Replicating Studying Adaptive Learning Efficacy using Propensity Score Matching and Inverse Probability of Treatment Weighting

Shirin Mojarad, Ryan S. Baker, Alfred Essa, Steve Stalzer

Abstract

Despite the importance of replication, it remains rare in the interactive learning research community. In this paper, we attempt to replicate recent quasi-experimental results suggesting that the ALEKS intelligent tutoring system is effective at improving student course outcomes in higher education (Mojarad et al., 2018). In this paper, we conduct a near replication, collecting a new data set of higher education students using ALEKS, at the same university as in that earlier paper. We investigate the robustness of the results found to the choice of quasi-experimental methodology. In the earlier work, the popular propensity score matching algorithm was used; a recent methodological paper challenges this method (King & Nielsen, 2019). We therefore investigate the impact of using another matching algorithm, inverse probability of treatment weighting (IPTW) instead of propensity score matching, and compare the results obtained by these two methods. We replicate the previous study: ALEKS is statistically significantly associated with better student course outcomes. The use of IPTW leads to the same qualitative result as PSM, but IPTW achieves superior matching, suggesting that this method should be preferred for future quasi-experiments within the interactive learning research community.

Keywords: adaptive learning, efficacy, causal inference, quasi-experimental design, replication studies

Replicating Studying Adaptive Learning Efficacy using Propensity Score Matching and Inverse Probability of Treatment Weighting

The climate surrounding the potential uptake of interactive learning systems has shifted over the last few years. There has been a big push in education for new teaching and learning products and methods to provide evidence of effectiveness. One particularly important development in the United States of America has been the move towards databases that list curricula demonstrated to be effective, such as the What Works Clearinghouse (https://ies.ed.gov/ncee/wwc/) and Evidence for ESSA (https://www.evidenceforessa.org/). These databases are increasingly used by districts in their educational decision making (Slavin, 2017), as a complement to – or increasingly, a substitute for -- more thematic and feature-based reviews of learning technologies (e.g. Neumann et al., 2019). This is a beneficial step towards avoiding the use of ineffective learning technologies, many of which make strong claims for their effectiveness despite a lack of evidence (e.g. Kroeze et al., 2015).

Interactive learning systems such as adaptive learning systems have a generally successful track record in terms of effectiveness. In one of the earliest formal efficacy studies on adaptive learning, Koedinger and colleagues (1997) found that the PUMP Algebra Tutor (later called Cognitive Tutor Algebra; now called MATHia) led to better learning outcomes than a traditional control condition, both on a test tailored the learning system and on subsections of two standardized tests of mathematics. Later work on the same system replicated this finding, but only when teachers had some experience in implementing the system (Pane et al., 2013). Mitrovic and colleagues (2004) present a set of evaluations of three different adaptive learning systems on databases, showing that their systems lead to better results, including by comparison to a non-adaptive interactive learning system. Multiple evaluations of the ANDES adaptive

learning system for learning physics indicate that using that system leads to better learning than completing the same mathematics problems (without adaptivity) on paper.

One of the key modern forms of adaptive learning is learning systems that select what content the student should work on next (out of all possible content) based on an assessment of student knowledge, in order to focus student time. The most widely-used system of this nature is ALEKS. ALEKS has been found to lead to better learners than control conditions for community college students (Mojarad et al., 2018), in an afterschool program (Craig et al., 2011), and in a non-traditional adult education program (Rivera et al., 2017), although some studies have found contrasting results (Fang et al., 2019). Other learning systems that use this form of adaptivity have also produced successful outcomes (Baker et al., 2018).

Overall, several meta-analyses have indicated that adaptive learning systems lead to better learning outcomes for students, compared to either traditional instruction or less adaptive interactive learning systems (VanLehn, 2011; Ma et al., 2014; Kulik & Fletcher, 2016). Despite the excellent overall track record of interactive learning systems in improving educational outcomes, however, relatively few such systems are found in the databases that track learning interventions. Although there have been literally dozens of published investigations into the effectiveness of adaptive learning systems, relatively few of these studies have made their way into these databases, or indeed into the broader societal discourse about learning effectiveness.

Beyond this, as most methodologists are aware, a single study – though sufficient for inclusion into these databases – is not really sufficient evidence to conclude that an intervention is definitively effective. The recent "replication crisis" across scientific fields has demonstrated that many or even most scientific findings do not obtain the same results even in the most precise of replications (Loken & Gelman, 2017). Replication can fail for many reasons. First, it may be

that a result was obtained by chance (and, perhaps, a contrary finding could not be published due to the difficulty of publishing null results (Rosenthal, 1979). It may also be that some feature of the methodology leads to a failure to replicate; one does not need to engage in unethical practices such as "p-hacking" (Head et al., 2015) to inadvertently adopt a method which produces a spurious positive result.

In this paper, we take a recent "success", where an interactive learning system was found to lead to positive outcomes for students. Specifically, we consider a recent quasi-experimental study, published in conference proceedings, which found that the ALEKS adaptive learning platform produced positive results for community college students (Mojarad et al., 2018). We attempt to conduct a "close replication", studying the same system with the same experimental protocol at the same college, but in a different semester. A year upon year replication may seem very easy to achieve, but even this standard can be difficult to achieve. Take, for instance, Pane et al.'s (2013) research on Cognitive Tutor Algebra. In that work, Pane and his colleagues found no effects of using the Cognitive Tutor in the first year of implementation but found strong evidence in support of a positive effect in the second year of implementation (Pane et al., 2013). They attributed this finding to improved implementation over time. In general, existing educational clearinghouses tend to treat evidence of effectiveness as permanent. This type of assumption is not valid in medicine – for example, antibiotics can lose their effectiveness over time (Goossens et al., 2005). It is also unlikely to be true in education research. Interventions are modified over time to adapt to changing state standards (Massell & Perrault, 2014) as well as changes in the culture of schools (Sugai & Horner, 2002). The students themselves change over time; observations by Schofield (1995) of urban American students skipping lunch and staying after class to use an intelligent tutoring system do not largely seem to be behaviors that

commonly occur in 2018. As such, it is important not to simply study a system's effectiveness once, but to continue to re-investigate the system's effectiveness over time.

In the current paper, therefore, we take a previous set of analyses (Mojarad et al., 2018) and conduct the same set of analyses on a new population of learners in the subsequent year, to see if the same findings hold. In doing so, we follow that paper's method of conducting several quasi-experimental comparisons of the same study data, verifying whether the same finding is obtained for different comparisons between treatment and control groups. Those comparisons consider multiple ways of defining each group. By doing so, we generate richer evidence as to whether the adaptive learning system being studied – ALEKS – is effective at promoting student learning as it continues to be used in the same context over time. As such, this paper's primary research question is:

RQ1: As ALEKS is used in an ongoing fashion by a community college, do students using ALEKS continue to obtain better results than students not using ALEKS?In following up this past study, we investigate an additional concern that emerged around the time of the first study's publication: whether a specific methodological choice in the analysis may have led to the positive results seen, or whether the results are robust to this methodological choice. More specifically, Mojarad and her colleagues (2018) used a statistical method – propensity score matching – which has become subject to recent criticism (King & Nielsen, 2019). Therefore, an additional statistical method -- inverse probability of treatment weighting (IPTW) – is investigated, and the results are compared between methods. As such, our secondary research question is:

RQ2: Is evidence on the effectiveness of ALEKS robust to several different ways of selecting a matched-comparison control group, including the use of a different statistical method?

Quasi-Experimentation

For many educational researchers, practitioners, and policy-makers, the preferred "gold standard" study design for establishing educational effectiveness is the randomized controlled trial (Cook & Payne, 2002; Slavin, 2002; Riehl, 2006; Silverman, 2009; Torgerson & Torgerson, 2001). The key defining attribute of an RCT, as compared to other types of studies, is the random assignment of individual subjects (in this case, students) to control and intervention groups.

RCTs are considered by many to be the most rigorous study design due to randomization, i.e. randomly assigning subjects to treatment and control groups (Cook & Payne, 2002; Riehl, 2006; Torgerson & Torgerson, 2001). If subjects were unevenly assigned to condition, in terms of some covariate, there would be challenges to causal inference. Randomization, for sufficiently large sample sizes, is thought to remove the biases that result from covariate imbalance and create a study where the assignment of subjects to treatment or control groups can be assumed to be random, when conditioned on observable characteristics of study subjects, and where missing data can be treated as occurring at random (Schneider et al., 2007). In other words, RCTs allow researchers to assume that all other factors except for the comparison of interest are equal and to make causal claims based on their experimental observations. As a result, the U.S. Department of Education has emphasized RCTs as a preferred type of evidence for educational research ("Evidence-Based Interventions Under the ESSA - Every Student Succeeds Act," n.d.), and recent national-level policies in the United States treat RCTs as the highest strength of evidence

among study designs ("Evidence-Based Interventions Under the ESSA - Every Student Succeeds Act," n.d.).

In observational studies, by contrast, assignment to control groups and treatment groups is not random, and because of that, factors other than the impact of the treatment may confound the result. However, the use of randomization does not solve all challenges for educational effectiveness researchers. Rubin notes that even randomized studies must be designed to collect extensive covariate values to test for and control for observed random imbalances in covariate distributions between treatment and control groups (Rubin & van der Laan, 2008). Bloom emphasizes that also controlling statistically for baseline covariates, especially pretests, improves the precision of experimental studies (Bloom et al., 2007). However, a surprising number of RCT studies, across fields, ignore the need to investigate whether covariates differ across randomized treatment and control units (Deaton & Cartwright, 2018).

These limitations – differences in key covariates which can be hard to avoid – can be addressed by instead designing studies that explicitly stratify similar students into different conditions. In other words, rather than randomly assigning students to condition and hoping for the best, one can explicitly identify key covariates and ensure that they are balanced between condition (Tipton et al., 2014). This balancing of students can even occur post-hoc, by identifying students from a broader population who match the students who participated in a specific treatment (Rosenbaum & Rubin, 1983).

This type of study has another important virtue – practicality. It can often be difficult for schools or universities to engage in true random sampling at the student level. Many considerations enter into class and teacher scheduling other than the convenience of education researchers. Random sampling can also lead to threats to validity such as compensatory rivalry

(teachers in the control condition working harder because they know they are in the control condition) and resentful demoralization (students who know they are in an inferior control condition becoming discouraged and putting in less effort) (Mcmillan, 2007). The challenges to subject recruitment and implementation seen for experimental studies can also lead to overly controlled studies that do not match genuine classroom conditions, or to small or unrepresentative populations, challenges that question the real-world and broader applicability of research conducted solely through RCTs (Feng et al., 2014). Overall, then, although RCTs are often considered the "gold standard" in evaluation research, they are still vulnerable to a range of biases.

In this paper, we present an example of an alternative to RCTs, post-hoc quasi-experimentation (QE) using causal inference. This method consists of taking a known treatment group, where there was not randomized assignment, and comparing it to a carefully selected comparison group, where covariates are matched between the two groups. This method has grown in popularity in education research over the last decade, leading to its use to study a range of research questions, from the effectiveness of special education services (Morgan et al., 2010), to the effects of school size on student attainment (Wyse et al., 2008), to the effectiveness of specific instructional and remedial programs (e.g. Bhatt & Koedel, 2012; Yamada & Bryk, 2016). Specifically, the study we seek to replicate in this article was an example of a quasi-experimental study.

Using causal inference and quasi-experimentation relieves some of the validity threats seen for RCTs. However, it opens researchers to several criticisms. In particular, it opens questions of "cherry-picking". If a researcher tries enough different comparisons and tests, it is quite plausible that one such comparison will produce the answer the researcher is looking for. If

the researcher then "cherry-picks" only that comparison and statistical test, they can produce the appearance of a positive result even when most other comparisons would produce a null result or even the opposite result (see discussion in Raudenbush, 2007).

This can be avoided, we argue, by presenting not one "theoretically best" quasi-experimental causal comparison, but by presenting several distinct comparisons. If several comparisons are made, and the comparisons are designed to be fairly different than one another, and the same result is obtained each time, it presents stronger evidence that the result obtained is valid than any single comparison could produce. In Mojarad et al. (2018), the researchers conducted five different distinct comparisons for the same experimental question, to avoid criticism that the analysis chosen was selected to produce the desired result. Four comparisons represented comparisons between groups with straightforward definitions and delineations. The fifth comparison used a statistical technique, propensity score matching (PSM) (Rosenbaum & Rubin, 1983), to align between the two groups.

In observational studies, subjects are not assigned randomly to treatment and control groups. Instead, treatment assignment is often influenced by subject characteristics. Thus, there are frequently systematic differences in baseline characteristics between treatment groups. This can result in confounding, in which differences in outcomes between treatment groups are due, at least in part, to systematic differences in baseline covariates between the treatment groups. Matching methods reduce or minimize the effects of confounding due to measured baseline covariates. In propensity score matching, for each member of the intervention group, we identify a member of the control group that is as similar as possible in terms of their propensity score. Then, the difference in outcomes between the matched pair is computed. The average of this difference over the observed pairs is an estimate of the mean causal effect of a particular

intervention on outcome. A propensity score is used to choose treatment and control groups with similar baseline characteristics. A propensity score is defined as the probability of the subjects being assigned to the treatment group, given a set of baseline characteristics (Rosenbaum & Rubin, 1983). Therefore, in PSM the subjects from treatment and control groups are matched using their probability of treatment assignment conditional on observed baseline covariates (Rosenbaum & Rubin, 1983). Commonly, a logistic regression model is used to calculate the propensity scores of students. However, since Mojarad et al. (2018) was published, a strong critique of propensity score matching has been published (King & Nielsen, 2019). To quote the authors of that critique, "propensity score matching… often accomplishes the opposite of its intended goal — thus increasing imbalance, inefficiency, model dependence, and bias". King and Nielsen argue that the mathematics underlying PSM attempts to approximate a randomized experiment rather than a fully blocked randomized experiment.

As a response to this criticism, we consider an alternate matching approach, inverse probability of treatment weighting (IPTW). IPTW has been argued to perform better at accounting for biases due to observed confounders (Austin, 2011). We apply IPTW as well as PSM, and investigate both whether the same qualitative findings are obtained, and what the properties of each match are.

IPTW uses weights based on the propensity score to create a sample in which the distribution of measured baseline covariates is independent of treatment assignment (Rosenbaum, 1987). Using IPTW, a subject's weight is equal to the inverse of the probability of receiving the treatment that the subject actually received. This probability is propensity score in case of treatment subjects and is one minus the propensity score in case of control subjects. For subject i, the assigned weight w_i can be defined as:

$$w_t = \frac{z_t}{p_t} + \frac{1 - z_t}{1 - p_t}$$

Where z_i is a binary variable, indicating whether the subject is treated or not, and p_i is the propensity score.

System

ALEKS (Assessment and LEarning in Knowledge Spaces) is an adaptive learning system designed for courses in science and mathematics. ALEKS has several mathematics courses that cover developmental mathematics for K-12, four year and two-year colleges.

ALEKS uses Knowledge Space Theory (KST) (Doignon & Falmagne, 2011) to determine what students know, what they don't know, and what they are most ready to learn. KST applies concepts from Combinatorics and stochastic processes to the modeling and empirical description of particular fields of knowledge. Within this theory, a mathematical language has been developed to delineate the ways in which particular elements of knowledge (concepts in Algebra, for example) can be gathered to form distinct knowledge states of individuals.

This framework enables the creation of computer algorithms for the construction and application of discipline-specific knowledge structures known as "Knowledge Spaces", used to map the details of each student's knowledge. ALEKS infers, at each moment, with respect to each individual topic, whether each individual student has mastered that topic. If the student has not yet mastered the topic, ALEKS infers whether she is likely to be ready to learn the topic at that moment. ALEKS uses this knowledge to make learning more efficient and effective by continuously offering the student a selection of only the topics she is ready to learn at the current time.

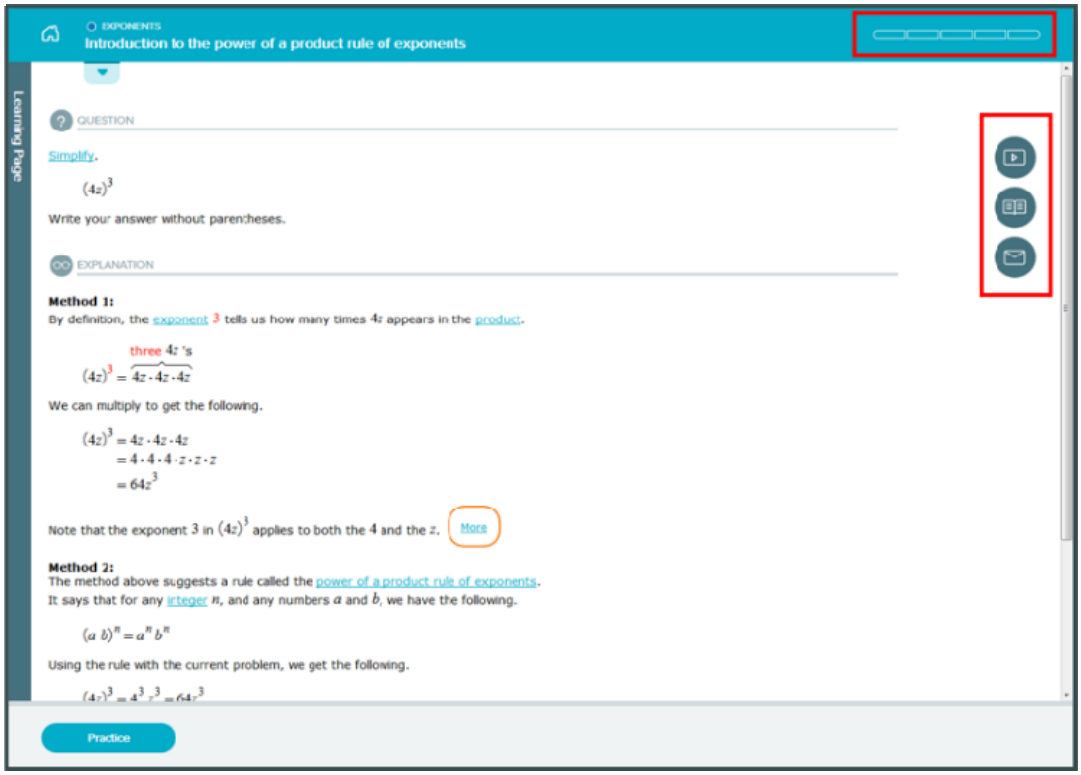
Once students register with ALEKS, they take a brief tutorial on how to use the system. Next, they complete a computer adaptive test called Initial Knowledge Check (IKC). The purpose of the IKC is to decide what they know and don't know, so that ALEKS can guide them to start with material they are ready to learn.

Once students are in Learning Mode, they will alternate between instruction and practice problems to learn each topic. They also have access to the resources such as worked examples in ALEKS to help them learn the topic. A sample of a Learning Mode page is shown in Figure 1.

In addition to the IKC, ALEKS regularly conducts progress knowledge checks to see if the students remember what they learnt and what they need to review again. These knowledge checks appear periodically throughout learning based on how instructors setup the course.

Figure 1

ALEKS Learning Mode

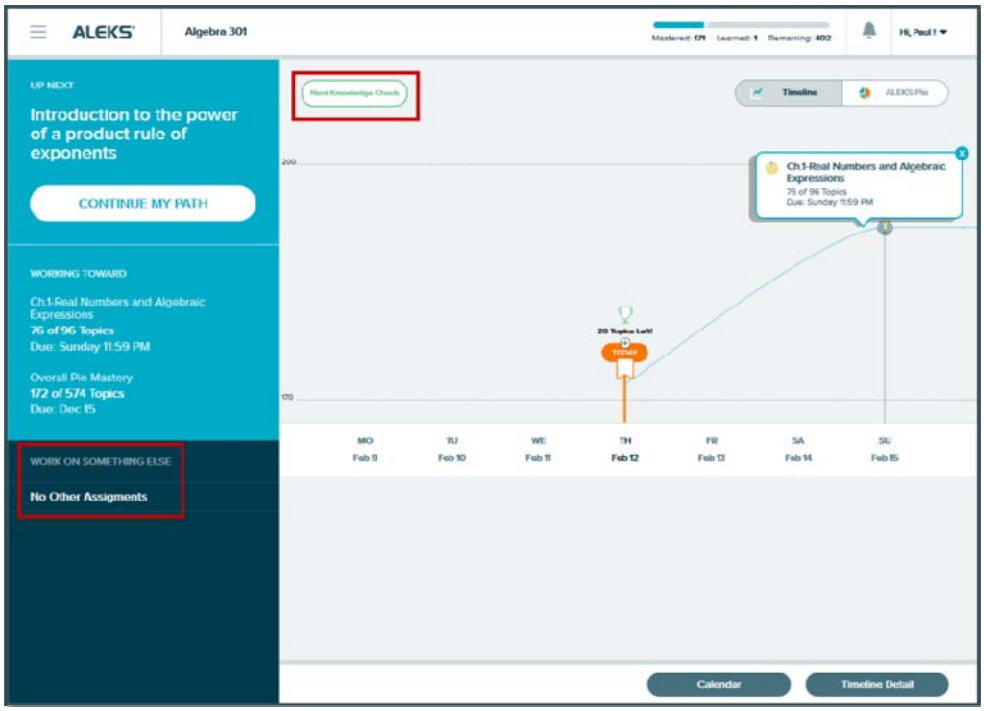


In the ALEKS Learning Mode, students can alternate between lesson pages and practice problems to learn each topic. The resources are located on the right side of each lesson page and practice problem as icons. There is also a gauge on the top right corner that tracks how many correct problems the student needs to finish learning the topic.

The ALEKS Calendar tracks students' weekly progress to help them manage their study time. It shows how much the student have left to do in order to finish their assignment on time. Students can also select next Knowledge Check to see when their next one will occur. There is an option to delay a Knowledge Check for up to 24 hours. The ALEKS Calendar shows how much the student have left to do in order to finish their assignment on time. Students can also select next Knowledge Check to see when their next one will occur A sample of this timeline is shown in Figure 2.

Figure 2

The ALEKS Calendar



Prior to Mojarad et al. (2018), Craig et al. (2011) studied the effectiveness of using ALEKS in improving mathematical skills of struggling students in an after-school program. Using random assignment of students to use ALEKS, they demonstrated that students using ALEKS significantly outperformed students assigned to a control condition on a state assessment test. The control condition included teacher-based lecture, while in ALEKS condition students only interacted with the program.

Method

Study

In this paper, we conduct analyses on data from three studies relevant to the success of students using ALEKS. The first two studies, previously published in a conference paper (Mojarad et al., 2018), involve two semesters of data from students in a community college

which used ALEKS in some sections, in Fall 2016 and Spring 2017. The third study, entirely new to this paper, was conducted in a subsequent semester, involving a different set of students, in Fall 2017. A very small number of students may have been included in multiple studies (semesters), due to failing the course the first time they took it.

Each of these studies investigate the effectiveness of ALEKS within the context of a large community college in the Midwestern United States. Within this specific community college, it was not practical to randomly assign instructors or classes to conditions, as the college has made a policy decision that eliminates the ability to use an RCT design to study the efficacy of its chosen product. Instead, the college's administration decided that instructors would be given the choice of adopting ALEKS in their courses, and many instructors chose not to use it. Even when an instructor did choose to adopt ALEKS, it was not required for students, and was counted minimally towards the final grade. Therefore, only a portion of students in classes adopting ALEKS ever used ALEKS, and many students may not have used ALEKS to the degree or in the fashion intended. Therefore, we probably should not simply compare ALEKS classes to non-ALEKS classes; there are both selection bias issues and valid concerns about implementation fidelity (Feng et al., 2014).

Data was collected retrospectively from the community college and the ALEKS platform. The data collected from the community college consisted of course enrollment and course outcomes, as well as data on student demographics including age, gender, Accuplacer score, and race. The course outcome was measured by a final exam developed and conducted by the school, without feedback from the ALEKS team. These students were enrolled in one of four courses at the college: pre-algebra, elementary algebra, intermediate algebra and college math. The school guidelines set the criteria for passing the course as grades C+ and above. Accuplacer is a

placement test, used for placement of post-secondary students into courses of the appropriate difficulty level (Mattern & Packman, 2009). Since a major number of students are placed into developmental math courses using Accuplacer score, we have used Accuplacer Arithmetic score as a proxy for students' initial knowledge. Accuplacer score is scaled between 20 and 120. In addition, we matched student records to their activity in ALEKS, where we identify whether they used ALEKS or not.

Table 1 shows the breakdown of number of total and ALEKS sections and students by semester. The original two studies in Fall 2016 and Spring 2017 included 3,925 students in 198 sections covering four courses. Amongst the 198 sections, 37 sections (19% of total sections) with a total 724 students (18% of total students) adopted ALEKS. ALEKS adoption was decided by the instructors who volunteered to use ALEKS in their classroom. From these 724 students, only 425 (59% of students in ALEKS sections) used ALEKS at least once, meaning that they at least completed the initial assessment in ALEKS. Note that this is a very minimal definition of usage – some students included in this category completed no learning content within ALEKS.

The replication study from Fall 2017 drawn from a new cohort taking the same four courses includes 2,072 students in 98 sections, from which 31 sections (32% of total sections) adopted ALEKS. This 32% adoption rate represents a statistically significant increase compared to the previous year's 19% adoption rate, $\chi^2(df=1,N=296) = 5.49$, $p= 0.019$, with a corresponding increase in the total number of students in ALEKS sections, $\chi^2(df=1,N=5999) = 84.9$, $p= <0.001$ and the proportion of students actually using ALEKS, again defining usage as using ALEKS at least once $\chi^2(df=1,N=5461) = 61.9$, $p= <0.001$.

These 31 sections included 598 students (29% of total students) from which 361 students (60% of students in ALEKS sections) used ALEKS, again defining usage as using ALEKS at

least once. The rate of students in ALEKS sections choosing to use ALEKS was almost identical between years (60% versus 59%). This pattern of results suggests that selection biases may have shifted between years at the instructor level but that selection biases were likely similar between years at the student level.

Table 1

The Breakdown of Number of Total and ALEKS Sections and Students by Semester

Semester	Dataset	# total sections	# total students	# ALEKS sections	# students in ALEKS sections	# ALEKS students
Fall '16	Original	98	2173	9	198	125
Spring '17	Original	100	1747	28	526	300
Fall '17	Replication	98	2072	31	598	361

Figure 3 shows a representation of ALEKS and non-ALEKS sections and students for the previously published (Fall 16/Spring 17) and replication (Fall 17) studies shown in Figure 3a and 3b, respectively.

Figure 3

Representations of ALEKS and Non-ALEKS Sections and Students for (a) original studies (Fall 2016 and Spring 2017) (b) replication study (Fall 2017)

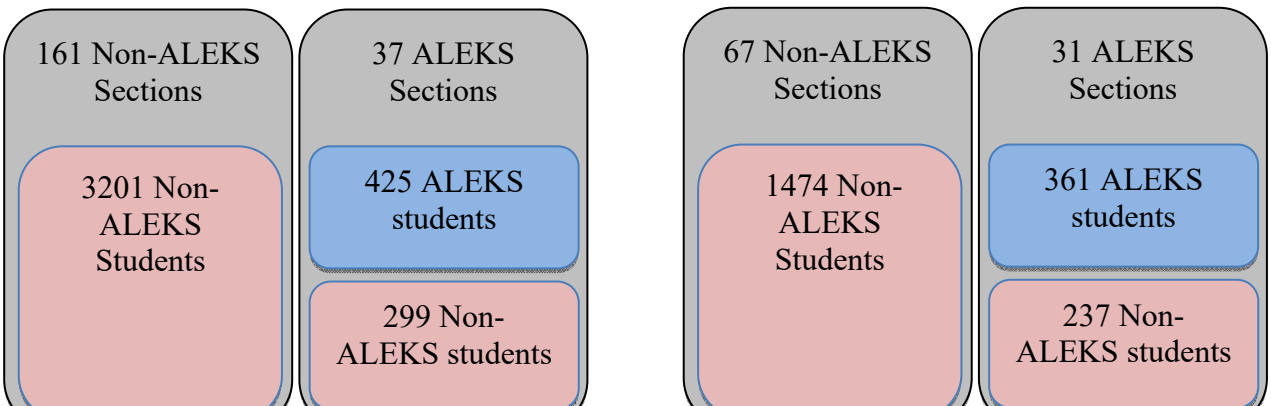


Figure 3a **Figure 3b**

Analysis Methods

In this section, we discuss the methods used in our analyses of whether students had higher pass rates when they used ALEKS than when they did not use ALEKS. In doing so, we investigate five different possible comparisons between ALEKS students and non-ALEKS students, in order to be confident that our results are not simply due to conducting a specific analysis. As mentioned above, we conduct these analyses both on data from two studies previously presented in a conference paper (Fall 2016 and Spring 2017) and also for a new replication study (Fall 2017), helping us to see whether the previously-obtained results are robust over time.

In Mojarad et al. (2018), comparisons were made for five possible breakdowns of ALEKS vs. Non-ALEKS students and sections, a structure we replicate here in our attempt to replicate that analysis approach on a new study. These breakdowns included:

1. ALEKS students vs. all Non-ALEKS students (in both ALEKS and non-ALEKS sections)
2. ALEKS students in ALEKS sections vs. Non-ALEKS students in ALEKS sections
3. ALEKS sections vs. Non-ALEKS sections
4. ALEKS students in ALEKS sections vs. Non-ALEKS students in Non-ALEKS sections

5. Matched ALEKS students vs. Matched Non-ALEKS students, using Propensity Score Matching

The first four of these comparisons are designed to investigate whether a difference between ALEKS and non-ALEKS students is seen for different student breakdowns. These four comparisons involve samples that may be subject to some degree of selection bias, and hence, a fifth study was conducted using a quasi-experimental design with propensity score matching to remove this bias. This fifth comparison had the goal of creating a fair comparison of ALEKS and Non-ALEKS students by balancing possible confounding attributes including age, gender, Accuplacer score and race.

These five comparisons are replicated in the replication study. In addition, a sixth comparison creates a matched comparison using Inverse Proportion of Treatment Weights instead of Propensity Score Matching, a method not applied in the original studies (and therefore representing new analysis rather than replication analysis). This method is discussed in detail in the section below “An Alternative to Propensity Score Matching”.

In this study, matching is done using three student characteristics: Accuplacer arithmetic score, age, and whether the student’s race is classified as minority or not. These attributes were chosen due to possible links to both outcomes and a student’s choice of whether to use ALEKS. The Accuplacer score is used by the college to decide whether to place students into developmental math courses and is used as a measure of students’ prior knowledge in the subject (Mattern & Packman, 2009). It is possible that prior knowledge could influence both a student’s choice of whether to use ALEKS (perhaps, for instance, a struggling student might be more likely to seek learning support) and their final outcomes. Therefore, we used Accuplacer as a proxy for initial knowledge and included it in our matching procedure. Age may influence the

choice of whether to use ALEKS (older, returning students may feel less comfortable with adaptive learning technology than younger students) and outcomes (e.g. Cantwell et al., 2001). Finally, minority status serves as a proxy for several factors, many of which may be associated with the choice to use ALEKS and success in college. Being a member of a minority group is known to be associated with different outcomes in college in the United States (Cameron & Heckman, 2001). The student group from which the control matches where identified included only non-ALEKS students in non-ALEKS sections. This naturally removes the student selection bias in the control group, since students in non-ALEKS sections do not have a choice to use ALEKS – we use matching to control for student selection bias in the experimental group. In this study, a logistic regression model is used to calculate the propensity score of students.

We compare students in terms of whether they pass a mathematics test; within each subject, the college gives the same test to every student in every class section at the end of the semester. Specifically, we use a chi-square (χ^2) contingency test to compare the pass rate between the two groups (Rao & Scott, 1984), as in the original paper being replicated. We also report effect size's, using Cohen's D, a measure of the distance between two group means, divided by their pooled standard deviation (Cohen, 1988). Cohen's guidelines for interpreting the magnitude of d in the social sciences is that an effect size around 0.2 represents a small effect, an effect size around 0.5 represents a medium effect, and an effect size around 0.8 represents a large effect (Cohen, 1988). However, Hill et. al. argue that effect sizes should be interpreted with respect to empirical benchmarks that are relevant to the intervention, target population, and outcome measure being considered (Hill et al., 2006). In a meta-analysis conducted on educational research, John Hattie argues that effect sizes of 0.00 to +0.15 are "developmental effects" (what students could probably achieve if there were no schooling), +0.15 to +0.40 are

"teacher effects" (what teachers can do without any special practices or programs), and +0.40 to +1.20 are the "zone of desired effects" (Hattie, 2008).

In the next section, we will discuss the differences between conditions according to the five comparisons, as well as evidence for the validity of causal conclusions drawn from this comparison.

Results of Replication

Tables 2 and 3 show the ALEKS and non-ALEKS group pass rates for each of the five comparisons, the difference in pass rates between the ALEKS and non-ALEKS groups, and the p-values for each of the comparisons. Table 3 shows the replication study, and Table 2 shows the original two studies for purposes of comparison (i.e. Mojarad et al., 2018). Table 4 shows the effect sizes (Cohen's D) associated with those comparisons. As discussed above, we conduct five comparisons. The first comparison is all students who at least took an initial assessment in ALEKS (ALEKS students) versus all students who did not use ALEKS in the course of the class (non-ALEKS students). Within this comparison, for both the original and replication studies, ALEKS students had statistically significantly higher pass rates. For the original pair of studies, χ^2 (df=1, N=3925) = 28.7, p<0.001, d = 0.29, with ALEKS achieving a boost (the increase in pass rates) of 14 points in pass rates. For the replication study, χ^2 (df=1, N=2072) = 49.8, p<0.001, d = 0.43, with ALEKS achieving a boost of 20 points in pass rates.

The second comparison is between ALEKS and non-ALEKS, but only within ALEKS sections. This comparison is important as it naturally controls for the instructor and class environment, by comparing students who did and did not use ALEKS within the same class. Within this comparison, ALEKS students had statistically significantly higher pass rates. For the original pair of studies, χ^2 (df=1, N=724) = 24.7, p<0.001, d = 0.38, with ALEKS achieving a

boost of 19 points in pass rates. For the replication study, χ^2 (df=1, N=600) = 21.8, p<0.001, d = 0.40, with ALEKS achieving a boost of 18 points in pass rates.

The third comparison considers assignment at the classroom level. In this comparison, all students within ALEKS sections, whether they did or did not use ALEKS, are compared against all the students in non-ALEKS sections. This comparison is perhaps the most common traditional quasi-experimental comparison, in the absence of modern statistical methods for causal inference in quasi-experimental studies, but may be vulnerable to issues of implementation fidelity within the experimental condition (O'Connell, 2008). Nonetheless, even within this comparison, ALEKS students had statistically significantly higher pass rates. For the original studies, χ^2 (df=1, N=3925) = 7.5, p=0.006, but only a modest effect size, d = 0.12, with ALEKS achieving a boost of 6 points in pass rates. For the replication study, χ^2 (df=1, N=2074) = 29.9, p<0.001, d = 0.27, with ALEKS achieving a boost of 13 points in pass rates.

The fourth comparison is between ALEKS students in ALEKS sections and non-ALEKS students in non-ALEKS sections. Within this comparison, we are excluding non-ALEKS students in ALEKS sections from this comparison as those are the students who chose not to use ALEKS, despite having the option of using it in the class. Including these students includes students who did not participate in the treatment, despite being assigned to the treatment group, creating concerns about implementation fidelity. Within this comparison, ALEKS students had statistically significantly higher pass rates. For the original studies, χ^2 (df=1, N=3626) = 26.7, p<0.001, d = 0.28 with ALEKS achieving a boost of 14 points in pass rates. For the replication study, χ^2 (df=1, N=1837) = 49.8, p<0.001, d = 0.43, with ALEKS achieving a boost of 20 points in pass rates.

Finally, comparison five attempts to avoid the biases inherent in the first four comparisons, by comparing ALEKS students who are matched with similar non-ALEKS students in non-ALEKS classes, using propensity score matching. The matching is done using Accuplacer, age and minority and as shown above, the students selected in the matching process have similar prior knowledge, age, and minority between conditions. All students in the matched treatment condition used ALEKS and all students in the matched control condition did not use ALEKS. Within this comparison, ALEKS students had statistically significantly higher pass rates. For the original studies, χ^2 (df=1, N=748) = 7.5, p=0.005, d = 0.20, with ALEKS achieving a boost of 15 points in pass rates. For the replication study, χ^2 (df=1, N=566) = 16.3, p<0.001, d = 0.35, with ALEKS achieving a boost of 16 points in pass rates.

As shown in Tables 2 and 3, all five comparisons are statistically significantly in favor of ALEKS, with a boost of 6 to 19 points in pass rates between ALEKS and non-ALEKS users across different comparisons for the original studies and 13 to 20 for the replication study. Some of the comparisons are likely to be biased in favor of ALEKS, others against ALEKS, but overall, they tell a common story – ALEKS is statistically significantly more effective at enhancing pass rates compared to the control condition.

Table 2

Pass Rates and Significance Level for ALEKS and Non-ALEKS Users, for Original Studies (Fall 2016 and Spring 2017)

Comparison	Pass Rates for ALEKS vs. Non-ALEKS	Boost	p-value
1. ALEKS students vs. all Non-ALEKS	71% vs 57%	+14	<0.001

students			
2. ALEKS students vs. Non-ALEKS students in ALEKS sections	71% vs 52%	+19	<0.001
3. ALEKS sections vs. Non-ALEKS sections	63% vs 57%	+6	0.004
4. ALEKS students in ALEKS sections vs. Non-ALEKS students in Non-ALEKS sections	71% vs 57%	+14	<0.001
5.Matched ALEKS students vs. Matched Non-ALEKS students (quasi-experimental study using Propensity Score Matching)	70% vs 60%	+10	<0.001

Table 3

Pass Rates and Significance Level for ALEKS and Non-ALEKS Users, for Replication Study (Fall 2017)

Comparison	Pass Rates for ALEKS vs. Non-ALEKS	Boost	p-value
1. ALEKS students vs. all Non-ALEKS students	74% vs 54%	+20	<0.001
2. ALEKS students vs. Non-ALEKS students in ALEKS sections	74% vs 56%	+18	<0.001
3. ALEKS sections vs. Non-ALEKS sections	67% vs 54%	+13	<0.001
4. ALEKS students in ALEKS sections vs. Non-ALEKS students in Non-ALEKS sections	74% vs 54%	+20	<0.001

5.Matched ALEKS students vs. Matched Non-ALEKS students (quasi-experimental study using Propensity Score Matching)	72% vs 56%	+16	<0.001

Table 4

Effect Size for Pass Rates Before and After Matching for Both Original and Replication Studies

Comparison	Previous studies (Cohen's D)	Replication study (Cohen's D)
1. ALEKS students vs. all Non-ALEKS students	0.29	0.43
2. ALEKS students vs. Non-ALEKS students in ALEKS sections	0.38	0.40
3. ALEKS sections vs. Non-ALEKS sections	0.12	0.27
4. ALEKS students in ALEKS sections vs. Non-ALEKS students in Non-ALEKS sections	0.28	0.44
5.Matched ALEKS students vs. Matched Non-ALEKS students (quasi-experimental study using Propensity Score Matching)	0.20	0.35

An Alternative to Propensity Score Matching

As discussed above, the method of propensity score matching has recently come under sharp criticism. In this section, we consider a popular alternative to propensity score matching, inverse probability of treatment weighting (IPTW). We first compare the quality of the match obtained by each approach, and then consider whether they produce different ultimate results in

terms of the efficacy of ALEKS. We apply this method to both the data from the original studies, and to the data from the replication study.

To ensure balance across baseline characteristics between the treatment and control groups, Austin (2009a) recommends that researchers report the mean, standard deviation and the effect size (Cohen's D) of each attribute across the two treatment and control groups. Cohen's D is recommended to be used to evaluate the degree of balance between conditions before and after PSM (Austin, 2011), in part because it is a standardized measure and therefore can be compared across attributes with different scales.

Tables 5-7 show a list of all the considered potential confounders' mean, standard deviation and SMD (effect size) across ALEKS and Non-ALEKS students before matching, after PSM matching, and after IPTW matching respectively for the original studies. Note that we had to exclude some students from the analysis due to missing Accuplacer scores. Therefore, the number of both non-ALEKS and ALEKS students is fewer than in the original studies.

Table 5

Confounders' Mean, Standard Deviation and Cohen's D Across ALEKS and Non-ALEKS Students Before Matching, for Original Studies

Variable	Non-ALEKS Students	ALEKS Students	Cohen's D
N	2519	374	
Age Average (std)	26.92 (8.95)	27.21 (9.02)	0.033
Accuplacer Arithmetic Average (std)	55.03 (22.40)	55.34 (22.83)	0.014
Minority Average (std)	0.73 (0.44)	0.66 (0.47)	0.150

Gender Average (std)	0.59 (0.49)	0.63 (0.48)	0.081

Table 6

Confounders' Mean, Standard Deviation and Cohen's D Across ALEKS and Non-ALEKS Students After PSM Matching, for Original Studies

Variable	Non-ALEKS Students	ALEKS Students	Cohen's D
N	374	374	
Age Average (std)	27.12 (9.01)	27.21 (9.02)	0.010
Accuplacer Arithmetic Average (std)	53.61 (22.35)	55.34 (22.83)	0.077
Minority Average (std)	0.66 (0.47)	0.66 (0.47)	<0.001
Gender Average (std)	0.63 (0.48)	0.63 (0.48)	<0.001

Table 7

Confounders' Mean, Standard Deviation and Cohen's D Across ALEKS and Non-ALEKS Students After IPTW Matching, for Original Studies

Variable	Non-ALEKS Students	ALEKS Students	Cohen's D
N	374	374	
Age Average (std)	27.22 (9.10)	27.21 (9.02)	0.001
Accuplacer Arithmetic Average (std)	55.31 (22.53)	55.34 (22.83)	0.002
Minority Average (std)	0.66 (0.47)	0.66 (0.47)	<0.001
Gender Average (std)	0.63 (0.48)	0.63 (0.48)	0.001

Austin (2009a) proposes that a Cohen's D of 0.1 denotes meaningful imbalance in a baseline covariate. The results in Table 5 show that in terms of this criterion, the two groups are initially not balanced in terms of minority groups (Cohen's D>0.1), with the non-ALEKS students group having higher percentage of minorities (73%) compared to the ALEKS students group (66%). In addition, the balance of gender across two students groups before matching is close to the threshold (Cohen's D > 0.08) with the non-ALEKS students group having a lower percentage of female students (59%) than the ALEKS students group (63%). After matching, both PSM and IPTW achieve acceptable balance across all attributes. However, with IPTW, we achieve similar or better matching for all confounders than with PSM. Figure 4 shows the balance for each confounder and method for the original studies. This figure represents a consolidation of attribute balances, in terms of Cohen's D, summarizing Tables 5-7; it can be used to observe and compare the balances before and after matching using different methods.

Tables 8-10 show similar data to Tables 5-7, but present descriptive statistics for the replication study. Again, some students were excluded from the analysis due to missing Accuplacer scores.

Figure 4

Balance of Attributes in the Original Studies Before Matching and After Matching Using Both IPTW and PSM

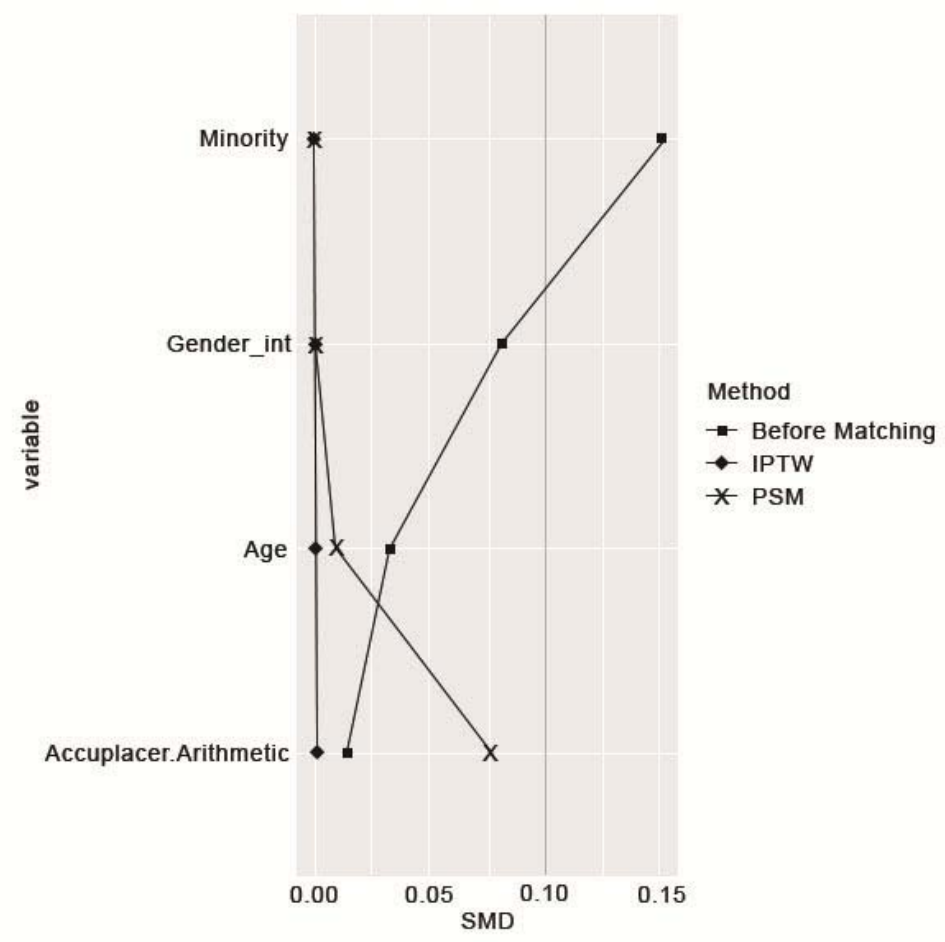


Table 8

Confounders' Mean, Standard Deviation and Cohen's D Across ALEKS and Non-ALEKS Students Before Matching, for Replication Study

Variable	Non-ALEKS Students	ALEKS Students	Cohen's D
N	1130	283	
Age Average (std)	26.20 (8.52)	27.16 (8.69)	0.112
Accuplacer Arithmetic Average	52.24 (23.41)	51.73 (23.64)	0.022

(std)			
Minority Average (std)	0.76 (0.43)	0.76 (0.43)	0.007
Gender Average (std)	0.59 (0.49)	0.62 (0.49)	0.061

Table 9

Confounders' Mean, Standard Deviation and Cohen's D Across ALEKS and Non-ALEKS Students After PSM Matching, for Replication Study

Variable	Non-ALEKS Students	ALEKS Students	Cohen's D
N	283	283	
Age Average (std)	27.41 (9.10)	27.16 (8.69)	0.028
Accuplacer Arithmetic Average (std)	52.63 (23.39)	51.73 (23.64)	0.038
Minority Average (std)	0.74 (0.44)	0.76 (0.43)	0.046
Gender Average (std)	0.59 (0.49)	0.62 (0.49)	0.061

Table 10

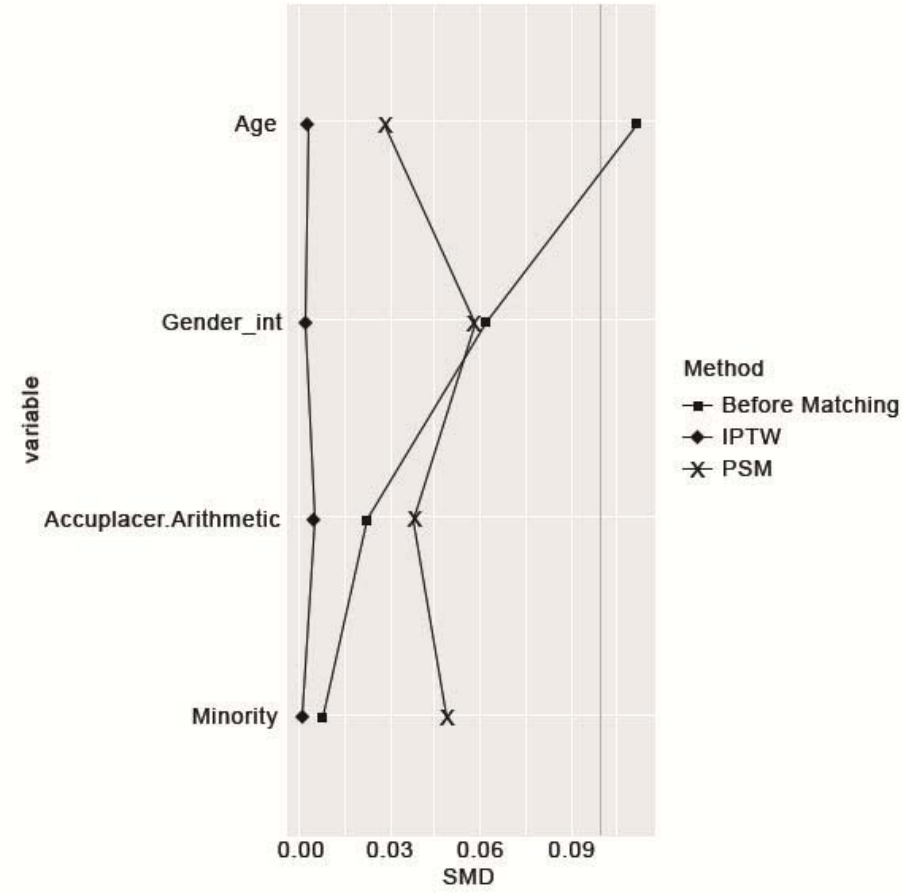
Confounders' Mean, Standard Deviation and Cohen's D Across ALEKS and Non-ALEKS Students After IPTW Matching, for Replication Study

Variable	Non-ALEKS Students	ALEKS Students	Cohen's D
N	283	283	
Age Average (std)	27.19 (9.40)	27.16 (8.69)	0.003
Accuplacer Arithmetic Average (std)	51.83 (23.17)	51.73 (23.64)	0.004

Minority Average (std)	0.76 (0.43)	0.76 (0.43)	<0.001
Gender Average (std)	0.62 (0.49)	0.62 (0.49)	<0.001

Figure 5

Balance of Attributes in Replication Study Before Matching and After Matching Using Both IPTW and PSM



Similarly, when looking at the replication study, we find that all of the confounders are reasonably balanced across ALEKS and Non-ALEKS users even before matching. The most imbalanced confounder is age, which is on average slightly higher amongst ALEKS students, compared to non-ALEKS students. After PSM, the balance on age is improved, but the balance

for the other confounders actually worsens slightly. For IPTW, we achieve substantially better matching for all confounders. Figure 5 shows the balance for each confounder and method for the original studies.

This result, in total, suggests that these studies probably did not need either PSM or IPTW to achieve acceptable balance, but that IPTW achieves better balance compared to PSM for both the original and replication studies. This result suggests that IPTW is a better method for creating pseudo-populations for treatment and control groups, in which confounders and treatment are unrelated to each other.

Our next question is whether the choice of IPTW versus PSM impacts our results. Table 11 and 12 show group pass rates, boost (the increase in pass rates) and *p*-value for both PSM and IPTW in the original and replication studies respectively. Note that we also report PSM results for all studies in tables 2 and 3 above – we re-present them here for easy comparison between PSM and IPTW. The results in these tables show that IPTW and PSM show similar results for ALEKS effectiveness after matching. Hence, although IPTW is arguably the better approach, the choice of method does not impact the results of the analysis.

Table 11

Pass Rates and Significance Level for ALEKS and Non-ALEKS Users, for Original Studies (Fall 2016 and Spring 2017)

Comparison	Pass Rates for ALEKS vs. Non-ALEKS	Boost	p-value
Matched ALEKS students vs. Matched Non-ALEKS students	70% vs 60%	+10	<0.001

Using PSM			
Matched ALEKS students vs. Matched Non-ALEKS students	70% vs 59%	+11	<0.001
Using IPTW			

Table 12

Pass Rates and Significance Level for ALEKS and Non-ALEKS Users, for Replication Study (Fall 2017)

Comparison	Pass Rates for ALEKS vs. Non-ALEKS	Boost	p-value
Matched ALEKS students vs. Matched Non-ALEKS students	72% vs 56%	+16	<0.001
Using PSM			
Matched ALEKS students vs. Matched Non-ALEKS students	72% vs 55%	+17	<0.001
Using IPTW			

To summarize, we can consider (in Table 13) the effect sizes for pass rates before and after matching, using both PSM and IPTW, for the both original and replication studies. When the comparison is conducted in this fashion, slightly higher Cohen's D values are obtained for IPTW than PSM (0.22 versus 0.20 for original studies; 0.36 versus 0.35 for replication study).

However, the same qualitative effect is obtained for both of these approaches (and indeed, before matching): students who used ALEKS performed better than students who did not use ALEKS.

Table 13

Effect Size for Pass Rates Before and After Matching for Original and Replication Studies, Using A Quasi-Experimental Comparison Based on IPTW Matching

Comparison	Original studies (Cohen's D)	Replication study (Cohen's D)
Before matching	0.28	0.44
PSM	0.20	0.35
IPTW	0.22	0.36

Discussion and Conclusions

Randomized Controlled Trials (RCTs) are considered by many to be the most rigorous study design due to randomization, i.e. randomly assigning subjects to treatment and control groups. Randomization, for sufficiently large sample sizes, is thought to remove selection bias. As a result, national-level policies in the United States treat RCTs as having the highest strength of evidence among study designs. In observational studies, by contrast, assignment to control groups and treatment groups is not random, and because of that, factors other than the impact of the treatment may confound the result. Though RCTs are reliable (if issues of implementation fidelity are properly accounted for), they are costly, time consuming. and increasingly are seen as raising ethical issues. These limitations can be addressed by instead designing quasi-experimental studies that utilize observed data for drawing causal relationships. In well-run

quasi-experimental studies, the researcher explicitly identifies key covariates and ensures that they are balanced between experimental groups (Tipton et al., 2014). This balancing of students can occur post-hoc, by identifying students from a broader population who match the students who participated in a specific treatment (Rosenbaum & Rubin, 1983).

In this paper, we demonstrate the potential of this approach, studying the efficacy of the learning system ALEKS for real-world learners in higher education, in a realistic setting where instructor and student decisions vary implementation. We conduct post-hoc quasi-experimentation using causal inference, taking a known treatment group, where there was not randomized assignment, and comparing it to a carefully selected comparison group, where covariates are matched between the two groups. We present two comparisons, one using data previously presented in a conference paper (Mojarad et al., 2018), and the other a replication study the following year. This second study represents a near-replication, conducted using the same system and protocol in the same university, but even in this near-replication, instructor implementation decisions were different, with substantially more instructors choosing to use the system in this year.

Our goal is to study whether students had higher pass rates when they used ALEKS than when they did not use ALEKS. In doing so, we investigate six different possible comparisons between ALEKS students and non-ALEKS students, in order to be confident that our results are not simply due to conducting a specific analysis. The first four of these comparisons, replications of the original studies, are designed to investigate whether a difference between ALEKS and non-ALEKS students is seen for different student breakdowns. These four comparisons involve samples that may be subject to some degree of selection bias, and hence, we conduct a fifth replication study where we use a quasi-experimental design and propensity

score matching, to create a fair comparison between ALEKS and Non-ALEKS students by balancing possible confounding attributes including age, gender, Accuplacer score and race. Based on recent criticisms of propensity score matching, we also conduct a sixth comparison, on both the original and replication studies, using an alternative statistical method, inverse probability of treatment weighting.

We find that ALEKS students appear to perform better than non-ALEKS students both in the original studies and in the replication study, across all six comparisons. This provides evidence that the previously reported findings hold in the later study as well – in other words, ALEKS's apparent effects on learning for this context seem to replicate.

Aside from providing data on statistical significance – all comparisons are statistically significant – we interpret the differences in terms of effect size (Cohen's D). In this study, we have reported the effect size of using ALEKS on pass rates for the two original and replication studies. While the effect size of using ALEKS in the original studies is 0.22, it is 0.36 in the replication study.

It is worth noting that ALEKS obtains higher effect sizes in the replication study, than in the original studies, which is not typically the pattern seen in replication research (Loken & Gelman, 2017). The higher effect size of using ALEKS in the replication study could be attributed to improved implementation (Pane et al., 2013), and higher participation rate. We observed that there has been a significant increase in the ALEKS adaption rate in Fall 2017 compared to Fall 2016 and Spring 2017. Comparing the original and replication studies, there has been a significant increase both in proportion of students in ALEKS sections and the proportion of students actually using ALEKS, though there is not actually a higher proportion of students adopting ALEKS *within* the ALEKS sections. These results indicate a higher adoption

rate by instructors, perhaps suggesting that instructors are choosing to adopt based on their colleagues' reports of using the system. However, there is not evidence that the factors influencing students to choose to use ALEKS have changed between years. One might expect the different population between studies to attenuate earlier effects, as ALEKS moves from use primarily by early adopters to a more general population of instructors, but instead the effect seems to be increasing. We cannot conclude at this point whether the change in effect size is due to a change in population or a change in implementation, as use matures at this institution.

We also investigate the balance across baseline characteristics between the treatment and control groups, for both statistical methods used, as part of better understanding the application of these two methods in evaluating the effectiveness of interactive learning systems. Balancing effectively between the two groups ensures that the selected treatment and control groups have similar confounder distributions and are drawn from similar populations. To investigate covariate balance, we report mean, standard deviation and the effect size for each covariate. We consider a Cohen's D of 0.1 or higher as a meaningful difference between the two groups, and hence an imbalance in the corresponding confounder across two groups, in line with recommendations in (Austin, 2009b).

Measuring the balance of attributes after matching using both IPTW and PSM, we observe that IPTW achieves better matching balance than PSM. In fact, PSM improves balance on some confounders relative to the original data, but actually worsens it slightly for other confounders. Since IPTW achieves better balance compared to PSM for both the replication and original studies, it is probably the more appropriate statistical method to use in this case. These results show an effect size of 0.23 for the original studies and 0.36 for the replication study.

Despite this general appearance of success for ALEKS, there are some limitations to the research presented in this paper. Despite our attempt to conduct the quasi-experiment study in the most rigorous manner, there are a few possible limitations that the current data cannot satisfy. The first possible limitation concerns the confounders we have used in this study to design balanced treatment and control groups. The set of confounders used here was reasonable and justified based on past research on the impact of these variables on the impact of curricula -- initial knowledge (Bright et al., 2008), age (Papastergiou, 2009) and whether students belong to a minority group (Stassen, 2003; Padgett et al., 2010). Nonetheless, other factors not available to the research team might have differentiated ALEKS users from non-ALEKS users in important ways, such as English Language Learner status, whether the students are first generation college students, national origin, parents' education, and high school GPA. In general, any quasi-experimental comparison is only as good as the data on confounders available to the researchers.

The second possible limitation arises from the fact that there are several frameworks and techniques in literature of causal inference. In this study we have used two common quasi-experimental design techniques, PSM and IPTW. Though the use of six different comparisons is fairly thorough compared to most papers published, it still might be valuable for future research to add additional comparisons beyond the six considered here, examining the impact of using other quasi-experimental designs. Some of the other popular quasi-experimental designs for causal inference are regression discontinuity designs, instrumental variable designs, and comparative interrupted time series designs (Kim & Steiner, 2016). There are also other matching approaches than IPTW or PSM (see King & Nielsen, 2019, for a review of other alternative statistical methods). There are, of course, a nearly limitless number of comparisons that can be made; the work in this paper represents a fairly broad set of comparisons, more than

are typically seen, but it is reasonable to consider making additional comparisons as well. Ultimately, comparing the data in multiple ways increases confidence that findings are genuine rather than spurious findings created through "p-hacking".

The third possible limitation is the relatively simple fashion in which usage was considered within this study. In this study we have defined ALEKS users as students who used ALEKS at least once, meaning that they at least completed the initial assessment in ALEKS. However, this is a very minimal definition of usage – some students included in this category completed no learning content within ALEKS. Although doing so goes substantially beyond the scope of the current study, in future studies it could be relevant to investigate the influence that time spent and usage patterns exert on learning outcomes (Pane et al., 2013). This analysis would help us to understand not just whether ALEKS is effective, but under what conditions.

In the end, we believe that no single scientific investigation in a complex field like education should be considered conclusive. This study shows that for multiple student cohorts with different adoption patterns, and for multiple possible post-hoc comparisons, students who use ALEKS succeed to a greater degree than students who do not use ALEKS, with the effect persisting across semesters, even as implementation conditions change. This finding suggests that ALEKS is likely to continue to be beneficial to learners in this community college, and comparable settings. While this finding should be investigated at longer durations still (i.e. use of ALEKS over the course of a decade by an institution), this paper's results represent promising evidence for ALEKS' longer-term usefulness. Overall, continued investigation of the efficacy of adaptive learning systems like ALEKS will be an important part of guaranteeing that adaptive learning continues to fulfill its promise for improving student learning outcomes.

References

Austin, P. C. (2009a). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine*, *28*(25), 3083–3107. https://doi.org/10.1002/sim.3697

Austin, P. C. (2009b). Using the Standardized Difference to Compare the Prevalence of a Binary Variable Between Two Groups in Observational Research. *Communications in Statistics - Simulation and Computation*, *38*(6), 1228–1234. https://doi.org/10.1080/03610910902859574

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

Baker, R., Wang, F., Ma, Z., Ma, W., & Zheng, S. (2018). Studying the effectiveness of an online language learning platform in China. *Journal of Interactive Learning Research*, *29*(1), 5-24.

Bhatt, R., & Koedel, C. (2012). Large-scale evaluations of curricular effectiveness: The case of elementary mathematics in Indiana. *Educational Evaluation and Policy Analysis*, *34*(4), 391-412.

Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using Covariates to Improve Precision for Studies That Randomize Schools to Evaluate Educational Interventions. *Educational Evaluation and Policy Analysis*, *29*(1), 30–59. https://doi.org/10.3102/0162373707299550

Bright, C., Bright, C., Lindsay, E., Lowe, D., Murray, S., & Liu, D. (2008). Factors that impact learning outcomes in both simulation and remote laboratories. *EdMedia + Innovate*

Learning, (1), 6251–6258. Retrieved from https://www.learntechlib.org/p/29248/

Cameron, S. V, & Heckman, J. (2001). The Dynamics of Educational Attainment for Blacks, Hispanics, and Whites. *Journal of Political Economy*, *109*(3). https://doi.org/10.3386/w7249

Cantwell, R., Archer, J., & Bourke, S. (2001). A comparison of the academic experiences and achievement of university students entering by traditional and non-traditional means. *Assessment & Evaluation in Higher Education*, *26*(3), 221-234.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Retrieved from https://6666957a-a-62cb3a1a-s-sites.googlegroups.com/site/ff07edownloadbooks/ff07/statistical-power-analysis-for-the-behavioral-sciences-2nd.pdf

Cook, T. D., & Payne, M. R. (2002). Objecting to the objections to using random assignment in educational research. In F. Mosteller & Boruch. Robert F. (Eds.), *Evidence Matters: Randomized Trials in Education Research*. Retrieved from https://books.google.com/books?hl=en&lr=&id=M9bosA7smQMC&oi=fnd&pg=PA150&dq=Objecting+to+the+objections+to+using+random+assignment&ots=Aro8r3TWLl&sig=FlDKRQ8f0fZDyWVr_kWLRDjwfJs#v=onepage&q=Objecting to the objections to using random assignm

Craig, S. D., Anderson, C., Bargagloitti, A., & Graesser, A. C. (2011). Learning with ALEKS: The Impact of Students' Attendance in a Mathematics After-School Program. *15th International Conference on Artificial Intelligence in Education (AIED 2011)*, *6738*(June), 435–437. https://doi.org/10.1007/978-3-642-21869-9

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, *210*, 2-21.

Doignon, J. P., & Falmagne, J. C. (2011). *Knowledge Spaces*. Berlin: Springer.

Fang, Y., Ren, Z., Hu, X., & Graesser, A. C. (2019). A meta-analysis of the effectiveness of ALEKS on learning. *Educational Psychology*, *39*(10), 1278-1292

Feng, M., Roschelle, J., Heffernan, N., Fairman, J., & Murphy, R. (2014). Implementation of an intelligent tutoring system for online homework support in an efficacy trial. *Proceedings of ITS 2014*, 561–566. https://doi.org/10.1007/978-3-319-07221-0_71

Goossens, H., Ferech, M., Vander, R., Md, S., Goossens, H., Ferech, M., … Elseviers, M. (2005). Outpatient antibiotic use in Europe and association with resistance: a cross-national database study. *The Lancet*, *365*(12), 579–587. https://doi.org/10.1016/S0140-6736(05)17907-0

Hattie, J. (2008). *Visible Learning*. https://doi.org/10.4324/9780203887332

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The Extent and Consequences of P-Hacking in Science. *PLoS Biology*, *13*(3). https://doi.org/10.1371/journal.pbio.1002106

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2006). *Empirical Benchmarks for Interpreting Effect Sizes in Research* (Vol. 2). Retrieved from http://www.ncaase.com/docs/HillBloomBlackLipsey2007.pdf

Kim, Y., & Steiner, P. (2016). Quasi-Experimental Designs for Causal Inference. *Educational Psychologist*, *51*(3–4), 395–405. https://doi.org/10.1080/00461520.2016.1207177

King, G., & Nielsen, R. (2019). Why Propensity Scores Should Not Be Used for Matching.

Political Analysis, 1–20. https://doi.org/10.1017/pan.2019.11

Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1997). Intelligent Tutoring Goes To School in the Big City. *International Journal of Artificial Intelligence in Education*, 8, 30-43.

Kroeze, K., Hyatt, K., & Lambert, C. (2015) Brain Gym: Let the user beware. *Journal of Interactive Learning Research, 26*(4), 395-401.

Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: a meta-analytic review. *Review of educational research*, *86*(1), 42-78.

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, *355*(6325), 582–584. https://doi.org/10.1126/science.aam5409

Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of educational psychology*, *106*(4), 901-908.

Massell, D., & Perrault, P. (2014). Alignment: Its Role in Standards-Based Reform and Prospects for the Common Core. *Theory into Practice*, *53*(3), 196–203. https://doi.org/10.1080/00405841.2014.916956

Mattern, K. D., & Packman, S. (2009, December 4). *Predictive Validity of ACCUPLACER Scores for Course Placement: A Meta-Analysis*. Retrieved from https://research.collegeboard.org/publications/content/2012/05/predictive-validity-accuplacer-scores-course-placement-meta-analysis

Mcmillan, J. H. (2007). Randomized Field Trials and Internal Validity: Not So Fast My Friend. *Practical Assessment Research & Evaluation*, *12*(1). Retrieved from https://pareonline.net/pdf/v12n15.pdf

Mitrovic, A., Suraweera, P., Martin, B., & Weerasinghe, A. (2004). DB-suite: Experiences with three intelligent, web-based database tutors. *Journal of Interactive Learning Research*, *15*(4), 409-432.

Mojarad, S., Essa, A., Mojarad, S., & Baker, R. S. (2018). Studying Adaptive Learning Efficacy using Propensity Score Matching. *Proceedings of the 8th International Conference on Learning Analytics and Knowledge (LAK18)*, 1–8. https://doi.org/10.1016/j.solmat.2011.03.007

Morgan, P. L., Frisco, M. L., Farkas, G., & Hibel, J. (2010). A propensity score matching analysis of the effects of special education services. *The Journal of special education*, *43*(4), 236-254.

Neumann, M., Wang, Y., Qi, G. Y., & Neumann, D. (2019). An Evaluation of Mandarin Learning Apps Designed for English Speaking Pre-schoolers. *Journal of Interactive Learning Research*, *30*(2), 167-193.

Padgett, R. D., Goodman, K. M., Johnson, M. P., Saichaie, K., Umbach, P. D., & Pascarella, E. T. (2010). The impact of college student socialization, social class, and race on need for cognition. *New Directions for Institutional Research*, *2010*(145), 99–111. https://doi.org/10.1002/ir.324

Pane, J. F., Griffin, B. A., McCaffrey, D. F., & Karam, R. (2013). Effectiveness of Cognitive Tutor Algebra I at Scale. *Educational Evaluation and Policy Analysis*, *36*(2), 127–144. https://doi.org/10.3102/0162373713507480

Papastergiou, M. (2009). Digital Game-Based Learning in high school Computer Science education: Impact on educational effectiveness and student motivation. *Computers &*

Education, *52*(1), 1–12. https://doi.org/10.1016/J.COMPEDU.2008.06.004

4Rao, J. N. K., & Scott, A. J. (1984). On Chi-Squared Tests for Multiway Contingency Tables with Cell Proportions Estimated from Survey Data. *The Annals of Statistics*, *12*, 46-60. https://doi.org/10.2307/2241033

Raudenbush, S. W. (2007). Designing field trials of educational innovations. *Scale up in education: Issues in practice*, *2*, 1-15.

Riehl, C. (2006). Feeling Better: A Comparison of Medical Research and Education Research. *Educational Researcher*, *35*(5), 24–29. https://doi.org/10.2307/3699784

Rivera, M. A., Davis, M. H., Feldman, A., & Rachkowski, C. (2017). An outcome evaluation of an adult education and postsecondary alignment program: the Accelerate New Mexico experience. *Problems and Perspectives in Management (Open-Access), 11* (4).

Rosenbaum, P. R. (1987). Model-Based Direct Adjustment. *Journal of the American Statistical Association*, *82*(398), 387–394.

Rosenbaum, P., & Rubin, D. (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

Rosenthal, R. (1979). The "File Drawer Problem" and Tolerance for Null Results. *Psychological Bulletin, 86*(3), 638.

Rubin, D. B., & van der Laan, M. J. (2008). Empirical efficiency maximization: improved locally efficient covariate adjustment in randomized experiments and survival analysis. *The International Journal of Biostatistics*, *4*(1), Article 5. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19381345

Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H., & Shavelson, R. J. (2007). *Estimating causal effects using experimental and observational design*. Washington, DC: American Educational Research Association.

Schofield, J. W. (1995). *Computers and classroom culture*. Retrieved from https://dl.acm.org/citation.cfm?id=526373

Silverman, S. L. (2009). From Randomized Controlled Trials to Observational Studies. *The American Journal of Medicine*, *122*(2), 114–120. https://doi.org/10.1016/j.amjmed.2008.09.030

Slavin, R. E. (2002). Evidence-Based Education Policies: Transforming Educational Practice and Research. *Educational Researcher*, *31*(7), 15–21. https://doi.org/10.3102/0013189X031007015

Slavin, R. E. (2017). Evidence-based reform in education. *Journal of Education for Students Placed at Risk (JESPAR)*, *22*(3), 178-184.

Stassen, M. L. A. (2003). Student Outcomes: The Impact of Varying Living-Learning Community Models. *Research in Higher Education*, *44*(5), 581–613. https://doi.org/10.1023/A:1025495309569

Sugai, G., & Horner, R. (2002). The evolution of discipline practices: School-wide positive behavior supports. *Child and Family Behavior Therapy*, *24*(1-2), 23-50. https://doi.org/10.1300/J019v24n01_03

Tipton, E., Hedges, L., Vaden-Kiernan, M., Borman, G., Sullivan, K., & Caverly, S. (2014). Sample Selection in Randomized Experiments: A New Method Using Propensity Score Stratified Sampling. *Journal of Research on Educational Effectiveness*, *7*(1), 114–135.

https://doi.org/10.1080/19345747.2013.831154

Torgerson, C. J., & Torgerson, D. J. (2001). The Need for Randomised Controlled Trials in Educational Research. *British Journal of Educational Studies*, *49*(3), 316–328. https://doi.org/10.2307/3122243

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, *46*(4), 197-221.

Vanlehn, K., Lynch, C., Schulze, K., Shapiro, J. A., Shelby, R., Taylor, L., ... & Wintersgill, M. (2005). The Andes physics tutoring system: Lessons learned. *International Journal of Artificial Intelligence in Education*, *15*(3), 147-204.

Wyse, A. E., Keesler, V., & Schneider, B. (2008). Assessing the effects of small school size on mathematics achievement: A propensity score-matching approach. *Teachers College Record*, *110*(9), 1879-1900.

Yamada, H., & Bryk, A. S. (2016). Assessing the first two years' effectiveness of Statway®: A multilevel model with propensity score matching. *Community College Review*, *44*(3), 179-204.

Author Note

We would like to thank Jenilyn Agapito and Alexandra Andres for assistance in preparing this document.