Exploring Causal Mechanisms in a Randomized Effectiveness Trial of the Cognitive Tutor

Adam C Sales
Carnegie Mellon University
Pittsburgh, PA, USA
acsales@cmu.edu

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

ABSTRACT

Cognitive Tutor Algebra I (CTAI), published by Carnegie Learning, Inc., is an Algebra I curriculum, including both textbook components and an automated, computer application that is designed to deliver individualized instruction to students. A recent randomized controlled effectiveness trial, found that CTAI increased students' test scores by about 0.2 standard deviations. However, the study raised a number of questions, in the form of evidence for treatment-effect-heterogeneity. The experiment generated student log-data from the computer application. This study attempts to use that data to shed light on CTAI's causal mechanisms, via principal stratification. Principal strata are categories of both treatment and control students according their potential CTAI usage; they allow researchers to estimate differences in treatment effect between usage subgroups. Importantly, randomization satisfies the principal stratification identification assumptions. We present the results of our first analyses here, following prior observational results. We find that students who encounter more than the median number of sections experience higher effects than their peers who encounter fewer, and students who need more assistance experience *lower* effects than their peers who require less.

Keywords

Causal Mechanisms, Principal Stratification, Intelligent Tutors, Bayesian Hierarchical Models

1. INTRODUCTION

The Cognitive Tutor Algebra I (CTAI) is a technology-based educational intervention that hopes to improve algebra I instruction by individualizing instruction to students needs, providing instant performance feedback, and implementing cognitive theories in mathematics education. [6]

Recently, a randomized controlled effectiveness trial, estimated the effect of a school's adoption of CTAI, under authentic conditions, on its students scores on an algebra proficiency exam. The results were reported in [4]. The study found that CTAI significantly increased test scores for 9th grade students in the second year of implementation, but was unable to detect effects in the experiment's first year, or in the 8th-grade group. These results raise a further question: by what mechanism, and for which students, does CTAI increase achievement? What usage patterns lead to higher effects? Can usage patterns explain the observed treatment effect heterogeneity?

The effectiveness trial produced extensive student usage data, as the computer program logged students' activity. In this paper, we begin use this data—in particular, usage data from the 2nd-year high school sample that apparently experienced a substantial CTAI effect—to explore the relationship between student usage and causal effects.

In doing so, we are guided by a previous study, [7] which (in one model specification) regressed post-test scores on CTAI usage variables, alongside student covariates and pretest scores. That paper was aimed at post-test prediction, not causal inference, but it is of use in generating causal hypotheses: are there different effects for students who use CTAI for different amounts of time? Or for students who require more assistance from the program? Or for students who encounter more sections? This paper is a preliminary inquiry into these questions—more an exposition of the types of results that are possible than a full analysis—future work will delve more deeply into the data.

The data from the CTAI effectiveness study is invaluable for testing these hypotheses: due to its randomization design, we can draw causal conclusions without heroic assumptions. To do so, we will make use of the statistical framework of principal stratification, which we will describe in the following section. The next section will describe our models in detail, and results and conclusions will follow.

2. PRINCIPAL STRATIFICATION

Following [8], we conceptualize causal inference in terms of counterfactuals: comparing what students would have experienced with CTAI with what they would have experienced in its absence. In particular, if Y is the outcome of interest, in our case, post-test scores, we may define two "potential outcomes" for each subject: $Y_i(0)$ is what a subject i would score on the post-test if i's school were assigned to the control condition, and $Y_i(1)$ is what I would score if her school were assigned to treatment.

Principal stratification (PS) [2] is an approach to modeling a categorical or discrete post-treatment variable M within the potential outcomes framework. When treatment assignment Z is binary, each subject i has two potential values of M: $M_i(0)$—the value of M that would be observed under the control condition—and $M_i(1)$, what would be observed under the treatment condition. These define subgroups—principal strata—within which causal effects may be defined. In particular, a principal causal effect is

$$Y(1) - Y(0)|M(1) = m, M(0) = m' \tag{1}$$

that is, the effect of Z on Y among those subjects with particular potential outcomes for M of m and m'.

In this study, following [3], we use principal stratification to examine some hypothesized causal mechanisms of CTAI. For instance, consider the usage variable $totalTime$: the total amount of time students spend working CTAI problems. Since $totalTime$ is continuous, we begin by dichotomizing it; for the sake of simplicity, let $\mu = median(totalTime)$ and $M = \mathbb{1}_{[totalTime>\mu]}$. We can define four principal strata. The first is comprised of those students who, if assigned to CTAI, would use it for more time than μ—$M(1) = 1$—but if assigned to the control condition would use it less, $M(0) = 0$. Next, consider the group $M(1) = 0$; $M(0) = 1$, those students who use CTAI for less time because of their treatment assignments. The remaining two groups are $M(1) = 0$; $M(0) = 0$ and $M(1) = 1$; $M(0) = 1$, those students who would use CTAI less for less, or more, time than m regardless of treatment assignment. By examining differences between the average treatment effects in the four groups, we can learn how CTAI's impact varies for different usage patterns.

Randomization allows us to estimate principal effects as the average treatment minus control difference in gain scores within each estimated stratum. That is, randomization of treatment assignment leads to identification of principal effects: the effect of Z within principal strata. On the other hand, the difference in treatment effects between principal strata does not necessarily estimate a causal quantity. Randomization does not identify students' counterfactual gain scores had they been in alternative principal strata. That being said, differences in treatment effects across strata can suggest causal mechanisms.

Fortunately, the CTAI study's design substantially simplifies the PS analysis, by eliminating two of the principal strata. Students in the control group had (for the most part) no access to the CTAI program. Therefore, we can safely assume that for all students, $M(0) = 0$. This leaves two principal strata, $M(0) = 0$; $M(1) = 0$, and $M(0) = 0$; $M(1) = 1$ —that is, the students who, if assigned to treatment, would use CTAI for more time than m and those who would not. Only one of the potential values of M is directly observed; in particular, $M(1)$ is unknown for subjects in the control group. Stated differently, the values $M(1)$ are missing for students in the control group, but they may be imputed because the "missingness mechanism," treatment assignment, is random, or ignorable. Therefore, randomization of treatment assignment allows us to identify members of each principal stratum, and effects of treatment within those strata.

3. MODELING STRATA AND OUTCOMES

In this preliminary study, we considered three of the usage variables previously modeled as predictors in [7]: $totalTime$, the total amount of time students spent working CTAI problems, $numSec$, the number of sections each student encountered, and $assistance$, the average sum of hints and errors per problem for each student. We ran a separate PS model for each usage variable, but all three PS models had the same form. Each PS model itself was a combination of two multilevel models. The first, fit only within the treatment group, modeled the usage variable M as a function of covariates X_t. This model was used to estimate the usage that control students would have experienced had they been assigned to treatment. The second model used the results of the first model, and a somewhat larger set of covariates X_y, to estimate the effect of random assignment to CTAI in each of the principal strata.

3.1 Usage Model

Modeling each usage variable was a four-step process: first, we calculated the variable's values from the available data; next, we transformed those values so that their observed distributions would be closer to a normal distribution; next, we modeled the transformed variables as a linear function of covariates X_t, and finally, we dichotomized the model's output, to define and estimate principal strata.

As students used CTAI, the program recorded timestamps at the beginning and end of each problem. The difference between these two is the amount of time the student spent on each problem, recorded in milliseconds. The sum of was the variable $totalTime$. The distribution of $totalTime$ was heavily skewed rightward, so we transformed it to ease the modeling process. The transformation that resulted in a distribution whose histogram appeared approximately normal was a box-cox transformation with a parameter of 0.3 [1].

Next, we modeled the transformed $totalTime$ as a function of a set of covariates X_t containing dummy variables for the state in which the school was located, the student's grade, race, sex, special education status, free or reduced-price lunch status and pretest scores, along with missingness indicators. Formally, the model was

$$totalTime_{ijk} = \alpha + X_{ti}^T\beta + \epsilon_{ijk} + \eta_{jk} + \nu_k \tag{2}$$

where α and β are, respectively, an intercept and a vector of coefficients estimated from the data, $\epsilon_{ijk} \sim N(0, \sigma_{st})$ is a student-level random error, $\eta_{jk} \sim N(0, \sigma_{tt})$ is a random effect for teacher, and $\nu \sim N(0, \sigma_{ut})$ is a random effect for school. The variance parameters σ_{st}, σ_{tt} and σ_{ut} are estimated from the data. In other words, $totalTime$ was modeled as multilevel, with students nested within teachers, nested within schools.

The transformed $totalTime$ values, or, in the case of the control sample, their predictions, gave rise to a dichotomous variable M, which took the value of 1 if $totalTime$ or its prediction is greater than its observed median of about 22 hours over the course of the year. The variable M defined two principal strata: those students with $M(1) = 1$ and those with $M(1) = 0$.

CTAI also automatically collected data on the number of hints and errors students request or make. Following [7], we normalized hints and errors by section. Next, we averaged the normalized values by student, producing average assistance per problem, or *assistance*. We transformed *assistance* in the same was as *totalTime*. Next, we modeled *assistance* with equation (2), and dichotomized the results using their observed median, 0.076, which, due to the prior normalization, is not a whole number.

The third usage variable we considered here is *numSec*, the number of sections students encountered on CTAI. We transformed *numSec* with a natural logarithm, modeled it with equation (2), and dichotomized it with its median, 27 sections.

3.2 Outcome Model

For each dichotomized usage variable M, we fit a multilevel linear model to estimate principal effects of CTAI treatment on post-test scores. The post-test from the CTAI effectiveness study is the Algebra Proficiency Exam. It was analyzed with item-response-theory, and its reported scores have a mean of 0 and a standard deviation of 1, so regression coefficients may be interpreted as effect sizes [4]. To account for pre-test scores, while avoiding measurement-error concerns, we modeled students' gain scores, $\texttt{diff}_{ijk}$, the difference between their post-test and pre-test scores. The student-level model, then, was

$$\begin{aligned}\texttt{diff}_{ijkm} =& \alpha' + X_{yi}^T\gamma + \lambda M_{ijkm} + \tau Z_{km} \\ &+ \kappa Z_{km} M_{ijkm} + \epsilon'_{ijkm} + \eta'_{jkm} \\ &+ \nu'_{km} + \zeta_m \qquad (3)\end{aligned}$$

Here α', ϵ', η', and ν' are, respectively, an intercept, and random effects for individual, teacher, and school. The apostrophes indicate that these are distinct from their analogues in equation (2). There is an additional random effect ζ for "match," accounting for the matched-pair randomization design. X_t is a vector of covariates equivalent to those in (2), with the addition of standardized test scores from the prior two years. The principal effects emerge from the coefficients τ and κ: τ is the average effect in the $M(1) = 0$ group, and $\tau + \kappa$ is the average effect in the $M(1) = 1$ group. Finally, λ is the difference in $Y(0)$ between the $M(1) = 1$ and $M(1) = 0$ groups.

Models (2) and (3) were fit simultaneously in JAGS [5], a Bayesian Gibbs sampler. To facilitate Bayesian model fitting, we provided weakly informative priors on all of the model parameters.

4. RESULTS

	Estimate	SE	95% Interval
$M(1) = 0$	0.12	0.06	(0.01,0.25)*
$M(1) = 1$	0.32	0.27	(-0.21,0.85)
Difference	0.20	0.27	(-0.32,0.74)

Table 1: Results for *totalTime*. Point estimates for effect size, standard errors and 95% credible intervals for the average treatment effects in two principal strata, denoted $M(1) = 1$ and $M(1) = 0$, as well as the difference between the two.

We present results for each of the three usage variables we considered. For each variable, we present the average treatment effect for subjects in the $M(1) = 0$ stratum—that is, students whose usage under the treatment condition was, or would be, less than the observed median—the effect for students in the $M(1) = 1$ stratum, and the difference between the two effects. For each effect, we present a point estimate, equivalent to the mean of the posterior distribution, a standard error—the standard deviation of the posterior—and a 95% credible interval, representing the 0.0275 and 0.975 quantiles of the posterior. Effects whose credible interval does not include 0 are marked with an asterisk.

Like [7], we were unable to establish that students who spend more time using CTAI gain more from its use. The relevant results are available in Table 1. We estimated an effect size of 0.12 in the low-time group $M(1) = 0$, and 0.32 in the high-time group $M(1) = 1$. However, the standard errors were too large to draw strong conclusions.

	Estimate	SE	95% Interval
$M(1) = 0$	-0.02	0.07	(-0.16,0.11)
$M(1) = 1$	0.30	0.13	(0.13,0.47)*
Difference	0.32	0.09	(0.14,0.49)*

Table 2: Results for *numSec*

On the other hand, as seen in Table 2, students who encountered a greater number of sections (or would have, had they been assigned to treatment) experienced a much larger effect than those who encountered fewer sections. The effect size for students who encounter more than the median number of sections is, with 0.95 probability, between 0.13 and 0.47—a very large effect. This is about 0.32 higher than for the students who encountered fewer sections, for whom there was no discernible effect at all.

	Estimate	SE	95% Interval
$M(1) = 0$	0.30	0.08	(0.14,0.45)*
$M(1) = 1$	0.12	0.09	(-0.07,0.30)
$M(1) = 2$	-0.08	0.08	(-0.23,0.06)
Difference (0–1)	-0.18	0.09	(-0.34,-0.00)*
Difference (1–2)	-0.20	0.09	(-0.36,-0.01)*

Table 3: Results for *assistance*

Lastly, we suspected that the relationship between assistance and CTAI effect might not be monotonic. That is, it might be that the effect of CTAI is low for students who request many hints and make many errors, but high for those with a medium amount, or vice versa. For that reason, we split the variable at the 1/3 and 2/3 quantiles, and estimated three principal effects. Our suspicion proved false, however, and the result was similar to what was reported in [7]: higher hints and errors corresponded to lower CTAI effects. For students who requested few hints and made few errors, the effect was between 0.14 and 0.45, while 95% intervals for the other two strata included 0. Ninety-five percent intervals on the difference from one strata two the next were entirely negative.

5. DISCUSSION

This work was a first look at causal modeling with usage variables from a randomized experiment of educational software. We showed that without additional identification assumptions, researchers can use log data to form a deeper understanding of their software's effect. That being said, this work is preliminary, both because the statistical models we used may be improved, and because much more information is available in the CTAI log data.

In this paper, we focused on three hypotheses that were suggested in [7]. That paper used a linear regression model, fit using a convenience sample of CTAI users, to show that certain usage variables, among which are the total amount of time students spend solving problems, the number of sections students encounter, and the assistance the software provides them, can predict standardized test scores, even after controlling for a number of baseline covariates. With some very strong assumptions, one may interpret [7]'s results as causal: that seeing more sections, for instance, causes students to achieve higher test scores.

In our design, by contrast, the estimated treatment effects—comparisons between treatment and control students—are inherently causal due to the randomization design. The principal stratification approach allows us to reliably estimate causal effects within the strata. That said, this approach largely replicates the results from [7]. Students who spend more time working on CTAI problems seem to experience a larger effect, but this conclusion is ultimately unclear: the credible interval of the difference in effects between students who use the program for more time and those who use it for less contains 0. On the other hand, we found that students who encounter more sections do indeed experience larger effects. One reason for this result may be that the effect a CTAI user feels is particular to the skills the user practices—students who encounter a wider array of sections learn more from CTAI, and their performance on a wider array of sections of the posttest is improved. At the same time, students who required more assistance per problem—that is, asked for more hints and made more errors—experienced a smaller effect than their peers who required less assistance. This may be for a number of reasons. For instance, perhaps students who need more assistance per problem are struggling more, and have a greater need for a teacher's help. Alternatively, students who ask for a lot of hints and make a lot of mistakes may not be trying their hardest on CTAI, and for that reason may not experience the same rewards from CTAI. More research and data analysis is necessary to properly interpret these results.

Along those lines, we plan a number of future analyses. First, improved models may help us understand the relationships that this paper explores. For instance, dividing the usage variables into three or more categories may be more illuminating than the two categories we explore here. Additionally, it may be useful to match section- or unit-specific usage to appropriate items on the posttest.

Further along, we hope to discover and define interesting multivariate principal strata, perhaps as the result of a cluster analysis of the high-dimensional usage data.

Finally, after cultivating a more complete understanding of the usage patterns that lead to higher CTAI effects, we can explore treatment-effect heterogeneity. In particular, we may be able to answer why in the first year of implementation CTAI did not seem to boost test scores, but in the second year it did. Was differential usage to blame?

In the meantime, this paper uses rigorous causal methods to confirm some previous hypotheses about CTAI's causal mechanisms, and points a way forward for future work modeling usage variables in experimental designs.

6. ACKNOWLEDGMENTS

This work is supported by the United States National Science Foundation Grant #DRL-1420374 to the RAND Corporation and by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B1000012 to Carnegie Mellon University. The opinions expressed are those of the authors and are not intended to represent views of the Institute or the U.S. Department of Education or the National Science Foundation. Thanks to Steve Fancsali, Steve Ritter, and Susan Berman for processing and delivering the CTAI usage data.

References

[1] G. E. Box and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 211–252, 1964.

[2] C. E. Frangakis and D. B. Rubin. Principal stratification in causal inference. *Biometrics*, 58:21–29, 2002.

[3] L. C. Page. Principal stratification as a framework for investigating mediational processes in experimental settings. *Journal of Research on Educational Effectiveness*, 5(3):215–244, 2012.

[4] 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*, 2013.

[5] M. Plummer. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, 2003.

[6] S. Ritter, J. R. Anderson, K. R. Koedinger, and A. Corbett. Cognitive tutor: Applied research in mathematics education. *Psychonomic bulletin & review*, 14(2):249–255, 2007.

[7] S. Ritter, A. Joshi, S. E. Fancsali, and T. Nixon. Predicting standardized test scores from cognitive tutor interactions. In *Proc. of the 6 th International Conf. on Educational Data Mining*, 2013.

[8] D. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology; Journal of Educational Psychology*, 66(5):688, 1974.