Student Usage Predicts Treatment Effect Heterogeneity in the Cognitive Tutor Algebra I Program

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

Asa Wilks
RAND Corporation
Santa Monica, CA, USA
awilks@rand.org

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

ABSTRACT

The Cognitive Tutor Algebra I (CTAI) curriculum, which includes both textbook and online components, has been shown to boost student learning by about 0.2 standard deviations in a randomized effectiveness trial. Students who were assigned to the experimental condition varied substantially in how, and how much, the used the online component of CTAI, but original analyses of the experimental data focused on estimating average effects, and did not examine whether the CTAI treatment effect varied by the amount of style of usage. This study leverages log data from the experiment to present a more nuanced analysis. It uses the framework of Principal Stratification, which estimates the varying CTAI treatment effect as a function of "potential" usage—either how students used the program, or how they would have used it had they been assigned to the treatment condition. With experimental data, Principal Stratification does not require that we assume that all relevant variables have been measured. With this framework, we find that students who receive a medium amount of assistance from the software (in the form of hints and error feedback) experience the largest effects, with lower effects for students who receive a lot or a little; and evidence that students who do not follow the curriculum order experience smaller treatment effects.

Keywords

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

1. INTRODUCTION

Intelligent tutors—computer programs designed to teach—claim to improve student achievement via a number of mechanisms, including a reliance on cognitive modeling, instant feedback, and individualized instruction. As the demand for intelligent tutors grows, so does the demand for evidence of their effectiveness, and the educational research community has kept apace, with a number of randomized field trials [e.g. 5, 9, 14]. Since intelligent tutors are computerized, it is relatively easy for experimenters to collect student log data, alongside traditional evaluation data. This paper will provide a template for how to evaluate the log data from an intelligent tutor experiment, to help elucidate the intelligent tutors' mechanisms and when and for whom they work.

A recent randomized study of Carnegie Learning's Cognitive Tutor Algebra I (CTAI) curriculum, under real-life conditions, was reported in [8]. In the second year of the experiment, in high school classrooms, the study found, that CTAI boosts student learning by about 0.2 standard deviation, on average. However, in the first year of the experiment CTAI's effect was close to nil. Surely one explanation for this heterogeneity is that students and teachers used the curriculum differently in the two years—but how? What aspects of student usage predict a treatment effect?

The effectiveness trial produced extensive student usage data, as the computer program logged students' activity. In this paper, we 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 future work, we will attempt to use these findings to explain the difference between the two years of the experiment.

A preliminary study, [17], argued that the best causal model for the usage data relies on the "principal stratification" framework [2, 7], under which students who used the CTAI software in a particular way are compared to control students who would have used it in the same way, had they been assigned to treatment. This study is the first full study that last year's preliminary study promised. It provides two sets of results exploring different aspects of CTAI's mechanisms: an analysis of assistance, which is calculated from the hints that students request and the errors they make, and an analysis of the the order at which students work on CTAI's sections. The paper also includes a more detailed discussion of the models, and a discussion of some issues with the results in [17].

2. DEFINING THE QUESTION: HOW DOES POTENTIAL USAGE MODERATE THE CTAI EFFECT?

As in [17], in this paper we model student usage under the principal stratification (PS) framework, a generalization of the Neyman-Rubin Causal Model [15] of potential outcomes. If Z is a binary treatment assignment, and Y

is an outcome, each subject has two potential outcomes: $Y(Z = 1)$ and $Y(Z = 0)$, the outcome she would present under the treatment condition, and under the control condition, respectively. Each of these is defined, though unobserved, prior to treatment assignment Z. After subjects have been assigned to treatment, exactly one of the potential outcomes is observable for each subject: for treatment subjects, the observed $Y = Y(Z = 1)$, and for control subjects, $Y = Y(Z = 0)$.

[2] generalized the potential outcomes framework, introducing the concept of principal strata. A principal stratum is a grouping of subjects based on potential values of intermediate outcomes. For example, if we call students' usage values U, each student has usage values $U(Z = 1)$ and $U(Z = 0)$—the usage they would exhibit under the treatment and control conditions, respectively. In the CTAI experiment, $U(Z = 0) = 0$ for all subjects, since no control subjects had access to the cognitive tutor. Say we model usage as a categorical value for K categories, $U = 1, ..., K$. Then there are k principal strata: $\{U(Z = 1) = k, U(Z = 0) = 0\}$ for $k = 1, ..., K$. In this framework, principal stratum membership is observed for students in the treatment group—we observe their usage once they are assigned to treatment, and we know from the experimental design that they would not have used the tutor had they been assigned to control. The potential usage for students in the control group, however, is unobserved, and must be estimated; the following section will discuss this process in more detail.

For each stratum, we can define a "principal effect": the average treatment effect $\tau_k = \mathbf{E}[Y(Z = 1) - Y(Z = 0)|U(Z = 1) = k, U(Z = 0) = 0]$ for subjects in principal stratum k. Although unobserved, these strata are defined prior to treatment assignment—if assigned to treatment, what *would* a student's usage be? That is, observed usage U is an intermediate outcome, or a mediator, but potential usage $U(Z = 0)$ and $U(Z = 1)$ is a pre-treatment covariate, or a moderator. The principal effects are, then, subgroup effects, for various levels of potential usage. Differences between principal effects are differences in the effect of CTAI for students who use (or would use) CTAI differently. To put it more precisely, consider the difference $\tau_j - \tau_k$. This is the difference in the effect of CTAI between the group of subjects who, if given the opportunity, would exhibit usage in the amount of j or the amount of k. While the effect estimates τ_j and τ_k are themselves causal (due to randomization) the difference between them could be due to the effect of usage, or to pre-treatment differences between students in the two groups. In other words, since usage values were not assigned randomly, the difference in CTAI effect between two usage principal strata are not necessarily causal. Still, estimating principal effects, and their differences, along with differences in the composition of principal strata, can shed light on the mechanisms of CTAI.

In one of our analyses below, usage is measured as a continuous, not categorical, variable, so the PS approach entails discretizing usage scores. [4] suggested an alternative: modeling potential usage as a continuous mediator, via an interaction in a regression analysis. They refer to this analysis as a "causal effect predictiveness" or CEP curve. CEP curves are directly analogous to principal strata effects, but with continuous intermediate variables.

3. ESTIMATING PRINCIPAL EFFECTS AND CEP CURVES

Estimating principal effects and CEP curves is a complex process, since first we must estimate unobserved principal strata membership or potential usage variables, and only then to estimate treatment effects. In fact, principal effects, in some circumstances, are only partially identified—even in an infinite sample, a Bayesian credible interval for a principal effect may have a finite width. This is especially the case when researchers attempt to estimate principal effects without covariates, and while relaxing traditional instrumental variables assumptions. However, in the presence of covariates that predict usage variables, we may estimate informative effects.

This section describes the models that we use to estimate principal effects and CEP curves. More details can be found in [16].

3.1 The Model

In general, the central challenge in PS modeling is that principal strata membership is unknown. In the CTAI experiment, since control students had no access to CTAI software, strata membership for the treatment group is known, but must be estimated for the control group. The distribution of the potential outcomes for Y, conditional on covariates, $p(Y(Z = 0)|X_i)$, can be decomposed into the probability distribution of Y given $U_i(Z = 1)$, which is the distribution of interest, times the distribution of $p(U_i(Z = 1)|X_i)$, which, due to random assignment, may be estimated from the treatment group. Then, we may estimate the parameters of $p(Y(Z = 0)|U(Z = 1) = a, X)$ and compare them to the analogous distribution $p(Y(Z = 1)|U(Z = 1) = a, X)$ yielding estimates of treatment effects within principal strata.

If we assume that outcomes are conditionally normally distributed, the result is a finite normal mixture model:

$$p(Y_i(Z = 0)|X_i) = \sum_{k=1}^{K} Pr(U_i(Z = 1) = k|X_i)\phi(\mu_k(Z = 0) + f_k(X_i), \sigma_k) \quad (1)$$

and

$$p(Y_i(Z = 1)|X_i, U(Z = 1) = k) = \phi(\mu_k(Z = 1) + f_k(X_i), \sigma_k) \quad (2)$$

where $\phi(\mu, \sigma)$ is the normal density with mean μ and standard deviation σ. Equations (1)-(2) additionally assume no interaction between covariates and treatment status within principal strata. The contribution of covariates X_i to the mean of $Y_i(Z = 1)$ can vary from stratum to stratum, but within stratum it does not vary with treatment status. In practice, we estimate $f_k(X_i)$ as linear in covariates:

$$f_k(X_i) = X_i^T \beta_k \quad (3)$$

where we estimate a different set of slopes β in each stratum k. The linearity assumption can be relaxed or adjusted based on the model's fit to the data. The effect of CTAI in the k^{th} principal stratum is $\tau_k = \mu_k(Z = 1) - \mu_k(Z = 0)$.

The model to estimate a CEP curve is broadly similar to the PS model, with one important difference. In the PS model, usage was parametrized as a categorical variable, and different effects were calculated for each stratum. In the CEP framework, usage is continuous, and its interaction with the effect of treatment must be modeled. As the next section will discuss, we chose to model the CTAI effect as quadratic in usage, for instance. The CEP outcome model, then, is

$$p(Y_i(Z=0)|X_i) = p_{U(Z=1)|X_i}(a)\phi(f_{U|Z=0}(a) + f_X(X_i), \sigma). \quad (4)$$

and

$$p(Y_i(Z=1)|X_i, U(Z=1)=a) = \phi(f_{U|Z=1}(a) + f_X(X_i), \sigma). \quad (5)$$

where $p_{U(Z=1)|X}(a)$ is the density of $U(Z=1)$ conditional on X, $f_{U|Z=0}(a)$ and $f_{U|Z=1}(a)$ are parametric functions of usage for treated and untreated subjects, respectively, and $f_X(X_i)$ is a model for covariates. The CTAI treatment effect is now a function of potential usage, $U(Z=1)$: $\tau(a) = f_{U|Z=1}(a) - f_{U|Z=0}(a)$.

Models (1), (2), (4), and (5) all require a model for the density of usage, as a function of covariates X. In our paper, the usage model, $p(U(Z=1)|X)$, is also linear in X. When the the usage variable is continuous, it is:

$$p(U(Z=1)|X) = \phi(X\gamma, \sigma_U) \quad (6)$$

normal-theory linear regression. In PS models, when we discretize U, we do so *after* fitting model 6.

When U is binary, we use a linear logistic regression to estimate $p(U(Z=1)|X)$:

$$Pr(U(Z=1)|X) = invLogit(X\gamma) \quad (7)$$

We fit all of the above models simultaneously with Markov Chain Monte Carlo (MCMC), using JAGS and R [10, 11]. Since MCMC is a Bayesian technique, it required priors; we put a normal prior with mean zero and standard deviation 3 on each of the model fixed effects—a prior that easily accommodates any plausible effect, but discourages outlandish estimates. We put a weakly-informative inverse-gamma(0.001, 0.001) prior on the variance parameters.

The models for assistance, described below in Section 5, were fit with the Stampede Supercomputer at the Texas Advanced Computing Center.

3.2 Some Potential Pitfalls

[17] presented a set of preliminary results from principal stratification analyses. They were presented as a first attempt at fitting principal stratification models, to illustrate the technique and its potential for helping us understand some of the factors behind CTAI's effect. However, since the EDM 2015 conference, a number of issues emerged with the preliminary results in that paper. It is instructive to discuss those results as an illustration of potential pitfalls in principal stratification analysis.

3.2.1 Model Convergence

One of the first checks of a Markov Chain Monte Carlo model is convergence. MCMC models (ideally) proceed through two stages: first, in the "burn-in" stage, parameter estimates fluctuate widely as the model converges on the posterior distribution for the parameters. After convergence, the algorithm draws from the posterior distribution of the parameters. From these draws, we can estimate the posterior's mean—a point estimate for the parameters—standard deviation, and quantiles. However, it is not always clear when the burn-in period has ended, and the model has begun sampling from the posterior. There are two principal ways of checking this. Both methods rely on running the MCMC separately in two or more chains. That is, start the Gibbs sampler c separate times, with c sets of starting values for the parameters, and let the c separate chains each take their own course. Then, the results from the c chains may be compared; if the model has converged, they should resemble one another, since they each would have converged on the true posterior distribution. One method of measuring whether this is the case is the Gelman-Rubin R-hat statistic, which compares the within-chain variance two the between-chain variance; since, after the burn-in stage, the chains should all be sampling from the same distribution, the between-chain variance should be small. At convergence, the R-hat statistic should be approximately one. Typically, values of R-hat less than 1.1 are acceptable. Additionally, analysts may inspect "traceplots": plots of the c chains for each parameter. If the chains are each stationary—that is, not changing in location or variance—and seem to share a location and scale with each other, the model has most likely converged. If the various chains converge on different distributions, the model might be non-identified, or multi-modal—several different estimates might be equally consistent with the data.

Some of the models in [17] may not have achieved convergence. In this paper, all of the models had clearly achieved convergence.

3.2.2 Gain-Score Modeling and Covariate Selection

A second concern with the model results from [17] emerged from our use of gain-scores—the difference between a post-test and a pre-test—as the outcome in the model, as opposed to the post-tests themselves. The problem with doing so is that the usage model was linear in the pre-test, by design. In the assistance model, for instance, assistance is anti-correlated with pretests, so the the control subjects who were estimated to have high levels of potential assistance also had high pre-test scores. On the other hand, pre-test scores are anti-correlated with gain scores, due to regression to the mean. So the control subjects with high estimated assistance will have lower gain scores on average. This can lead to an overestimate of an effect in the high-assistance stratum, especially if the usage model is misspecified. In principle this is an easy problem to correct, simply by including pre-test scores as a covariate in the outcome model as well. However, doing so would undermine the rationale of gain score modeling. For these reasons, we relied exclusively on post-test modeling in this paper, with the pre-test as a covariate in both the usage and outcome sub-models.

3.2.3 Student-Level Averages as Usage Variables

[17], and an earlier version of this manuscript, estimated the variation of the CTAI effect as a function of the av-

erage number of hints and errors each student requested or committed (called "assistance").[1] These averages were taken over all of each student's worked problems. Subsequent analysis revealed a curious phenomenon: the students with the most extreme average assistance values worked very few problems–almost uniformly so. Interpreting the CEP curve, in this case, becomes nearly impossible, since average assistance is so closely related to the amount of usage. The reason for the close relationship is straightforward: sample averages are random variables, and the variance of a sample mean is directly proportional to the sample size. The average assistance values for the group of students who worked very few problems had a high variance; conversely, the variance of average assistance for students who worked a large number of problems was much smaller.

The solution we chose for this issue was to run the model not on student-level average assistance values, but on problem-level data directly, adding another level into the multilevel structure. That way, the model considers student-level usage variables to be latent, as opposed to manifest (i.e. directly observed). Extreme values of latent variables estimated from a small number of problems enter into the model less as students with extreme usage patterns, and more as students whose usage is poorly-determined. In other words, from one MCMC draw to another, the estimate for each low-usage student's assistance value would vary considerably, so low-usage students would contribute little to the overall estimate of the CEP curve. We discuss the problem-level assistance model in Section 5.

3.2.4 Model Validation

The difficulty of constructing correct principle stratification models, and the ease of constructing models that yield misleading results, suggests that PS models should undergo rigorous specification checking before they are believed. [1], an excellent example of careful principal stratification analysis, provided guidance on how to validate a PS model, which we followed. We conducted three types of checks with each model:

- Estimating each effect with multiple different models and checking for concordance. In the assistance analysis, we estimated MCMC models treating the usage variable as either categorical or continuous. In both analyses we estimated both a normal-distribution model, as discussed in in Section 3.1, and a "robust" model, in which we substituted student's t-distribution for normal distributions in the model, allowing for outliers.
- Inspecting residual plots to assess model fit, for both the usage model and the outcome model.
- Estimating models with made-up outcome data. We did this primarily with a placebo outcome, generated by adding random noise to the pre-test variable. We then hoped not to find any treatment effects.

[1]The original manuscript also included an analysis of each student's average number of problems per section, which fell prey to the same issues as the assistance analysis. We will revisit the problem-per-section analysis in future work.

In this paper, due to space constraints, we included estimates from alternative methods, but not residual plots or placebo results; these, though, are available upon request.

Unfortunately, we cannot claim, at this point, that a method or model exists that will always recover the correct answer and never mislead—each model needs to be carefully tailored to its data, and then validated.

4. THE DATA

The CTAI experiment is described in [8]. The study was conducted in 73 high schools and 74 middle schools in 52 urban, suburban, and rural school districts in seven states, encompassing nearly 18,700 high school students and 6,800 middle school students. The schools were matched on a set of covariates prior to randomization, and were subsequently randomized to treatment or control conditions within matched pairs.

The study was an effectiveness trial, where the intervention must be adopted in as naturalistic conditions as possible. This means the study is supposed to capture common implementation variation resulting from imperfect implementation or even refusal to implement certain instructional materials. The naturalistic design of the experiment is particularly important for our analysis of student usage—usage patterns in the experiment plausibly correspond with what we may expect in general.

For the current study, we used only data from the second cohort in high schools. This is because that was the stratum in which overall effects were detected at the 5% level. Indeed, in the first year of implementation point estimates for the effect were close to zero. It may be the case that the difference in effect between the first and second years (a difference which itself is statistically significant) is due to different usage patterns. We hope that our larger project of estimating treatment effect heterogeneity by usage will help explicate the heterogeneity by cohort.

Software usage data is available for only a subset of the students in the treatment group. Considering only students who were present at post-test and are thus a part of outcomes analyses, we have usage logs for 83%. Students not present at post-test are considered to have attritted from the study.

The percentage of non-attritted students for whom we have usage data varies by school, from 0% (n=3 schools) to 100% (n=20 schools). We assume that schools that have 0% coverage did not implement the CTAI curriculum, despite being assigned to the treatment group. Carnegie Learning was unable, for technical reasons, to retrieve software usage log data for that school.

4.1 Imputing Missing Data

As described above, there were missing data values in the covariates, as well as in the student log scores. We used the `missForest` package in `R` [18, 11] to impute missing covariate values. The out-of-box normalized root mean-squared-error for the imputation was 0.02. Since this value is so low, since there was a relatively small amount of missing data, and since covariates play a merely predictive role in our analy-

sis, we assumed that the uncertainty from other aspects of the model would dominate the uncertainty due to covariate imputation and only imputed one dataset, rather than a full multiple imputation.

Missing usage data presents a more serious problem. First, some schools in the CTAI study were not included in the usage dataset. We deleted these schools from the analysis, along with their matched pairs. Since a matched randomized experiment is an aggregate of a randomized trial in each matched pair, discarding the matched pairs with missing data is nearly benign.

We classified within-school missing usage data into two groups: some students did not have usage data because they did not use the software. Since absolute software usage is driven primarily by teachers, we calculated the proportion of students with missing data for each teacher. If almost all of a teacher's students were missing from the usage dataset, we assumed that they did not use the tutor in their classroom.

The rest of the missing student usage data was due to our inability to match students to their records. We assumed that these data were missing at random [6]—that their missingness was ignorable conditional on their measured covariates. The missingness was likely not missing completely at random, since students who were difficult to match generally did not fill out their student information thoroughly, and thoroughness may correlate with post-test scores or usage patterns. The imputation strategy for these missing data points was identical to the imputation of unobserved potential usage for the control students. That is, the same model that estimated densities for usage variables for control students also estimated missing usage data for some treated students. The missing data strategy in this case was, therefore, either full-information maximum likelihood or MCMC, depending on the analysis.

5. HIGHS AND ERRORS

5.1 Assistance Scores

	#Errors=0	#Errors>0	Sum
#Hints=0	0.42	0.34	0.76
#Hints>0	0.01	0.23	0.24
Sum	0.43	0.57	1.00

Table 1: The proportions of problems in our dataset in which students make at least one error or request at least one hint.

[12] defined assistance as the sum of the number of hints students request and the number of errors they make, which together represent the feedback CTAI gives the students. High assistance indicates that a student is struggling.

Hints and errors vary from problem to problem, from section to section, and from student to student. Table 1 shows the joint probability of requesting at least one hint and making at least one error in our dataset. In 58% of worked problems, the student requested at least one hint or one error. Further, hints and errors tend to accompany each other: in only 1% of worked problems the student requested a hint without making an error. In many problems, hints and errors occur

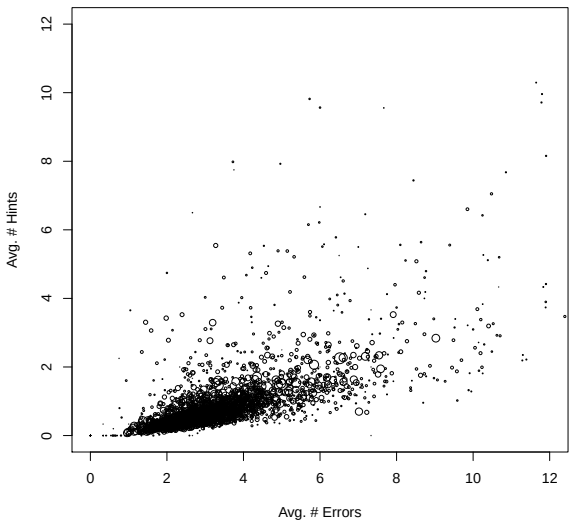

Figure 1: The average number of hints and errors requested for each student. The size of the plotted points is proportional to the square root of the number of problems they completed—and hence to the standard deviation of the plotted averages.

sequentially: a student will work part of the problem, perhaps make an error and receive feedback, perhaps request a hint, and then move on to the rest of the problem. It is important to keep in mind, then, that hints do not always precede errors—sometimes, they are the result of a prior error made while working the same problem.

Figure 1 plots the average number of hints a student requests as a function of the average number of errors he makes. While most students request between 0 and two hints per problem, and make between one and eight errors per problem, some students request far more hints or make far more errors. Further, students who request more hints are much more likely to make more errors. The size of the points in Figure 1 is proportional to the square root of the number of problems they completed—and hence to the standard deviation of the plotted averages. The extreme values in the figure typically come from students who work very few problems, as described in Section 3.2.3, complicating the interpretation of a model that uses average hints or errors as a mediator variable.

For that reason, we incorporated a problem-level sub-model for assistance into our larger principal stratification model. Rather than model the total number of hints and errors per problem, which would necessitate a complex, and possibly misspecified, count-data model, we modeled the probability of a student requesting a hint or making an error (or both) on each problem. The model was as follows:

$$Pr(A_{ip} \geq 1) = invLogit(U_i + \delta_{s[p]}) \quad (8)$$

Where A_{ip} is the total amount of assistance, i.e. hints and errors, that student i experiences from problem p. U_i is a

random student effect, representing the student's propensity to receive assistance on a problem, and $\delta_{s[p]}$ is a section random effect.[2]

The variable U_i, student i's "assistance score," is the mediator that we use to predict her CTAI treatment effect.

U_i is itself predicted, in turn, by a set of covariates including pretest scores, demographics, and teacher random effects nested within school random effects. The results of this usage model are available upon request. They show that prior test scores and "gifted" status are inversely correlated with assistance scores—higher performing students are less likely to make errors or request hints. Special education students are more likely to receive assistance, and males are less likely than females.

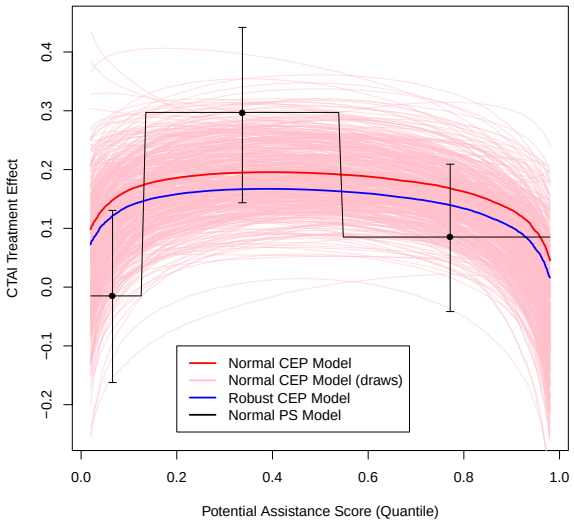

Figure 2: Assistance model results: $\mathbf{E}[Y(Z=1) - Y(Z=0)|U(Z=1)]$, CTAI treatment effect as a function of potential assistance $U(Z=1)$ quantiles. Results are shown for an MCMC CEP normal-distribution model that treats assistance as continuous, a "robust" CEP model based on the t-distribution that allows for outliers, and a normal-distribution PS model that breaks assistance scores into high, medium, and low categories. To display statistical uncertainty, we also plotted 500 draws for the effect function from the CEP model, and 95% credible intervals (error bars) for the three PS effects. The treatment effect is in effect size units.

Figure 2 shows the results for three models—a normal- distribution CEP model, a robust CEP model, and a normal-distribution PS model—which roughly agree that treatment effects are highest for students with assistance scores in the center of the distribution, and lower for students who used a high or low amount of assistance. The PS model, in which assistance scores were discretized, reports more exaggerated differences between treatment effects for students with medium assistance scores and those with high or low scores; these differences, moreover, are highly significant—the probabilities that the average effect for medium students is higher than that for low and high students are 1 and 0.987, respectively. However, when the estimation error is taken into account, it is apparent that the CEP and PS models do not necessarily disagree.

There are a number of ways to interpret these results. The results reflect varying CTAI effects for various usage patterns. One of CTAI's selling points is the instant feedback it provides students as they work through and complete problems. Students who under-utilize this service—in the low assistance stratum—are then likely to experience a smaller CTAI effect. This may be because they began as excellent students—assistance is anti-correlated with pretest scores—and hence did not need the extra help that CTAI provides. Alternatively, students with low assistance scores may be under-utilizing the service for a different reason; perhaps they feared that requesting too many hints, or making too many mistakes, would slow their progress through the tutor, so they were overly cautious.

Students who request hints or make errors quickly, without slow deliberation, may not be able to learn from the problems they work. Some students "game" the system, by requesting hints until they are provided with the correct answer, or they simply do not try very hard to figure out the answer themselves. It may be that the students in the CTAI experiment with very high assistance scores, experience lower treatment effects for some of these reasons Alternatively, they might have struggled with the material in general, and required more personalized help from a teacher, as opposed to a computerized tutor.

However, students in the middle of the assistance distribution experienced large CTAI effects, suggesting an assistance "sweet spot." In future trials, teachers could be instructed to encourage their students to use a medium number of hints, and complete problems with a moderate amount of caution—trying hard to answer problems correctly, but also allowing themselves to make mistakes. If this strategy leads to higher CTAI effects, it suggests that part of the CTAI effect heterogeneity across usage patterns is causal—that using the system differently leads to higher effects.

6. SKIPPING SECTIONS

An important part of the design of CTAI is the scaffolding of skills and knowledge. The skills that students learn in Algebra I build on each other, so the order in which students learn material and master skills matters—at least in theory. The design of CTAI accounts for this order, by insisting that students master certain skills before moving on to others. Indeed, that is the notion that lies behind the sections of the CTAI curriculum.

We attempted to test the hypothesis that this scaffolding matters—that is, do students who the CTAI curriculum learn more from CTAI than students who do not? To answer

[2]The conventional item response theory model in this case would have a problem effect instead of a section effect. We chose section effects rather than problem effects since there are 5438 problems (that only appear once in the dataset, making problem effects difficult to estimate.

this question, we compared the order in which students in the CTAI experiment worked on sections to the intended order. About 80% of students worked on the sections in order. However, 20% of students skipped at least one section. Did the students who skipped one or more sections experience the same CTAI effect as those who completed the sections in the intended order? More precisely, is the CTAI effect the same in the principal stratum of students who, if assigned to CTAI, would complete the section in order, and in the principal stratum of students who, if assigned to CTAI, would skip at least one section?

A complication in estimating counterfactual stratum membership for control students in this case was that in the CTAI setup, teachers, not students, control which sections the students work on. Indeed, there were 38 teachers in the treatment group for whom we had data on whether students skipped a section. Of those 38 teachers, 17 teachers did not have any students who skipped any sections at all, while there were five teachers more than 80% of whose students skipped sections. Since such a large proportion of the variation in section-skipping occurred at the teacher level, we included a set of teacher-level predictors in our usage model. An anonymous reviewer alerted us to the threat of over-fitting; hence, due to the small number of teachers in the treatment group, we chose only two teacher level covariates in the model: percent ESL, and average pre-test. The small covariate-to-sample size ratio at both the student and the teacher levels, combined with the informative priors [See 3], should alleviate concerns of over-fitting.

The usage model, whose results are available upon request, was unsuccessful in estimating precise effects for any covariate, but in aggregate was able to predict stratum membership. One exception is that students with higher pretest scores are more likely to skip sections, as are teachers whose students have higher pretest scores on average.

Stratum	Effect (Normal)	Effect (Robust)
Do Not Skip	0.27	0.19
	0.09	*0.07*
	(0.06–0.44)	(0.05–0.33)
Skip $\geq$ 1 If Treated	-0.09	-0.07
	0.13	*0.11*
	(-0.33,0.17)	(-0.28,0.48)
Difference	-0.36	-0.26
	0.12	*0.11*
	(-0.59,-0.12)	(-0.48,-0.03)

Table 2: The CTAI effect in the two principal strata defined by whether a not a student would skip a section if they were assigned to the treatment. We estimated principal effects with both an MCMC model based on the normal distribution, based on the more robust student's t-distribution. Standard deviations of the posteriors are in italics, and 95% credible intervals (MCMC) are provided in parentheses under the estimates.

The results of our analysis are in Table 2 and Figure 3. Both models detect significantly greater treatment effects in the principal stratum of students who would not skip sections if assigned to the treatment, than in the stratum of students who would. This might be taken as evidence that the order in which students complete sections plays a large role in the effectiveness of CTAI. Alternatively, it may be that teachers who tinker with the order of sections that their students work are likely to tinker with other aspects of the CTAI design as well, to deleterious effect (perhaps along the lines of [13]). In either reading, the effect of CTAI is not merely due to the practice it gives students, or immediate feedback, but also to its underlying pedagogical and cognitive theory.

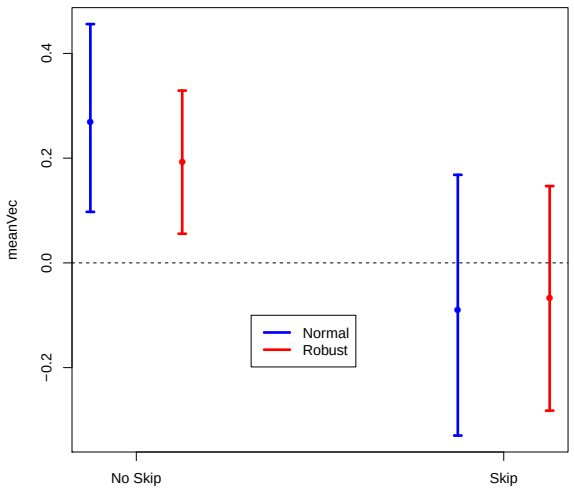

Figure 3: Estimates, and 95% credible intervals, for the CTAI effect in the principle stratum of students who would not skip sections, and in the stratum of students who would. The results plotted for both the normal and t-distribution ("robust") models.

A third possibility is that the entire difference is driven by an underlying teacher or student characteristic, such as ability; students with higher pretest scores are more likely to skip sections—perhaps the treatment effect is significantly lower for them, as well.

7. DISCUSSION

We showed that without additional identification assumptions, researchers can use log data to form a deeper understanding of their software's effect. However, we also discussed some of the difficulties in estimating these models correctly.

We updated and clarified a result from our preliminary study [17]. We find that the relationship between the amount of assistance students receive from CTAI and the CTAI treatment effect they experience is not monotonic. The highest effects appear for the students who receive a medium amount of assistance; those who receive much more or less experience smaller treatment effects, on average. This may be the result of student attributes—that the students at the margins are either too advanced or gaming the software—or it may be that certain modes of software usage are better than

others.

Next, we investigated if students who skip a section in the recommended curriculum, working on sections out of order, may experience lower effects. The result may confirm part of the motivating theory behind CTAI: that Algebra I skills build on each other, so the order at which students work on material can contribute or detract from their success.

Along those lines, we plan a number of future analyses. We hope to update the preliminary study's results that suggested that the CTAI treatment effect increases with the amount of usage, and to investigate the dependence of the CTAI effect on students' mastery of sections. 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.

8. 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. The authors acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing HPC resources that have contributed to the research results reported within this paper. `http://www.tacc.utexas.edu`. Thanks to Steve Fancsali, Steve Ritter, and Susan Berman for processing and delivering the CTAI usage data. Thanks to Brian Junker for helpful advice and guidance.

References

[1] A. Feller, T. Grindal, L. W. Miratrix, and L. Page. Compared to what? variation in the impact of early childhood education by alternative care-type settings. *Annals of Applied Statistics*, 2016. in press.

[2] C. E. Frangakis and D. B. Rubin. Principal stratification in causal inference. *Biometrics*, 58(1):21–29, 2002.

[3] A. Gelman, A. Jakulin, M. G. Pittau, and Y.-S. Su. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, pages 1360–1383, 2008.

[4] P. B. Gilbert and M. G. Hudgens. Evaluating candidate principal surrogate endpoints. *Biometrics*, 64(4):1146–1154, 2008.

[5] J. B. Heppen, K. Walters, M. Clements, A.-M. Faria, C. Tobey, N. Sorensen, and K. Culp. Access to algebra i: The effects of online mathematics for grade 8 students. ncee 2012-4021. *National Center for Education Evaluation and Regional Assistance*, 2011.

[6] R. J. Little and D. B. Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2014.

[7] 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.

[8] 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.

[9] J. F. Pane, D. F. McCaffrey, M. E. Slaughter, J. L. Steele, and G. S. Ikemoto. An experiment to evaluate the efficacy of cognitive tutor geometry. *Journal of Research on Educational Effectiveness*, 3(3):254–281, 2010.

[10] M. Plummer. *rjags: Bayesian Graphical Models using MCMC*, 2016. R package version 4-5.

[11] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

[12] S. Ritter, A. Joshi, S. Fancsali, and T. Nixon. Predicting standardized test scores from cognitive tutor interactions. In *EDM*, pages 169–176, 2013.

[13] S. Ritter, M. Yudelson, S. E. Fancsali, and S. R. Berman. How mastery learning works at scale. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale*, pages 71–79. ACM, 2016.

[14] J. Rochelle, M. Feng, N. Heffernan, and C. Mason. Preliminary findings from an efficacy study of online mathematics homework, 2015. Poster presented at an US Dept of Education, Institute for Educational Sciences meeting of investigators of funded projects.

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

[16] A. C. Sales and J. Pane. Modeling the treatment effect from educational technology as a function of student usage, 2016. Conference Paper for AEFP Annual Conference 2016.

[17] A. C. Sales and J. F. Pane. Exploring causal mechanisms in a randomized effectiveness trial of the cognitive tutor. In *EDM*, 2015.

[18] D. J. Stekhoven. Missforest: nonparametric missing value imputation using random forest. *Astrophysics Source Code Library*, 1:05011, 2015.