The effect of teachers reassigning students to new Cognitive Tutor sections

Adam C Sales
University of Texas College of Education
Austin, TX, USA
asales@utexas.edu

John F Pane
RAND Corporation
Pittsburgh, PA, USA
jpane@rand.org

ABSTRACT

The design of the Cognitive Tutor Algebra I (CTA1) intelligent tutoring system assumes that students work through sections of material following a pre-specified order, and only move on from one section to the next after mastering the first section's skills. However, the software gives teachers the flexibility to override that structure, by reassigning students to different sections of the curriculum. Which students get reassigned? Does reassignment hurt student learning? Does it help? This paper used data from the treatment arm of a large effectiveness study of the CTA1 curriculum to estimate the effects of reassignment on students' scores on an Algebra I posttest. Since reassignment is not randomized, we used a multilevel propensity score matching design, along with assessments of sensitivity to bias from unmeasured confounding, to estimate the effects of reassignment. We found that reassignment reduces posttest scores by roughly 0.2 standard deviations——about the same as the overall CTA1 treatment effect—that unmeasured confounding is unlikely to completely explain this observed effect, and that the effect of reassignment may vary widely between classrooms.

1. INTRODUCTION

Two closely related pillars of intelligent tutoring systems are sequencing and mastery learning. It has long been obvious that the sequence in which students learn different topics is an important component of a curriculum, due to prerequisites—for instance, students must master arithmetic in order to learn how to solve algebraic equations. A related example is scaffolding, in which learners gradually achieve independence over a sequence of problems; scaffolding "consists essentially of the adult 'controling' those elements of the task that are initially beyond the learner's capacity, thus permitting him to concentrate upon and complete only those elements that are within his range of competence" [33]. However, measuring the effects of sequencing [21] [9] and determining prerequisites or optimal sequences [29] [31] [17] remains an active area of research.

By "mastery learning," we mean the idea that students should "progress through topics as they master them," [22] as opposed to at a fixed pace. This typically results in students within the same classroom working on different parts of a curriculum at the same time.

The Cognitive Tutor Algebra I (CTA1) system [8] includes both features. A particular Algebra I curriculum is programmed into the software, so that students, if left alone, will encounter topics in a specific, intentional sequence. Mastery learning governs how they progress from one section to the next: an underlying knowledge tracing model estimates the probability students have mastered a set of pre-defined skills as they work through problems that incorporate those skills. Students ideally progress from one section to the next only after demonstrating mastery on the previous section's skills.

Mastery learning does not always proceed this way in the CTA1 software. After a student has worked a certain, prespecified number of problems in a particular section, he or she is automatically promoted to the next section, even if he or she has not mastered its skills [28]. Teachers can also reassign students working on one section to work on an entirely different section. If a teacher reassigns a student to a section other than the next one in the sequence, reassignment violates the intended sequencing as well as mastery learning.

There are a number of reasons teachers may want to meddle in the automatic progress of students through a curriculum [16]. If a teacher observes an advanced student spending time on basic skills, the teacher may move the student to more advanced sections. If certain skills will be on a standardized test, and a teacher wants all students to have had exposure to those skills before the test, the teacher may reassign all of his or her students to work on a section covering those skills. If a teacher notices a student falling behind his or her peers in the classroom, the teacher may choose to reassign the student to the section that the rest of the class is working on, even if the student has not demonstrated mastery on prerequisite skills (at least, within the tutor). If a teacher disagrees with the method a certain CTA1 section employs in teaching an Algebra topic, the teacher may reassign students out of that section, perhaps to the next unit or section in the curriculum.

It is unclear whether reassignment benefits students. On the

one hand, it violates the design principles of the software. On the other hand, it allows teachers flexibility to teach the material as they see fit, and use the tutor to meet the particular needs of their classrooms.

This paper uses data from a large randomized trial of the CTA1 curriculum to estimate the effect of reassignment. Unfortunately for our purposes, reassignment itself was not randomized—the study was designed to estimate CTA1's effectiveness, so access to the tutor was randomized instead. Still, log data from study participants includes data on how often each student was reassigned from one section to another, and posttests measure their algebra skills at the end of the study. For those reasons, this data provides a rare opportunity to measure the effect of reassignment, and, by extension, the (joint) importance of topic sequencing and mastery learning.

The following section gives background on the effectiveness trial and describes the data we will use for the study. Section 3 describes propensity score matching, the method we employ. Section 4 describes the propensity score models, which in turn describe characteristics of students who are reassigned. Section 5 describes the matching algorithm and covariate balance. Section 6 gives our main results on the effects of reassignment, including sensitivity analysis to confounding from unmeasured covariates and between-classroom effect heterogeneity. Section 7 concludes.

2. DATA: THE RAND CTA1 EFFECTIVENESS STUDY

In the years 2007–2010, the RAND Corporation conducted a randomized study to test the effectiveness of the CTA1 curriculum relative to business as usual. The study tested CTA1 under authentic, natural conditions—that is, oversight and support of CTA1's use was the same as it would have been outside of an RCT. The study population consisted of over 25,000 students in 73 high schools and 74 middle schools located in 52 diverse school districts in seven states. Students in Algebra I classrooms in participating schools took an algebra I pretest and a posttest, both from the CTB/McGraw-Hill Acuity series. The pretest was the Algebra Readiness Exam, a 40-item multiple-choice exam testing students' algebra I prerequisite skills. The posttest was the Algebra Proficiency Exam, a 32-item multiple-choice exam testing algebra I skills including solving equations for an unknown, graphing linear and quadratic functions, calculating complex algebraic expressions and other skills. Data from both exams were scored with a three-parameter item response theory (IRT) model.

Results [19] were reported separately for middle and high schools, in the first and second years of implementation. In the first year, estimated effects were close to zero in middle schools and slightly negative in high schools, with confidence intervals including negative, null, and positive effects in both cases. In the second year, estimated effects were positive—roughly one fifth of a standard deviation—in both middle and high schools, and were statistically significant in high schools. In the high school sample, the difference between the effects in the first and second years was statistically significant as well.

Table 1: The number and percent of students in each study year of the dataset who were never reassigned, or reassigned once, twice, three times, or four or more times

		# Reassignments				
year		0	1	2	3	4+
1	n	1621	552	133	43	34
	%	68	23	6	2	1
2	n	1056	297	193	95	194
	%	58	16	11	5	11

As part of the study, RAND collected basic demographic data from students, including gender, race/ethnicity, prior standardized test scores, and special education, free or reduced-price lunch, and English language learner status.

Carnegie Learning collected computer log data from most users in the treatment arm of the study. At the problem level, this dataset records which problems students attempted, along with timestamps and the numbers of hints and errors for each attempted problem. The dataset also contains data on which sections of CTA1 students attempted, and the result: whether the student mastered the section, was promoted automatically without mastery, was reassigned by the teacher to a new section, or stopped using the tutor altogether midway through the section.

The current study analyzes data from the high school treatment group only, assessing the effect of teachers reassigning students from one CTA1 section to another. Since students in the control arm of the study did not have access to the tutor, section reassignment is not relevant for them. We focus on high school, as opposed to middle school, since the characteristics of Algebra I students tend to differ between the two levels: 8th-grade students only take Algebra I if they are sufficiently advanced, whereas most 9th grade students (who have not taken it already) take Algebra I regardless. Thus, the high school sample was not only larger but also more broadly representative than middle school sample.

Unfortunately, log data was not available for every student in the treatment arm of the study, primarily for two reasons: some students in CTA1 schools nevertheless did not use the tutor, and some students used the tutor but their log data was irretrievable or could not be reliably linked to posttest scores and covariates. This study omitted schools in which data was missing for over 20% of students in either year, leaving 18 schools. Among the students at these schools, we omitted 164 who had no log data, and 242 who worked—but did not complete—only one section or who had no section completion data for some other reason. A total of 4,218 students in 282 classrooms remained in the analysis sample, roughly 70% of the full treatment group.

Table 1 shows the number of included students in each year of the experiment who were reassigned zero, one, two, three, or four or more times. Since the sample size decreases quickly with the number of reassignments, and for the sake of simplicity, we chose to dichotomize reassignment, esti-

mating the effect of being reassigned at least once versus never.

3. STATISTICAL APPROACH

For subjects $i = 1, \ldots, N$ in the treatment arm of the CTA1 trial, let Y_i denote subject i's posttest score, and let $Z_i \in \{0, 1\}$ indicate whether i was ever reassigned. Following [18] and [25], let y_i^0 and y_i^1 denote i's posttest score were $Z_i = 0$ or 1—i.e., had i not been reassigned, or had i been reassigned, perhaps counterfactually—and let $\tau_i = y_i^1 - y_i^0$ be the effect of reassignment on i's posttest score. Since y_i^1 and y_i^0 are never simultaneously observed, τ_i is unidentified; however, weighted average treatment effects of the form $\tau^w = \sum_i w_i \tau_i$, with $w_i \geq 0$ and $\sum_i w_i = 1$ may be identified under the right causal assumptions. For instance, had Z been randomized, the average treatment effect, τ^w with $w_i = 1/N$, could be estimated without bias by the difference in the mean of Y between subjects with $Z = 1$ and with $Z = 0$. Of course, reassignment Z was not random, so identifying average treatment effects requires some combination of control for observed covariates and assumptions about unobserved covariates.

Let $\boldsymbol{x}_i$ denote a vector of covariates for subject i. These include pretest scores, special education, gifted, and English language learner (ELL) status, race/ethnicity (white, black, Latinx[1]), received free or reduced-price lunch (FRL). Let $Class_i$ be i's classroom; since reassignment occurred within classrooms, $Class$ is a covariate as well. If reassignment were randomly assigned, the (theoretical) distribution of $\boldsymbol{x}$ and $Class$ would be equal between reassigned and not-reassigned students—$\boldsymbol{x}$ and $Class$ would be balanced. Our strategy will be to construct a randomization scheme in which $\boldsymbol{x}$, and, to the extent possible, $Class$ are balanced, and conduct inference under that randomization scheme.

Specifically, we use propensity score matching [23] [27]. The propensity score for subject i, $e_i(\boldsymbol{x}_i, Class_i) = Pr(Z_i = 1|\boldsymbol{x}_i, Class_i)$ is the probability of i being reassigned conditional on covariates $\boldsymbol{x}$ and classroom. [24] showed that under two conditions, described below, estimates of the average treatment effect conditional on $e(\boldsymbol{x}, Class)$ are unbiased. To estimate effects, we first estimate propensity scores (Section 4), then identify groups of reassigned and not-reassigned students with similar estimated propensity scores—a "match"—and verify that covariates are sufficiently balanced within the matched sample (Section 5), and, finally, estimate effects within the matched sample 6.

The first condition for propensity score matching is that there is some randomness in the treatment assignment:

$$0 < e_i(\boldsymbol{x}_i, Class_i) < 1 \text{ for all } i. \quad (1)$$

When (1) fails for a subset of the analysis sample, common practice is to drop that subset and estimate average effects for the remainder of the analysis sample, i.e. the subset for which (1) holds; this subset is referred to as the "region of common support" [4] [30]. In this study, including $Class$ among the covariates leads to violations of (1). Of the 282 classrooms over the two years of the study, 95 contained no reassigned students, and in 52 classrooms every student was reassigned at least once. In this subset of the data, including 44% of students, $Pr(Z = 1|Class) = 0$ or 1. Our solution is to drop classrooms in which no one or everyone was reassigned, and only estimate effects for students in classrooms with some reassignment variance, a student-level analysis.

We attempted a parallel classroom-level analysis, in which we matched classrooms in which all students were reassigned to classrooms in which no one was. However, we were unable to construct a match with adequate covariate balance (there were few no-reassigned classrooms with similar mean pretest scores to the all-reassigned classrooms that were of similar sizes). For that reason, we dropped the classroom-level analysis.

The second condition for propensity score matching is that there are no unmeasured confounders:

$$(y^1, y^0) \perp\!\!\!\perp Z|\boldsymbol{x}, Class \quad (2)$$

Assumption (2) is well known as the Achilles heel of causal inference outside of RCTs. (2) is untestable; its believability depends on what is understood about the process that underlies treatment assignment Z, and what covariates are available for control. In our case, reassignment is poorly understood, and appears highly idiosyncratic [16]. Fortunately, our study includes a pretest measure, and observational studies controlling for pretest scores tend to perform well, and replicate experimental estimates [6] [7]. Section 6.1 discusses a sensitivity analysis that relaxes 2 and assumes reasonable levels of unmeasured confounding.

Our attitude towards propensity score matching is agnostic. If the propensity score models in the following section were approximately correct, and yielded good estimates of the true propensity scores, then the theory underlying propensity score adjustment holds. If not, the process of propensity score matching may still result in a set of matched reassigned and not reassigned students that, on average, resemble each other on all measured covariates. In other words, the (mis)estimated propensity scores $\hat{e}$ may still be approximate "balancing" scores, satisfying

$$\boldsymbol{x} \perp\!\!\!\perp Z|\hat{e}. \quad (3)$$

Causal inference based on comparisons within these matched sets will still be plausible; indeed, [24] showed that in order to estimate average treatment effects, it is sufficient to condition on a balancing score, rather than the propensity score itself.

Following that logic, we choose propensity score models, and matching schemes based on the fitted models, in order to satisfy (3). Since posttest scores play no role in propensity score estimation and matching, the process may be iterative without affecting the objectivity of the final causal estimate. That is, we may try a series of candidate propensity score models and matches, and choose the one that results in the best covariate balance. Only then do posttests enter the picture, so that we may estimate effects.

[1]For the sake of parsimony, these categories were collapsed from a larger set in the original dataset, so that 8 American Indian/Alaskan Native students were categorized as Latinx, 23 Asian/Pacific Islander students and 118 students with missing data were categorized as white, and 22 Other/Multiracial students were categorized as black.

All data analysis was done in `R` [20] using the `tidyverse` suite of packages [32] for data manipulation, plotting, and other tasks. This document was produced dynamically with `knitr` [34]. Source code is available at www.github.com/[Redacted]

4. PROPENSITY SCORES: WHO GETS REASSIGNED?

We use multilevel logistic regression [10] to estimate student level propensity scores. The multilevel regression accounts for the nesting of students within classrooms, classrooms within teachers, and teachers within schools. In constructing the model, we give special consideration to the role of pretest scores, a proxy for student mathematical ability at the beginning of the school year, in predicting reassignment. First, we decompose pretest scores into student- and classroom-level components. If w_i is student i's pretest score, let $w_i = \bar{w}_{j[i]} + \tilde{w}_i$, where $\bar{w}_{j[i]}$ is the average pretest score in i's classroom $j[i]$, and $\tilde{w}_i$ is the difference between i's pretest score and the classroom mean. This decomposition was motivated by the possibility that reassignment patterns may differ between high- and low-achieving classrooms, and that a teacher's decision to reassign a student depends on the student's ability relative to the classroom than his or her absolute ability. Second, we modeled the effect of $\tilde{w}$ on Z as linear in the logit scale, but allowed the slope to vary by classroom. This was motivated by the possibility that some teachers use reassignment to help struggling students catch up to their peers, so lower $\tilde{w}$ would predict Z, and other teachers use it to help high-achievers skip sections related to basic skills, so higher $\tilde{w}$ would predict Z. We also considered models incorporating non-linear effects of $\tilde{w}$, via natural splines [14] but found no evidence that the non-linearity improved the model fit. We fit the model using the `lme4` package in `R` [1].

All in all, the propensity score model was:

$$\begin{aligned} &logit\left\{Pr(Z_i = 1|\boldsymbol{x}_i, Class_i = j)\right\} = \\ &\quad \beta_{0state[i]} + \beta_1\tilde{w}_i + \beta_2\bar{w}_{j[i]} + \\ &\quad \beta_3 Black_i + \beta_4 Latinx_i + \beta_5 Male_i + \\ &\quad \beta_6 Freshman_i + \beta_7 SpEd_i + \beta_8 gifted_i + \\ &\quad \beta_9 ESL_i + \beta_{10} FRL_i + \beta_{11} FRLmis_i + \beta_{12} year_i + \\ &\quad \gamma_{j[i]}\tilde{w}_i + \epsilon^{Cls}_{j[i]} + \epsilon^{Teach}_{k[i]} + \epsilon^{Schl}_{l[i]} \end{aligned} \tag{4}$$

where $logit(x) = log(x/(1-x))$ is the logit function, $\beta_{0state[i]}$ is a (fixed) intercept for each state in the sample, $FRLmis_i$ is an indicator for missing data in FRL (which was mode-imputed), and $year_i = 1, 2$ is the study year for subject i.

Finally, $\gamma_{j[i]}$, $\epsilon^{Cls}_{j[i]}$, $\epsilon^{Teach}_{k[i]}$, and $\epsilon^{Schl}_{l[i]}$ are random effects. The subscripts j, k and l refer to classroom, teacher, and school, respectively; the $[i]$ refers to student, so that $j[i]$ is i's classroom, $k[i]$ is i's teacher, and $l[i]$ is i's school.

$\gamma_{j[i]}$ is a random slope for $\tilde{w}_i$, varying at the classroom level. This is essentially an interaction term, allowing the slope for (classroom centered) pretest scores to vary from one classroom to the next. However, unlike standard regression interactions, random slopes are modeled as being drawn from a normal distribution, with a standard deviation estimated from the data. This is a form of regularization, shrinking the classroom-level slopes towards a common value, and allowing stable estimation even with very few observations from each classroom [10] [26]. The set of random slopes γ_j has a mean of zero—the average slope across classrooms is the fixed intercept β_1. Therefore, the slope for pretest in classroom j is $\beta_1 + \gamma_j$.

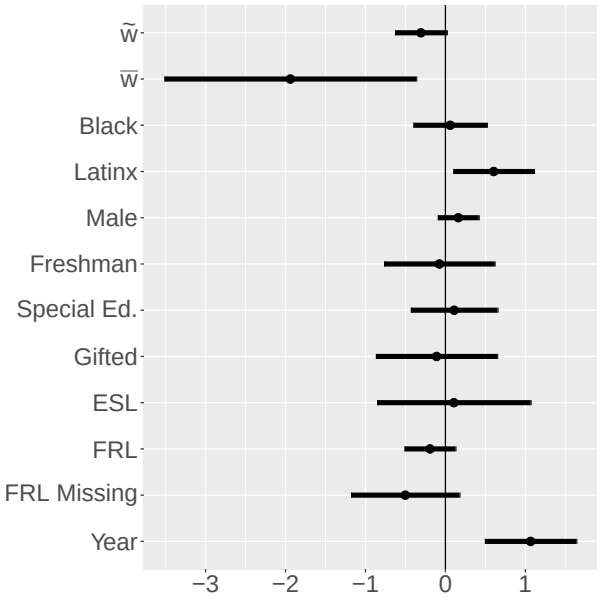

Figure 1: Estimated coefficients and 95% confidence intervals for student and class-level covariates from model (4)**.**

ϵ^{Cls}_j, ϵ^{Teach}_k, and ϵ^{Schl}_l are random intercepts for classroom, teacher, and school. These were also modeled as normal with a mean of zero and a standard deviation estimated from the data. Including them in the regression accounts the fact that two students in the same classroom or with the same teacher or in the same school may be more likely to have the same Z—either both be reassigned or neither—than two students in different classrooms, with different teachers, or in different schools.

Figure 1 gives estimated coefficients and 95% confidence intervals for the propensity score model (4). Reassignment was much more prevalent in the second year of implementation than in the first, and classrooms with low average pretest scores reassigned students more often—though the magnitude of this trend is hard to determine, ranging from moderate to very large (the coefficients for $\bar{w}$ and $\tilde{w}$ were scaled by the standard deviations of these variables in the data). Latinx students were reassigned more often than their White classmates.

Students with lower pretest scores were reassigned more frequently than their classmates with higher scores. However, this may vary by classroom. On average, classroom-specific β_{1j} was approximately -0.31 standard deviations, but the 95% confidence interval for the mean includes slightly positive values as well. The standard deviation of β_{1j}, varying by classroom, was estimated as 0.83, suggesting that in some classrooms the slope on $\tilde{w}$ was moderately positive, and in

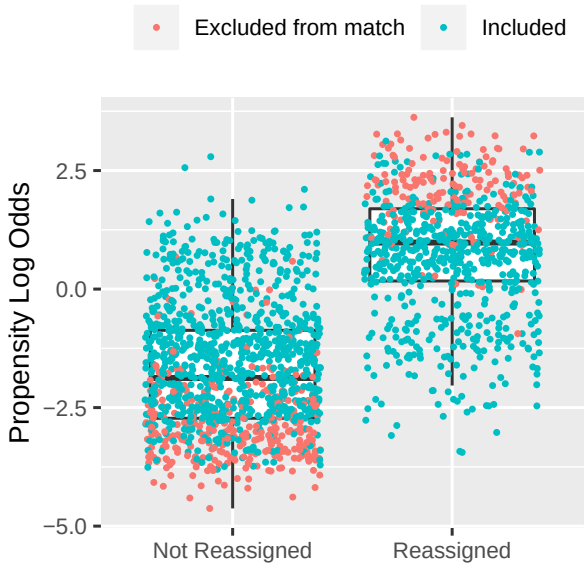

Figure 2: Estimated propensity scores for reassigned and not-reassigned students. Scores for students who were excluded from the ultimate match are colored red.

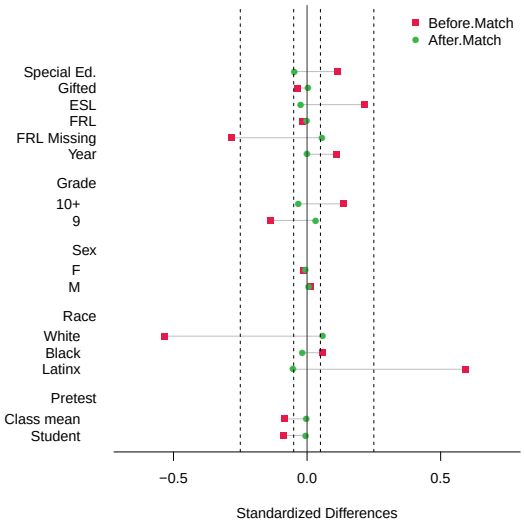

Figure 3: Covariate balance (standardized differences) before and after matching, for student level data. Dotted lines indicate standardized differences of 0.25 and 0.05, following the What Works Clearinghouse standards.

others it was negative. However, the model was not able to estimate the variance of β_{1j} precisely; the p-value testing the null hypothesis of zero variance was 0.07.[2] When model (4) was modified so that β_1 was not allowed to vary by classroom, it was estimated as -0.32±0.27.

5. MATCHING AND COVARIATE BALANCE

We construct a student-level match based on propensity scores on the log-odds scale, i.e. $log(\hat{e}/(1-\hat{e})$. Instead of a pair-matching design, which would necessitate discarding non-reassigned students who would make good matched comparisons, we use a restricted full match design [11]. In this design, the numbers of reassigned and not-reassigned students in each matched set is allowed to vary, so that in some cases several reassigned students may be matched with a single non-reassigned student, and vice-versa. We use the `R` package `optmatch` [13] to choose the matched sets optimally. The `fullmatch()` routine takes a matrix of discrepancies (e.g. differences in propensity scores) between treatment and control subjects, and arranges them into matched sets so that the sum of absolute discrepancies between matched subjects is minimized.

As described at the end of Section 3, the post-test scores played no role in this process. Hence, we were able to iteratively match students, check covariate balance, modify the propensity score model and/or the matching routine if necessary, and repeat until adequate balance was achieved. Here we present the final match; a record of attempts is available on the first author's github site.

[2]This hypothesis was tested with a likelihood ratio χ^2 test comparing (4) to a model in which β_1 did not vary by classroom.

The initial full match based on the log-odds propensity scores yielded decent covariate balance. However, pretest scores were slightly unbalanced, and since we consider pretest to be the most important covariate, we decided to match on the Mahalanobis distances between reassigned and not reassigned students combining propensity scores and pretest scores. Additionally, as displayed in Figure 2, the distributions of propensity scores among reassigned and not-reassigned students do not entirely overlap. Although this is at least partially due to overfitting the propensity score model (4), matching students with highly discrepant propensity scores may hinder the believability of the result. Hence, in our final match we imposed a caliper of 0.3 pooled standard deviations of the Mahalanobis distances. This prevented students with very different pretest scores or propensity scores to be matched. On the other hand, matches were unavailable for 25% of the students in the sample (21% of reassigned students and 28% of not-reassigned students). Propensity scores for these students are colored red in Figure 2. Our effect estimates pertain only to the remaining 75% of students—all in all, 1480 students, 604 reassigned and 876 not reassigned.

Covariate balance after matching was excellent. Figure 3 and Table 2 give covariate balance (standardized differences) before and after matching. They were produced with the `RItools` package in `R` [3]. Before matching, several covariates were unbalanced, especially race. Table 2 shows stars reflecting p-values from individual covariate balance tests; nearly all covariates were unbalanced at the $\alpha = 0.1$ level. An omnibus balance test [12] gives $p < 0.001$. Figure 3 shows, as benchmarks, standardized differences of $\pm$ 0.25 and 0.5, corresponding to thresholds given in the What Works Clearing-

Table 2: Balance (standardized differences) on student level covariates before and after propensity score match. Omnibus p-values testing covariate balance are p<0.001 before matching and p=0.95 after matching.

	Before Match		After Match
	std.diff		std.diff
Pretest			
Class Mean	-0.09	.	0.00
Class Centered	-0.09	.	-0.01
Race/Ethnicity			
White	-0.53	***	0.06
Black	0.06		-0.02
Latinx	0.59	***	-0.05
Sex			
F	-0.01		-0.01
M	0.01		0.01
Grade			
10+	0.14	**	-0.03
9	-0.14	**	0.03
Special Ed.	0.12	*	-0.05
Gifted	-0.04		0.00
ESL	0.22	***	-0.02
FRL	-0.02		0.00
FRL Missing	-0.28	***	0.06
Year	0.11	*	0.00

house (WWC) handbook[3] [5]. Before matching, imbalances in race and FRL missingness exceeded 0.25, and most other imbalances were greater than 0.05.

Matching improved nearly all of these imbalances. Most importantly, pretest measures were nearly exactly balanced. None of the individual covariate balance tests was significant at the 10% level or had standardized differences greater than 0.25, and, with the exception of race, and FRL missingness none of the covariates was imbalanced with a standardized difference greater than 0.05. The omnibus p-value testing overall balance was 0.95.

The match also balanced classroom indicators. Before matching, the omnibus p-value testing balance of classroom indicators was < 0.001; after matching it was 0.99.

6. THE EFFECT OF REASSIGNMENT

Table 3 gives five estimates for the effect of reassignment in classrooms where some students, but not all, were reassigned at some point. The first column gives the estimate itself, the second gives the sample size N for that estimate, the third, "Std Error" gives the standard error, and the fourth, "CI," gives a 95% confidence interval. The last two columns contain sensitivity analyses, described in the following section. All the estimates used a regression routine from the `esti-matr` package in `R` [2], with "HC2" heteroskedasticity-robust standard errors.

The first row, labeled "Raw," is an an unadjusted estimate, comparing all students in the sample who were reassigned to all students who weren't. There is little difference in their average posttest scores.

The next row, labeled "Matched+Regression," gives the effect estimate based on the match from Section 5. The lower sample size 1480 reflects the fact that some students were excluded from the match; this estimate only pertains to those who were included. To estimate the effect, we regress posttests on Z including a fixed effects for each match. Let $\hat{\tau}_m$ be the estimated effect in match m. If m is a pair—one reassigned student matched with one non-reassigned student—then $\hat{tau}_m$ is the difference between the two students' posttest scores. If there are more than two students in the match, $\hat{\tau}_m$ is difference in posttest means between reassigned and not-reassigned students within matched-set m. If treatment assignment is unconfounded within each match, $Z \perp\!\!\!\perp \{y_C, y_T\}|match$, then $\hat{\tau}_m$ is unbiased for the average effect of Z on posttest scores in match m. Then the regression estimate is a weighted average of $\hat{\tau}_m$, with weights $w_m \propto (1/n_{1m} + 1/n_{0m})^{-1}$; this weighing scheme minimizes the standard error under standard linear regression assumptions (if the regressions assumptions do not hold, but Z is still unconfounded within the match, then the estimate is still unbiased but the weights are sub-optimal).

The next row, labeled "Match+Regression" uses the same regression model as the "Matched" estimator, but additionally controls for pretest scores (with a natural spline with five degrees of freedom), and indicators for special education status, missing free or reduced-price lunch data, and race. This strategy controls for differences in these covariates left over after the match, accounting for the fact that the match was imperfect.

The "Matched" and "Match+Regression" estimates were almost identical—effect sizes of -0.2 and -0.19, respectively, with 95% confidence intervals of [-0.29,-0.12] an [-0.28,-0.11]. These negative effect estimates suggest that reassignment hurts student learning. The effect size of a fifth of a standard deviation is roughly the same as the overall average effect of CTA1 in high schools in the second year of implementation, as estimated in [19], suggesting that reassignment may negate most of the positive effect of using CTA1.

The next two rows of Table 3, however, suggest that the effect of reassignment may depend on context. Each row uses the "Match+Regression" approach, but separately in data from implementation years 1 and 2. It appears that reassignment may have hurt students' posttest scores more in the first than in the second year of implementation—in the first year, we estimate an effect of -0.24 and in the second year we estimate an effect of -0.11. That said, the difference between the two effects is not itself statistically significant—that is, it may be the result of statistical noise.

The final row of Table 3, labeled "Within-Class," uses a different confounder control strategy altogether. This estimate matches students by classroom, as if reassignment were ran-

[3]In the context of a randomized experiment with attrition, covariate imbalances with standardized differences greater than 0.25 invalidate a study, whereas differences between 0.05 and 0.25 require statistical adjustment and differences less than 0.05 are acceptable as is.

	Estimate	N	Std. Error	CI	[Pretest]	[State]
Raw	-0.04	1981	0.03	[-0.11,0.03]	[-0.15,0.07]	[-0.16,0.08]
Matched	-0.20	1480	0.04	[-0.29,-0.12]	[-0.35,-0.06]	[-0.36,-0.05]
Match+Regression	-0.19	1480	0.04	[-0.28,-0.11]	[-0.33,-0.05]	[-0.34,-0.04]
Year 1	-0.24	1008	0.06	[-0.34,-0.13]	[-0.42,-0.06]	[-0.43,-0.04]
Year 2	-0.11	472	0.07	[-0.25,0.02]	[-0.33,0.11]	[-0.35,0.12]
Within-Class	-0.17	1981	0.04	[-0.24,-0.10]	[-0.29,-0.05]	[-0.30,-0.04]

Table 3: Estimates of the effect of reassignment without controlling for confounding ("Raw"), controlling for confounding with propensity score matching ("Matched"), with matching and further regression adjustment ("Match+Regression"), overall and separately for each year, and matching by classroom, with further regression adjustment ("Within-Class"). The table gives estimates, standard errors, 95% confidence intervals, and 95% sensitivity intervals assuming an unobserved confounder with properties similar to pretest scores ("[Pretest]") and to State ("[State]")

domized within classrooms. To weaken that assumption, the "Within-Class" estimate incorporates additional regression controls: a natural spline with five degrees of freedom for pretest, and indicator variables for the remaining covariates. This strategy estimates a similar negative effect as the others, NA, with a 95% confidence interval of [-0.24,-0.10].

6.1 Unobserved Confounding

The estimates in Table 3 all assumed (2), that there was no unobserved confounding. This assumption is strong, untestable, and could undermine all of the inference in Section 6. For instance, the estimated negative effect may be due to baseline differences in ability, beyond what is captured in pretest scores.

[15] suggest a method of estimating the sensitivity of a regression to an omitted confounder based on benchmarking from observed confounders. Roughly speaking, the idea is to widen the confidence interval from an ostensibly causal linear model to account for the possibility of a hypothetical unmeasured confounder, U, that predicts reassignment and posttests to the same extent as one of the observed covariates. These "sensitivity intervals" account for uncertainty from two sources: random error, and systematic error due to the omission of a confounder.

In order to confound the causal relationship between reassignment and posttests, a confounder would have to predict both. Capturing these two requirements, the method of [15] is based on two sensitivity parameters: first, T_Z encodes the extent to which U predicts Z, after accounting for observed covariates $\boldsymbol{x}$. Formally, T_Z is the t-statistic on the U coefficient from an ordinary least squares regression of Z on U and $\boldsymbol{X}$. The second parameter is ρ^2, the squared partial correlation between posttest scores and U, conditional on $\boldsymbol{x}$. Of course, since U is unobserved, neither T_Z nor ρ^2 is known; [15] suggest benchmarking them using observed covariates. That is, imagine each observed covariate, in turn, were unobserved, and calculate its T_Z and ρ^2 given the rest of the observed covariates.

Table 3 includes two such sensitivity intervals. The column labeled "[Pretest]" includes sensitivity intervals for an unobserved confounder that predicts reassignment and posttests as well as do pretest scores—typically the most important confounder. That is, these intervals are 95% confidence intervals that assume the possible existence of an unmeasured covariate as important as pretest. It turns out, in the current analysis, that omitting state indicators would cause more bias than omitting pretest scores; for that reason, the column labeled "[State]" gives sensitivity intervals for an unobserved confounder that predicts reassignment and posttest scores as well as state indicators. Both sets of sensitivity intervals are considerably wider than the corresponding confidence intervals, including both large and small negative effects. Sensitivity intervals for the "Matched", "Match+Regression," "Year 1," and "Within-Class" estimates, whose confidence intervals excluded zero, excluded zero as well. That is, confounding from an unobserved variable as important as pretest or state may have led us to over-estimate the negative effect of reassignment; it may have also led us to under-estimate the effect. However, such confounding cannot explain the sign of the effect we estimated—even assuming the existence of an unobserved confounder as important as our most important covariates, the effect must be negative.

That said, an even stronger confounder, or more complex confounding from several unobserved covariates, may explain the observed results. Without a randomized trial, it is impossible to entirely rule out unobserved confounding.

6.2 Treatment Effect Heterogeneity

Previous research [16] has found evidence for a wide variety of uses for reassignment. In some cases, teachers reassign students who are falling behind their classmates, in other cases teachers reassign nearly the entire class to work on a particular section of the tutor, and in other cases teachers will simultaneously reassign all students working on a particular section *out* of that section. Along similar lines, our (inconclusive) evidence for variance between classrooms in the relationship between pretest scores and the probability of a student being reassigned points towards varying uses for reassignment.

If reassignment is used differently from classroom to classroom, it stands to reason that it might have different effects in different classrooms, as well. To test that assumption, we fit a multilevel model with random effects for reassignment, varying by classroom. The model had the same fixed effects as model underlying the "Match+Regression" results described above, as well as random intercepts for classroom and random slopes for reassignment, varying by classroom.

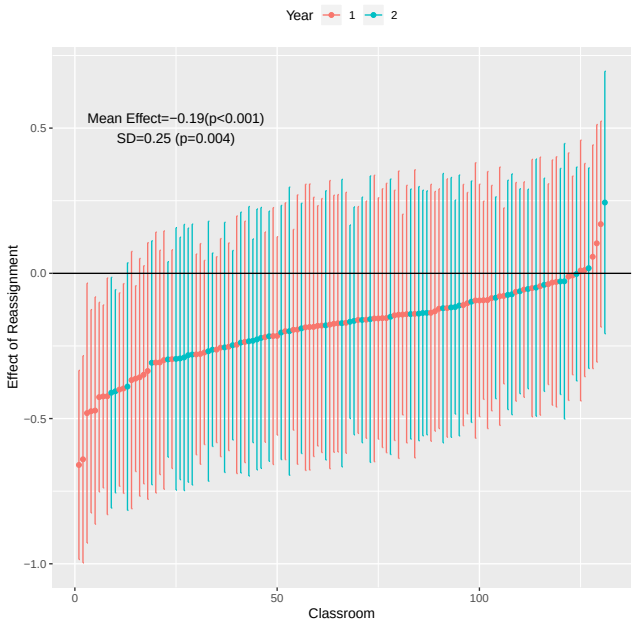

Figure 4: Classroom-specific effects of reassignment ($\hat{\beta}_1 + \hat{\gamma}_j$) from model (5). Error bars represent standard errors.

Formally, the model is:

$$\begin{aligned} Posttest_i = \beta_{0,m[i]} + \beta_1 Z_i + \beta_2 SpEd_i + \\ \beta_3 frlMIS_i + \beta_4 Black_i + \beta_5 Hisp_i + \\ ns^5(pretest_i, \boldsymbol{\alpha}) + \gamma_{j[i]} Z_i + \epsilon^{Cls}_{j[i]} + \epsilon^{Ind}_i \end{aligned} \quad (5)$$

where $\beta_{0,m[i]}$ is a fixed intercept for each match, $ns^5(pretest_i, \boldsymbol{\alpha})$ is a natural spline for pretest, with five degrees of freedom and coefficient vector α, and $\gamma_{j[i]}$, $\epsilon^{Cls}_{j[i]}$, and ϵ^{Ind}_i are random effects, modeled as normal with mean zero and standard deviation estimated from the data. Symbols α, β, γ, and ϵ do not represent the same quantities as in equation (4). $\gamma_{j[i]}$ is the random slope for reassignment, varying by classroom; the effect of reassignment in classroom j is estimated as $\hat{\beta}_1 + \hat{\gamma}_j$. That is, β_1 represents the effect of reassignment, averaged over all classrooms, and γ_j represents the difference between classroom j's effect and the average. While precisely estimating the effect of reassignment in any particular classroom is beyond the scope of our data, this model allows us to estimate the variance across those effects, as the variance of γ_js.

The results are displayed in Figure 4. The effect of reassignment in an average classroom is estimated as similar to the effects in Table 3. This effect varies with a standard deviation of approximately 0.25. To test for between-classroom variance, we compared the fit of the multilevel model to an analogous model without random slopes, with a likelihood ratio χ^2 test; the p-value was 0.004. This standard deviation is large enough to imply that the effect will be positive in some classrooms—indeed, Figure 4 shows a number of classrooms with positive effects. That said, the confidence intervals (based on estimates for the conditional variance of random slopes, combined with the standard error of the main effect of reassignment) are all rather wide and nearly all contain zero.

Therefore, while the effect of reassignment was negative, on average, it may have been positive in some classrooms.

This variation could be due to a number of factors, including differences in the composition of classrooms and in when or how reassignment is used. We considered two simple hypotheses about classroom-level predictors of heterogeneous treatment effects. The first hypothesis was that variance in students' pretest scores within a classroom predicts the effect of reassignment in that classroom. The idea is that some teachers may use reassignment as a tool to address varying student ability—for instance, they may reassign lagging students to help them keep up with their classmates. Classrooms with higher variance in pretest scores afford more opportunities for teachers to use this reassignment strategy. If the strategy is widely used, and either particularly effective or ineffective at boosting students' posttest scores, there will be a correlation between classroom-level variance in pretest scores and the effect of reassignment.

Our second hypothesis was that the proportion of students in a classroom who have been reassigned may predict classroom-level effects. The idea here is that in classrooms with a low proportion of students reassigned, teachers use reassignment in a more targeted fashion, so it may be more beneficial.

Figure 5 plots random effects $\hat{\gamma}_j$ from model (5) as a function of classroom level pretest variance and the proportion of students reassigned, respectively, with simple OLS fits. A positive relationship between pretest variance and $\hat{\gamma}$, and a negative relationship between proportion reassigned and $\hat{\gamma}$ are apparent, but with wide standard errors. To test these hypotheses more formally, we re-fit model (5), adding fixed effects for the variance in pretest scores and proportion reassigned, as main effects and interacted with Z_i. The model reduced the unexplained variance in classroom-level effects from 0.25 to 0.21—these variables explained about 27% of the unexplained variance in treatment effects. The coefficient on the interaction beween pretest variance and reassignment—measuring the extent to which pretest variance explains treatment effects—was estimated as 0.09, with a 95% confidence interval of [-0.75,0.93], so the data are compatible with large associations in either direction between pretest variance and treatment effects. No firm conclusions may be drawn. The coefficient on the interaction beween proportion reassigned and reassignment—measuring the extent to which classroom proportion reassigned explains treatment effects—was estimated as -0.3, with a 95% confidence interval of [-0.57,-0.03], and a p-value of $p = 0.03$. This suggests that the effect of reassignment may be lower—more negative—in classrooms in which a higher proportion of students were reassigned. This aligns with our second hypothesis.

These effect heterogeneity analyses assume (2), no unmeasured confounding. Unfortunately, we are not aware of methods for sensitivity analysis of the type presented in Section 6.1, applied towards estimates of effect heterogeneity. In particular, unobserved confounding may vary by class-

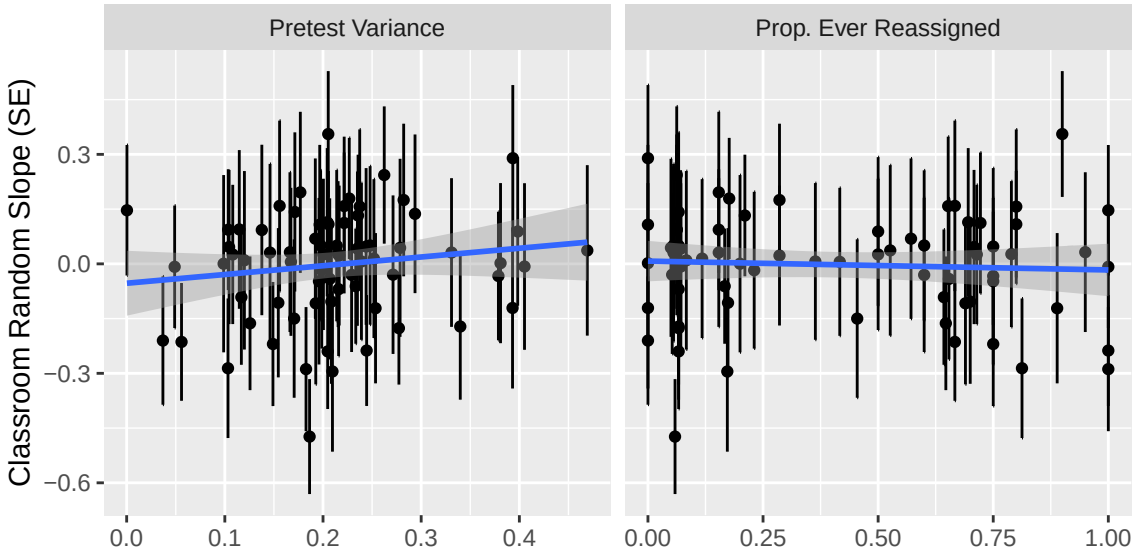

Figure 5: The random effects γ_j from model (5) (with error bars for one standard error) as a function of classroom-level variance in pretest scores and the proportion of students in a classroom who were ever reassigned. OLS fits are added for interpretation.

room; for instance, the structure of the propensity score match may vary with the proportion of students ever reassigned, since within-classroom matches will be scarce when this proportion is high. For those reasons, the conclusions in this section should be taken as suggestive and exploratory.

7. DISCUSSION

A deeper understanding of the use of reassignment and its effects can yield practical and theoretical dividends. Teachers would benefit from clear guidelines as to when and whether reassigning students to a new section may benefit that student's learning. A better understanding of if and when reassignment helps or hurts student learning can contribute to our understanding of the importance of sequence and mastery learning in intelligent tutoring systems.

Here, we estimate that, on average, reassignment hurts student learning, perhaps as much as CTA1 helps. That conclusion comes with two important caveats: first, although it appears unlikely that the entire reassignment effect we estimated is due to confounding from unmeasured variables, a large portion of the effect might be. That is, the magnitude of the reassignment effect we estimated may be an artifact of unmeasured confounding—reassignment may not be as bad as we estimate, or it may be worse. (Of course, we cannot rule out that the entire effect is due to confounding, or that the direction of our estimated effect is wrong.)

Secondly, there is evidence that the effect of reassignment varies widely between classes. Even if it hurts on average, used properly it may help.

More broadly, these issues illustrate the opportunities and perils of analyses of log data from randomized trials of educational technology. Even when the randomization itself does not contribute to an analysis, the combination of log data collected under natural conditions and a long period of time and a posttest measuring student ability at the end of the study can be used to gain insights on tutor use and effects. On the other hand, log data, even from a randomized trial, is observational, and therefore messy and subject to confounding and other threats. Causal modeling of log data from randomized experiments is crucial, but difficult.

8. ACKNOWLEDGMENTS

This work was partially funded by NSF grant #DRL-1420374.

9. REFERENCES

[1] D. Bates, M. Mächler, B. Bolker, and S. Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

[2] G. Blair, J. Cooper, A. Coppock, M. Humphreys, and L. Sonnet. *estimatr: Fast Estimators for Design-Based Inference*, 2019. R package version 0.20.0.

[3] J. Bowers, M. Fredrickson, and B. Hansen. *RItools: Randomization Inference Tools (Development Version)*, 2017. R package version 0.2-0.

[4] M. Caliendo and S. Kopeinig. Some practical guidance for the implementation of propensity score matching. *Journal of economic surveys*, 22(1):31–72, 2008.

[5] W. W. Clearinghouse. Standards handbook, version 4.1, 2020.

[6] T. D. Cook, W. R. Shadish, and V. C. Wong. Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management: The Journal of the Association for Public Policy Analysis and Management*, 27(4):724–750, 2008.

[7] T. D. Cook, P. M. Steiner, and S. Pohl. How bias reduction is affected by covariate choice, unreliability,

and mode of data analysis: Results from two types of within-study comparisons. *Multivariate Behavioral Research*, 44(6):828–847, 2009.

[8] A. T. Corbett, K. R. Koedinger, and W. Hadley. Cognitive tutors: From the research classroom to all classrooms. *Technology enhanced learning: Opportunities for change*, pages 235–263, 2001.

[9] S. Doroudi, K. Holstein, V. Aleven, and E. Brunskill. Sequence matters but how exactly? a method for evaluating activity sequences from data. *Grantee Submission*, 2016.

[10] A. Gelman and J. Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2006.

[11] B. B. Hansen. Full matching in an observational study of coaching for the sat. *Journal of the American Statistical Association*, 99(467):609–618, 2004.

[12] B. B. Hansen and J. Bowers. Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science*, 23(2):219–236, 2008.

[13] B. B. Hansen and S. O. Klopfer. Optimal full matching and related designs via netwrk flows. *Journal of Computational and Graphical Statistics*, 15(3):609–627, 2006.

[14] T. J. Hastie. Generalized additive models. In *Statistical models in S*, pages 249–307. Routledge, 2017.

[15] C. A. Hosman, B. B. Hansen, P. W. Holland, et al. The sensitivity of linear regression coefficients' confidence limits to the omission of a confounder. *The Annals of Applied Statistics*, 4(2):849–870, 2010.

[16] A. Israni, A. C. Sales, and J. F. Pane. Mastery learning in practice: A (mostly) descriptive analysis of log data from the cognitive tutor algebra i effectiveness trial, 2018.

[17] K. R. Koedinger. *Toward evidence for instructional design principles: Examples from Cognitive Tutor Math 6*. ERIC Clearinghouse, 2002.

[18] J. Neyman. Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10:1–51, 1923.

[19] J. F. Pane, B. A. Griffin, D. F. McCaffrey, and R. Karam. Effectiveness of cognitive tutor algebra i at scale. *Educational Evaluation and Policy Analysis*, 36(2):127–144, 2014.

[20] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.

[21] F. E. Ritter, J. Nerb, E. Lehtinen, and T. M. O'Shea. *In order to learn: How the sequence of topics influences learning*, volume 2. Oxford University Press, 2007.

[22] S. Ritter, A. Joshi, S. Fancsali, and T. Nixon. Predicting standardized test scores from cognitive tutor interactions. In *Educational Data Mining 2013*, 2013.

[23] P. Rosenbaum. *Observational studies*. Springer, 2002.

[24] P. Rosenbaum and D. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[25] D. B. Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, pages 34–58, 1978.

[26] A. Sales, T. Patikorn, and N. Heffernan. Bayesian partial pooling to improve inference across a/b tests in edm. In *Proceedings of the 11th International Conference on Educational Data Mining. International Educational Data Mining Society*, 2018.

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

[28] A. C. Sales, J. F. Pane, et al. The role of mastery learning in an intelligent tutoring system: Principal stratification on a latent variable. *The Annals of Applied Statistics*, 13(1):420–443, 2019.

[29] R. Scheines, E. Silver, and I. M. Goldin. Discovering prerequisite relationships among knowledge components. In *EDM*, pages 355–356, 2014.

[30] W. R. Shadish and P. M. Steiner. A primer on propensity score analysis. *Newborn and Infant Nursing Reviews*, 10(1):19–26, 2010.

[31] A. Vuong, T. Nixon, and B. Towle. A method for finding prerequisites within a curriculum. In *EDM*, pages 211–216, 2011.

[32] H. Wickham, M. Averick, J. Bryan, W. Chang, L. D. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. L. Pedersen, E. Miller, S. M. Bache, K. Müller, J. Ooms, D. Robinson, D. P. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, and H. Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019.

[33] D. Wood, J. S. Bruner, and G. Ross. The role of tutoring in problem solving. *Journal of child psychology and psychiatry*, 17(2):89–100, 1976.

[34] Y. Xie. *knitr: A General-Purpose Package for Dynamic Report Generation in R*, 2020. R package version 1.28.