with little or no usage. Further, the structure of Zearn data makes it difficult to disentangle which students have and have not been "offered" Zearn Math. Many paid accounts centrally roster students such that student accounts are created for students whose teachers never utilize the product in the classroom. Conversely, accounts may be set-up with IDs that do not match the data provided by the district, thereby limiting universal match between usage and student achievement data.

[4] Generally, academic pre-scores are state assessment scores from the spring prior to the study year or the fall of the study year in districts that implement standardized fall assessments.

biased and less generalizable. Closeness on math is prioritized over ELA when necessary because of its greater relevance to outcomes.

Within strata, treatment students are matched to control students with whom they share a majority of characteristics on factors such as school, gender, race/ethnicity, special education status, English language proficiency, gifted status and economic disadvantage.

Zearn's matching method, called optimal matching, utilizes Bertsekas' auction algorithm to produce combinatorial optimization such that treatment individuals are matched to others closest to them in the control pool. When controls are the best-fit match for more than one treatment individual, the pairing goes to the individual from whom the next best pairing is the farthest (1981; Rosenbaum, 2020).[5]

If a treatment student has no match within their grade and score strata with whom they share a majority of characteristics, they are excluded from the treatment group. The caliper that limits match difference is selected to maximize inclusion in the sample and prevent biasing the sample through uneven patterns of exclusion, while ensuring similarity between groups.

Analysis

Once consistent Zearn Math users are matched to a similar group of low- or non-users, a difference of means analysis is conducted to quantify the impact of Zearn Math on student achievement. Means are calculated on users and non- or low-users overall as well as for groups disaggregated by starting math proficiency and demographics.

Difference in means t-tests are run on the average academic gains of consistent users vs. the average academic gains of low- or non-users to determine if the impact of treatment is statistically significant. Given *SD*=standard deviations and *n*=number of observations per group, t-tests are conducted as:

$$t = \frac{mean_{treatment} - mean_{control}}{\sqrt{\frac{SD^2_{treatment}}{n_{treatment}} + \frac{SD^2_{control}}{n_{control}}}} \quad \text{(Figure 1)}$$

[5] In other words, if Control Student A is the best match for Treatment Student 1 and Treatment Student 2, sharing 6 out of 7 characteristics with each, Control Student A can still only be matched with either Treatment Student 1 or Treatment Student 2. If the next best match for Treatment Student 1, Control Student B, shares 4 characteristics, and the next best match for Treatment Student 2, Control Student C, shares 5 characteristics, then Treatment Student 1 would be matched with Control Student A and Treatment Student 2 would be matched with Control Student C. In this way, the algorithm of optimal matching balances the closeness of any individual match with its impact on the closeness of the overall group match.

Effect size is calculated with *Cohen's d,* which divides the difference in means between treatment and control by the pooled standard deviations:

$$Cohen's\ d = \frac{mean_{treatment} - mean_{control}}{pooled\ SD}$$ (Figure 2)

In addition to capturing changes in student achievement across all users, efficacy analyses examine how Zearn Math use impacts the performance within demographic and academic student subgroups. When possible, these subgroups include gender, race/ethnicity, special education status, English-language proficiency, gifted status, economic disadvantage, grade and any other district or state specific groups or academic programs. Subgroup analysis may be limited by low sample size or data availability.

Zearn Math users and low- or non-users do not have to fully align on demographic characteristics. Thus, subgroups do not always align on starting proficiency, which is why demographic subgroup proficiencies are reported as difference-in-difference rather than as absolute scores or percentiles.

Limitations

Zearn's approach to efficacy research utilizes quasi-experimental methods to control for observed confounders and zoom in on the relationship between Zearn Math use and academic achievement. While these methods build promising evidence of Zearn Math's effectiveness, limitations to retroactive analyses exist that cannot fully replicate experimental conditions.

First, even with robust quasi-experimental methods, accuracy of estimates is limited by the ability to model all variables relevant to selection into treatment and control. Unobserved confounders may mediate the relationship between Zearn Math use and academic achievement, biasing results.

In addition, district- and state-level analyses examine the impact of fidelity implementation of Zearn Math rather than utilizing an intention-to-treat analytic framework that would define the treatment group as all students to whom Zearn Math was available (McCoy, 2017). The focus on fidelity implementation better aligns with the interest of partner districts, many of whom have made Zearn Math available to all students, but are observing more limited implementation in schools. Also, Zearn data does not well-capture the distinction between students who have and have not been offered access to the platform, making it difficult to accurately implement an ITT criteria.

Utilizing fidelity usage as the benchmark for treatment means that estimates may be biased as this usage represents the best version of implementation which may exceed "typical usage." In addition, focusing on lesson completion as a usage metric means that those who struggle more and therefore complete fewer lessons, even with the same amount of time on the platform, will be systematically

excluded from the analysis.

In the fall of 2022, a two-year randomized control trial (RCT) was launched to measure Zearn Math's impact. Many of the limitations associated with quasi-experimental analyses will be addressed in this RCT, which will compare outcomes between randomly assigned treatment and control schools under real-world usage conditions.

Alongside the RCT, Zearn is building broad and rigorous evidence of Zearn Math's efficacy with robust quasi-experimental methods and an expansion of efficacy studies to multiple districts and states across the country. Replication of trends and findings will provide even stronger evidence of Zearn Math's efficacy across a wide range of implementation contexts.

Works Cited

Bertsekas, DP. (1981). "A new algorithm for the assignment problem." Math. Prog. 21:152–71

Blackwell, M., Iacus, S., King, G., & Porro, G. (2009). cem: Coarsened exact matching in Stata. *Stata Journal*, *9*(4), 524-546.

Iacus, S. M., G. King, and G. Porro. (2011). "Multivariate Matching Methods that are Monotonic Imbalance Bounding." Journal of the American Statistical Association, 106, Pp. 345–361.

Imai, K., King, Gary & Nall, C. (2009). "The Essential Role of Pair Matching in Cluster-Randomized Experiments, with Application to the Mexican Universal Health Insurance Evaluation." Statistical Science, 24, 1, Pp. 29–53.

King, Gary, Nielsen, Richard, Coberley, Carter, Pope, James E., & Wells, Aaron. (2011). "Comparative Effectiveness of Matching Methods for Causal Inference".

King, Gary, & Nielsen, Richard. (2019). "Why Propensity Scores Should Not Be Used for Matching." Political Analysis, 27, 4, Pp. 435-454.

Ling, A., Montez-Rath, M., Mathur, M., Kapphahn, K., & Desai, M. (2020). How to Apply Multiple Imputation in Propensity Score Matching with Partially Observed Confounders: A Simulation Study and Practical Recommendations. Journal of Modern Applied Statistical Methods, 19(1), eP3439. https://doi.org/10.22237/jmasm/1608552120

McCoy C. E. (2017). Understanding the Intention-to-treat Principle in Randomized Controlled Trials. *The western journal of emergency medicine*, *18*(6), 1075–1078. https://doi.org/10.5811/westjem.2017.8.35985

Pearl, J. (2010). "The Foundations of Causal Inference." Sociological Methodology, 40, 1, Pp. 75–149.

Rosenbaum, P. R. (2020). Modern Algorithms for Matching in Observational Studies. *Annual Review of Statistics and Its Application*, *7*(1), 143-176. https://doi.org/10.1146/annurev-statistics-031219-041058

Stuart, E. A., & Rubin, D.B. (1983). "The Central Role of the Propensity Score in Observational Studies for Causal Effects." Biometrika 70:41–55.

Stuart, E. A., & Rubin, D.B. (2008). "Matching with Multiple Control Groups with Adjustment for Group Differences." Journal of Educational and Behavioral Statistics, 33, 3, Pp. 279–306.

Thum, Y. M., & Kuhfeld, M. (2020a). NWEA 2020 MAP Growth Achievement Status and Growth Norms for Students and Schools. NWEA Research Report. Portland, OR: NWEA

Thum, Y. M., & Kuhfeld, M. (2020b). NWEA 2020 MAP Growth Achievement Status and Growth Norms Tables for Students and Schools. NWEA Research Report. Portland, OR: NWEA.

Zearn. (2022a) *Zearn Impact: For students who scored below grade level in math, consistent Zearn usage tied to 2 grade levels of growth in 2 years of pandemic learning—nearly double the gains of curriculum alone.*
https://webassets.zearn.org/Implementation/Zearn_Impact_for_Students_Below_Grade_Level.pdf

Zearn. (2022b) *Consistent Zearn Usage Dramatically Reduces Learning Loss.*
https://webassets.zearn.org/Implementation/Zearn_Impact_for_Students_Below_Grade_Level.pdf