The Effect of an Intelligent Tutor on Performance on Specific Posttest Problems

ABSTRACT

This paper drills deeper into the documented effects of the Cognitive Tutor Algebra I and ASSISTments intelligent tutoring systems by estimating their effects on specific problems. We start by describing a multilevel Rasch-type model that facilitates testing for differences in the effects between problems and precise problem-specific effect estimation without the need for multiple comparisons corrections. We find that the effects of both intelligent tutors vary between problems– the effects are positive for some, negative for others, and undeterminable for the rest. Next we explore hypotheses explaining why effects might be larger for some problems than for others. In the case of ASSISTments, there is no evidence that problems that are more closely related to students' work in the tutor displayed larger treatment effects.

Keywords

Causal impact estimates,multilevel modeling,intelligent tutoring systems

1. INTRODUCTION: AVERAGE AND ITEM-SPECIFIC EFFECTS

The past decade has seen increasing evidence of the effectiveness of intelligent tutoring systems (ITS) in supporting student learning [7][13]. However, surprisingly little detail is known about these effects such as which students experience the biggest benefits, under what conditions. This paper will focus on the question of which areas of learning had the largest impact in two different year-long randomized trials: of the Cognitive Tutor Algebra I curriculum (CTA1) [17] and of the ASSISTments ITS [22].

Large-scale efficacy or effectiveness trials in education research, including evaluations of ITS [17][18][22], often estimate the effect of an educational intervention on student scores on a standardized test. These tests consist of many items, each of which tests student abilities in, potentially, a separate set of skills. Prior to estimating program effects, analysts collapse data across items into student scores, often using item response theory models [25] that measure both item- and student-level parameters. Then, these student scores are compared between students assigned to the intervention group and those assigned to control.

This approach has its advantages, in terms of simplicity and (at least after aggregating item data into test scores) model-free causal identification. If each item is a measurement of one underlying latent construct (such as "algebra ability") aggregating items into test scores yields efficiency gains. However, in the (quite plausible) case that posttest items actually measure different skills, and the impact of the ITS varies from skill to skill, item-specific impacts can be quite informative.

In the case of CTA1 and ASSISTments, we find that, indeed, the ITS affect student performance differently on different posttest items, though at this stage it is unclear why the affects differed.

The following section gives an overview of the two large-scale ITS evaluations we will discuss, including a discussion of the available data and of the two posttests. Next, Section 3 will discuss the Bayesian multilevel model we use to estimate item-specific effects, including a discussion of multiple comparisons; Section 4 will discuss the results—estimates of how the two ITS impacted different posttest items differently; Section 5 will present a preliminary exploration of some hypotheses as to why ASSISTments may have impacted different skills differently; and Section 6 will conclude.

2. THE CTA1 AND ASSISTMENTS TRIALS

This paper uses data from two large-scale field trials of ITSs CTA1 and ASSISTments. The CTA1 intervention consisted of a complete curriculum, combining the Cognitive Tutor ITS, along with a student-centered classroom curriculum. CTA1 was a created and run by Carnegie Learning; an updated version of the ITS is now known as Mathia. The Cognitive Tutor is described in more detail in [2] and elsewhere, and the effectiveness trial is described in [17]. ASSISTments is a free online-homework platform, hosted by Worcester Polytechnic Institute, that combines electronic versions of textbook problems, including on-demand hints and immediate feedback, with bespoke mastery-based problem sets known as "skill builders." ASSISTments is described in [10] and the efficacy trial is described in [22].

This section describes the essential aspects of the field trials and the data that we will use in the rest of the paper.

2.1 The CTA1 Effectiveness Trial

From 2007 to 2010, the RAND Corporation conducted a randomized controlled trial to compare the effectiveness of the CTA1 curriculum to business as usual (BaU). The study tested CTA1 under authentic, natural conditions, i.e., oversight and support of CTA1's use was the same as it would have been if there was not a study being conducted. Nearly 20,000 students in 70 high schools ($n = 13,316$ students) and 76 middle schools ($n = 5,938$) located in 52 diverse school districts in seven states participated in the study. Participating students in Algebra I classrooms took an algebra I pretest and a posttest, both from the CTB/McGraw-Hill Acuity series.

Schools were blocked into pairs prior to randomization, based on a set baseline, school-level covariates, and within each pair, one school was assigned to the CTA1 arm and the other to BaU. In the treatment schools, students taking algebra I were supposed to use the CTA1 curriculum, including the Cognitive Tutor software; of course, the extent of compliance varied widely [12][11].

Results from the first and second year of the study were reported separately for middle and high schools. In the first year, the estimated treatment effect was close to zero in middle schools and slightly negative in high schools. However, the 95% confidence intervals for both these results included negative, null, and positive effects. In the second year, the estimated treatment effect was positive–roughly one fifth of a standard deviation—for both middle and high schools, but it was only statistically significant in the high school stratum.

In this study, we make use of students' overall scores on the pretest, anonymized student, teacher, school, and randomization block IDs, and an indicator variable for whether each student's school was assigned to the CTA1 or BaU, along with item-level posttest data: whether each student answered each posttest item correctly. For the purposes of this study, skipped items were considered incorrect.

2.1.1 Posttest: The Algebra Proficiency Exam

The RAND CTA1 study measured the algebra I learning over the course of the year using the McGraw-Hill Algebra Proficiency Exam (APE). This was a multiple choice standardized test with 32 items testing a mix of algebra and pre-algebra skills. Table 1, categorizes the test's items by the algebra skills they require, and gives an example of a problem that would fall into each category. The categorization was taken from the exam's technical report [6].

2.2 The Maine ASSISTments Trial

From 2012–2014, SRI International conducted an randomized field trial in the state of Maine to estimate the efficacy of ASSISTments in improving 7th grade mathematics achievement. Forty-five middle schools from across the state of Maine were randomly assigned between two conditions: 23 middle schools were assigned to a treatment condition; mathematics teachers in these schools were instructed to use ASSISTments to assign homework, receiving support and professional development while doing so. The remaining 22 schools in the BaU condition were barred from using ASSISTments during the course of the study but were offered the same resources and professional development as the treatment group after the study was over. The study was conducted in Maine due to the state's program of providing every student with a laptop, which allowed students to complete homework online.

The 45 participating schools were grouped into 21 pairs and one triplet based on school size and prior state standardized exam scores; one school in each pair, and two schools in the triplet, were assigned to the ASSISTments condition, with the remaining schools assigned to BaU. Subsequent to random assignment, one of the treatment schools dropped out of the study, but its matched pair did not. Although the study team continued to gather data from the now-unmatched control school, that data was not included in the study. However, we are currently unable to identify which of the control schools was excluded from the final data analysis, so the analysis here includes 44 schools, while [22] includes only 43.

The study measured student achievement on the standardized TerraNova math test at the end of the second year of implementation, and estimated a treatment effect of 0.18 ± 0.12 standard deviations.

In this study, we make use of anonymized student, teacher, school, and randomization block IDs, and an indicator variable for whether each student's school was assigned to the ASSISTments or BaU, along with item-level posttest data: whether each student answered each posttest item correctly. For the purposes of this study, skipped items were considered incorrect. The initial evaluation included a number of student-level baseline covariates drawn from Maine's state longitudinal data system, include prior state standardized test scores. We do not currently have access to that data; the only covariate available was an indicator of whether each student was classified as special education.

2.3 The TerraNova Test

The primary outcome of the ASSISTments Maine trial was students' scores on the TerraNova Common Core assessment mathematics test, published by Data Recognition Corporation CTB. The TerraNova assessment includes 37 items, 32 of which were multiple choice and 5 of which were open response. Actually, item number 37 has three parts, labeled 37a, 37b, and 37c, which are scored separately, so it is more accurate to describe the test as having 39 items. The items are supposed to align with the Common Core State Standards, but the research team was not given a document aligning CCSS with the test items. Instead, a member of the ASSISTments staff with expertise in middle school education aligned them according to her best judgment. Table 2 gives this alignment. More information on specific standards can be found at the CCSS website [16].

3. METHODOLOGY: MULTILEVEL EFFECTS MODELING

Objective	Items	Example
Functions and Graphs	6, 8, 19, 20, 22, 23, 27, 31, 32	Which of these points is on the graph of [function]
Geometry	12, 18, 24, 29	Find the length of the base of the right triangle shown below
Graphing Linear Equations	5, 9, 15, 17, 26	Which of the lines below is the graph of [linear equation]?
Quadratic Equations and Functions	2, 25, 28, 30	Which of these shows a correct factorization of [quadratic equation]?
Solving Linear Equations and Linear Inequalities	1, 4, 11, 13, 16	Solve the following system of equations
Variables, Expressions, Formulas	3, 7, 10, 14, 21	Which of these expressions is equivalent to the one below?

Table 1: Objectives required for the 32 items of the Algebra Proficiency Exam, the posttest for the CTA1 Evaluation

CCSS	Items
Expressions and Equations	17,28,36
Functions (8G)	26,27
Geometry	12,16,19,21,23,31,35
Make sense of problems and persevere in solving them (MP)	13
Ratios and Proportional Relationships	22,24,25,29,33
Reason abstractly and quantitatively (MP)	15,20
Statistics and Probability	10,11,32,34,37a,37b,37c
The Number System	1,2,3,4,5,6,7,8,9,14,18,30

Table 2: Common Core State Standards (CCSS) for the 37 TerraNova items, as identified by the ASSISTments team. Standards are from grade 7 except where indicated–grade 8 (8G) or Mathematical Practice (MP)

In principal estimating program effects on each posttest item is straightforward: the same model used to estimate effects on student overall scores could be used to estimate effects on each item individually (perhaps—but not necessarily—adapted for a binary response). However, estimating 32 separate models for each stratum of the CTA1 study, and 39 separate models for the ASSISTments study ignores multilevel structure of the dataset, and leads to imprecise estimates. Moreover, doing so invites problems of multiple comparisons—between the four strata of the CTA1 study and the ASSISTments study, there are 167 separate effects to estimate. If each estimate is subjected to a null hypothesis test at level $\alpha = 0.05$, even if neither ITS affected test performance at all, we would still expect to find roughly eight significant effects.

Instead, we estimated item-specific effects with a multilevel logistic regression model model [8], based roughly on the classic "Rasch" model of item response theory [25][20]. That is, we estimated all item-specific effects for a particular experiment simultaneously, with one model, in which the item-specific effect estimates are random effects. The separate effects were modeled as if drawn from a normal distribution with a mean and standard deviation estimated from the data. This normal distribution can be thought of as a Bayesian prior distribution; the fact that its parameters are estimated from the data puts us in the realm of empirical Bayes [5]. This prior distribution acts as a regularizer, shrinking the several item-specific effect estimates towards their mean [15]. Although doing so incurs a small amount of bias, it reduces standard errors considerably while maintaining the nominal coverage of confidence intervals [23].

Gelman, Hill, and Masanao [9] argue that estimating a set of different treatment effects within a multilevel model also obviates the need for multiplicity corrections. Generally speaking, the reason for spurious significant results is that as a group of estimates gets larger, so does the probability that one of them will exceed the test's critical value. In other words, as a the set of estimates grows, so does their maximum (and their minimum, in magnitude). Multilevel modeling helps by shrinking the most extreme estimates towards their common mean. Since extreme values are less likely in a multilevel model, so are spuriously significant effect estimates.

A small simulation study in the Appendix (mostly) supports Gelman et al.'s argument. As the number of estimated effects grows, the familywise error rate (i.e. the probability of *any* type-I error in a group of tests) grows rapidly if effects are estimated and tested separately, but not if they are estimated simultaneously in a multilevel model. However, the error rates for the multilevel model effect estimates are slightly elevated—hovering between 0.05 and 0.075 throughout. There is good reason to believe that a fully Bayesian approach will improve these further (see, e.g., [21], p. 425).

3.1 The Model for the CTA1 Posttest

For the CTA1 RCT, we estimated a separate model for high school and middle school, but we combined outcome data across the two years. Let $Y_{ij} = 1$ if student i answered item j correctly, and let $\pi_{ij} = Pr(Y_{ij} = 1)$. Then the multilevel

logistic model was:

$$\begin{aligned} logit(\pi_{ij}) &= \beta_0 + \beta_1 Year2_i + \beta_2 Trt_i + \beta_3 Pretest_i \\ &+ \beta_4 Year2_i Trt_i + \beta_5 Year2_i Pretest_i \\ &+ \gamma_{j0} + \gamma_{j1} Trt_i + \gamma_{j2} Year2_i + \gamma_{j3} Year2_i Trt_i \\ &+ \delta_i + \eta_{cls[i]} + \epsilon_{sch[i]} \end{aligned} \quad (1)$$

Where $Year2_i = 1$ if student i was in the 2nd year of the study and 0 otherwise, $Trt_i = 1$ if student i was in a school assigned to treatment, and $Pretest_i$ is i's pretest score. The coefficients β_0–β_5 are "fixed effects," that is, they are not given any probability model. γ_{j0}–γ_{j3} vary with posttest item j, and are modeled jointly as multivariate normal: $\boldsymbol{\gamma} \sim MVN(\mathbf{0}, \Sigma)$, where Σ is a 4×4 covariance matrix for the γ terms. Similarly, the random intercepts δ_i, $\eta_{cls[i]}$, and $\epsilon_{sch[i]}$, which vary at the student, classroom, and school level, are each modeled as univariate normal with mean 0 and a standard deviation estimated from the data.

Collecting like terms in model (1), note that for a student in the first year of the study, the effect of assignment to the CTA1 condition is $\beta_2 + \gamma_{j1}$ on the logit scale; in other words, the effects of assignment to CTA1 in year 1 are modeled as normal with a mean of β_2 and a variance of Σ_{22}. The variance Σ_{22} estimates the extent to which the effect of assignment to the CTA1 condition varies from one problem to another. If the effect were the same for every posttest problem, we would have $\Sigma_{22} = 0$. For students year 2, the effect on problem j is $\beta_2 + \beta_4 + \gamma_{j1} + \gamma_{j3}$ on the logit scale—the effects are normally distributed with a mean of $\beta_2 + \beta_4$ and a variance of $\Sigma_{22} + \Sigma_{44} + 2\Sigma_{24}$. The Σ matrix also includes the covariance between the effects of the intervention on items in year 1 and the effects on the same items in year 2 as

$$Cov(\gamma_{j1}, \gamma_{j1} + \gamma_{j3}) = Var(\gamma_{j1}) + Cov(\gamma_{j1}, \gamma_{j3}) = \Sigma_{22} + \Sigma_{23}$$

Likelihood ratio tests using the χ^2 distribution can test the null hypothesis that the variance of treatment effects are 0. For simplicity, we did so using separate models for the two years, rather than the combined model (1).

The treatment effects themselves are estimated using the BLUPs (best linear unbiased predictors) for the random effects $\boldsymbol{\gamma}$. In many contexts, random effects are considered nuisance parameters, and primary interest is in the fixed (unmodeled) effects $\boldsymbol{\beta}$. However, there is a long tradition, mostly in the Bayesian and empirical Bayes literature, of using BLUPs for estimation of quantities of interest. The models were fit in `R` [19] using the `lme4` package [3], which provides empirical Bayesian estimates of the conditional (or posterior) variance of the BLUPs, which we use (in combination with the estimated standard errors for fixed effects) in constructing confidence intervals for item-specific effects.

3.2 The Model for the ASSISTments Posttest

The model for estimating item-specific effect of ASSISTments on TerraNova items was highly similar to model (1). There were four important differences: first, there was only one year of data. Second, we did not have access to pretest scores, but we did include an indicator for special education status as a covariate. Third, we found large differences in effects between multiple choice and open response items, and hence decided to include an item-type fixed effect. Lastly, the hierarchical variance structure for student errors was somewhat different—we included an error term for teacher instead of classroom, and included random intercepts for randomization block.[1]

All in all, the model was:

$$\begin{aligned} logit(\pi_{ij}) &= \beta_0 + \beta_1 Trt_i + \beta_2 SpEd_i + \beta_3 Type_j \\ &+ \beta_4 Type_j Trt_i \\ &+ \gamma_{j0} + \gamma_{j1} Trt_i \\ &+ \delta_i + \eta_{tch[i]} + \epsilon_{sch[i]} + \zeta_{pair[i]} \end{aligned} \quad (2)$$

where $Type_j = 1$ if item j is open-ended and 0 if multiple choice, $SpEd_i = 1$ if student i is classified as needing special education, $\eta_{tch[i]}$ is a random intercept for i's teacher, and $\zeta_{pair[i]}$ is a random intercept for i's school's randomization block. The rest of the parameters and variables are defined the same as in (1). The treatment effect on problem j is modeled as $\beta_1 + \gamma_{j1}$ for multiple choice items and $\beta_1 + \beta_4 + \gamma_{j1}$ for open ended response items. The random effects $\boldsymbol{\gamma} \sim N(\mathbf{0}, \Sigma)$ where Σ is a 2×2 covariance matrix.

4. MAIN RESULTS: ON WHICH ITEMS DID ITSS BOOST PERFORMANCE?

4.1 CTA1

Figure 1 gives the results from model (1) fit to the middle school and to the high school sample. Each point on the plot represents the estimated effect of assignment to the CTA1 condition on the log odds of a correct answer on one posttest item. The estimates are accompanied by approximate 95% confidence intervals.

It is immediately clear that the effect of assignment to CT vary between posttest items–indeed the χ^2 likelihood ratio test rejects the null hypothesis of no treatment effect variance with $p < 0.001$ in all four strata.

In the middle school sample, the average treatment effect across items was close to 0 for both years (-0.08 in year 1 and 0.03 in year 2 on the logit scale), and not statistically significant. However, the standard deviation of treatment effects between problems was much higher—0.31 in year 1 and 0.29 in year 2, implying that assignment to CTA1 boosted performance on some problems and hurt performance on others. To interpret the standard deviation of effects on the probability scale, consider that for a marginal student, with a 1/2 probability of answering an item correctly, a difference of 0.3 between two treatment effects would correspond to a difference in the probability of a correct answer of about 7.5% (using the "divide by 4 rule" of [8] p. 82). The effects are also moderately correlated across the two years, with $\rho \approx 0.4$—items that CTA1 impacted in year 1 were somewhat likely to be similarly impacted in year 2.

Many of the treatment effects in the upper pane of Figure 1 are estimated with too much noise to draw strong conclusions—the sample size was substantially smaller in

[1] In linear models it is typically recommended to include fixed effects for randomization block [4]. In logistic regression, including a large number of fixed effects violates the assumptions underlying the asymptotic [1]. We tried it both ways and found that it made little difference.

Figure 1: Estimated treatment effects of CTA1 for each level—high school or middle school—implementation year, and posttest item, with approximate 95% confidence intervals

the middle school stratum than in the high school stratum. However, some effects are discernible: in year 1, effects were negative, and on the order of roughly 0.4 on the logit scale (0.1 on the probability scale for a marginal student) on items 1, 2, 9, 10, 12, 19, 22, and 25, and on the order of approximately 0.7 for item 17 (which asks students to match a linear equation to its graph), and similarly-sized positive effects on items 27, 30, and 32. In year 2 there were fewer clearly negative effects—on items 1 and 7—and more positive effects, such as on items 16, 18, 22, 29, and 32. There is a striking difference between the year 1 and year 2 effects on item 22, which asks students to match a quadratic expression to its graph—the effect was quite negative in year 2 and quite positive in year 2.

In the high school sample, the average treatment effect across items was roughly -0.1 in year 1 and 0.13 in year 2, on the logit scale, neither statistically significant–though the difference between the average effect in the two years was significant ($p < 0.001$). The effects varied across items, though less widely in high school than in middle school—in both years the standard deviation of item-specific effects was roughly 0.17. Item-specific effects were more highly correlated across years ($\rho \approx 0.69$)—at some points in the lower pane of Figure 1 it appears as though the curve from year 2 was simply shifted up from year 1.

The item-specific effects in the high school sample were estimated with substantially more precision than in the middle school sample, due to a larger sample size. In year 1, there were striking negative effects on items 2, 14, and 25 which ask students to manipulate algebraic expressions, and on item 12, which ask students to calculate the length of the side of a triangle. In year 2, these negative effects disappeared. Instead, there were positive effects, especially on items 8 and 22, which both ask about graphs of algebraic functions, and on a stretch of items from 15–22. The difference in the estimated effects between years was positive for all items and highest for problems 2, 20, and 25, which ask students to manipulate or interpret algebraic expressions, and 12, the triangle problem. In items 2, 12, and 25, the effect was significantly negative in year 1 and closer to zero in year 2, while for item 20 the effect was close to zero in year 1 and positive in year 2.

Figure 2 plots the estimated effect on each posttest item as a function of the item's objective in Table 1. Some patterns are notable. There was a wide variance in the effects on the four geometry problems for middle schoolers in year 1, but in year 2 all the effects on geometry items were positive and roughly the same size. The geometry items in the high school sample follow a similar, if less extreme, pattern. Across both middle and high school, the largest positive effects were for Functions and Graphs problems, especially item 22 for year 2; on items 23, 27, 31 and 32, middle schoolers—especially in year 2—saw positive effects while high schoolers saw effects near 0.

4.2 ASSISTments

Figure 3 gives the results from model (2), plotting item-specific effect estimates with approximate 95% confidence

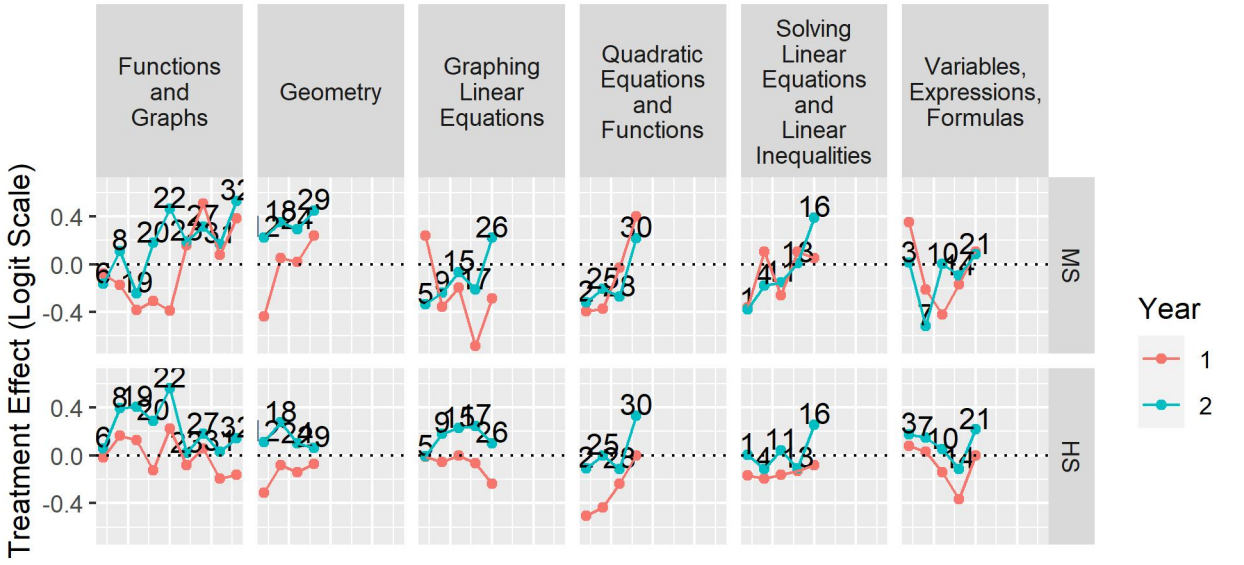

Figure 2: Estimated treatment effects of CTA1 posttest items arranged by the group of skills each item is designed to test. See Table 1 for more detail.

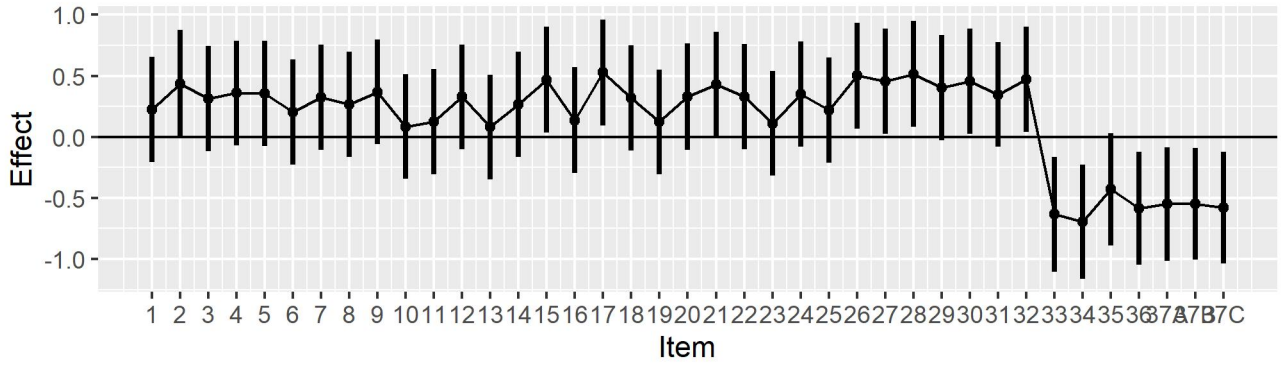

Figure 3: Estimated treatment effects of ASSISTments for each posttest item, with approximate 95% confidence intervals

intervals for each TerraNova posttest item. It is immediately apparent that there were a negative effects on open-ended questions, 33–37C, and (somewhat more ambiguous) positive effects on multiple choice items. The model estimated an average effect of 0.32, with a standard error of 0.20, for multiple choice problems, and of -0.52, with a standard error of 0.21, for open-ended questions. After accounting for that difference (i.e. within item type categories) the standard deviation of item-specific effects was positive ($p<0.001$) but less than for the CTA1 items: it was estimated as 0.15 on the logit scale. The confidence intervals in Figure 3 are also much wider than those for CTA1; we suspect that a large part of the reason is that we did not have access to pretest scores, an important covariate.

The largest effects on the multiple choice items were 28 and 17, which both required students to plug in values for variables in algebraic expressions. The confidence intervals around the effects for items 15, 26, 27, 30, and 32 also exclude 0. On the other end, effects on all open-ended items other than 35 were statistically significant, and fairly similar to each other.

Figure 4 plots item-specific effects for multiple choice TerraNova items grouped according to their CCSS, as in Table 2, with the non-grade-7 standards grouped together as "Other." Interestingly, the largest effects tended to be for items in this "Other" category—as did the smallest effect, for item 13. Effects for problems in the "Number System" and "Ratios and Proportional Relationships" categories had the most consistent effects, between 0.2 and 0.4 on the logit scale.

5. EXPLORING HYPOTHESES ABOUT *WHY* ASSISTMENTS EFFECTS DIFFERED

Researchers on the ASSISTments team have built on the CCSS links of Table 2, linking TerraNova posttest items to data on student work within ASSISTments, for students in the treatment condition. This gives us an opportunity to use student work within ASSISTments to explain some of the variance in treatment effects.

Like TerraNova items in Table 2, ASSISTments problems are linked with CCSS. By observing which problems treatment students worked on, and using this linkage, we could observe which Common Core standards they worked on the most within ASSISTments. We hypothesized that treatment effects might be largest for the TerraNova problems that were linked with the Common Core standards students spent the most time working on. In other words, we linked TerraNova items with worked ASSISTments problems *via* Common Core standards. The Common Core linkage we used in this segment was finer-grained than Table 2, so TerraNova items in the same category in Table 2 may not be linked with the same problems in this analysis.

We examined our hypothesis in two ways: examining the relationships between treatment effects and the number of related ASSISTments problems students in the treatment group worked, and the number of related ASSISTments problems students in the treatment group worked *correctly*. This analysis includes two important caveats: first, the linkages, both between TerraNova items and CCSS, and between ASSISTments problems and CCSS, were subjective and error-prone, possibly undermining the linkage between TerraNova items and ASSISTments problems. Secondly, student work in ASSISTments is necessarily a post-treatment variable—it was affected by treatment assignment. If the treatment randomization had fallen out differently, different schools would have been assigned to the ASSISTments condition and different ASSISTments problems would have been worked. Including the number of worked or correct related problems as a predictor in a causal model risks undermining causal interpretations [14].

Figures 5 and 6 plot estimated item-specific effects for multiple choice TerraNova items against the number of ASSISTments problems that students in the treatment arm worked or worked correctly, respectively, over the course of the RCT. The X-axis is on the square-root scale, and a loess curve is added for interpretation. Little, if any, relationship is apparent in either figure, suggesting either the lack of a relationship between specific ASSISTments work and posttest items, or issues with the linkage. This is hardly surprising, given both the difficulty in linking ASSISTments and TerraNova problems, and given the fact that topics in mathematics are inherently connected, so that improving one skill tends to improve others as well.

6. CONCLUSIONS

Education researchers are increasingly interested in "what works." However, the effectiveness of an intervention is necessarily multifaceted and complex—effects differ between students, as a function of implementation [24], and, potentially, as a function of time and location. In this paper we explored a different sort of treatment effect heterogeneity—differences in effectiveness for different outcomes—specifically, different posttest items measuring different skills. Collapsing item-level posttest data into a single test score has the advantage of simplicity (which is nothing to scoff at, especially in complex causal scenarios) but at a cost. Analysis using only summary test scores squanders a potentially rich source of variability and information about intervention effectiveness that is already at our fingertips. There is little reason *not* to examine item-specific effects.

In this paper, we showed how to estimate item specific effects using a Bayesian or empirical Bayesian multilevel modeling approach that, we argued, can improve estimation precision and avoid the need for multiplicity corrections. The estimates we provided here combine maximum likelihood estimation and empirical Bayesian inference; there is good reason to suppose that a fully Bayesian approach would provide greater validity, especially in standard error estimation and inference. However, fitting complex multilevel models using Markov Chain Monte Carlo methods is computationally expensive, and can be very slow, even with the latest software. We hope to explore this option more fully in future work.

While estimating item-specific effects is relatively straightforward, interpreting them presents a significant challenge. This is due to a number of factors: first, when looking for trends in treatment effects by problem attributes, the sample size is the number of exam items, not the number of students, so patterns can be hard to observe and verify. Secondly, there is a good deal of ambiguity and subjectiv-

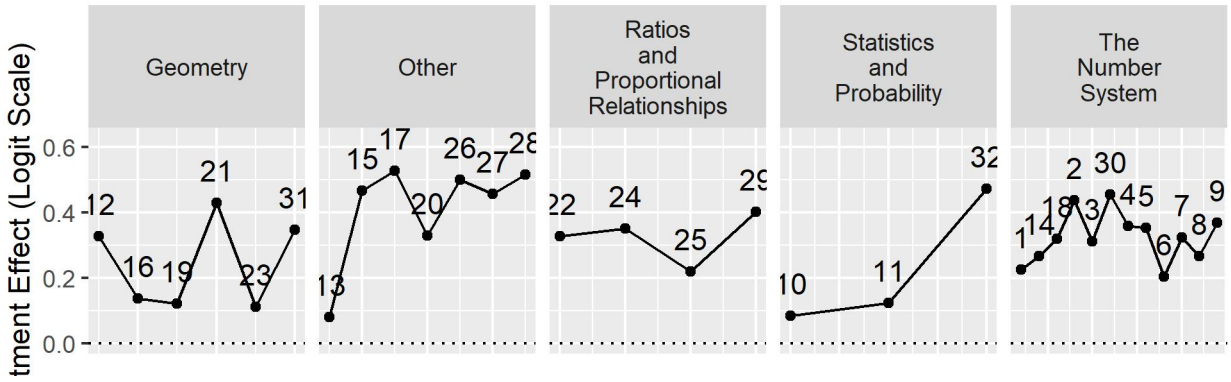

Figure 4: Estimated treatment effects of ASSISTments for each multiple choice posttest item, arranged according to CCSS, as in Table 2. The "Other" category includes Functions and the two Mathematical Practice standards, "make sense of problems and persevere in solving them" and "reason abstractly and quantitatively".

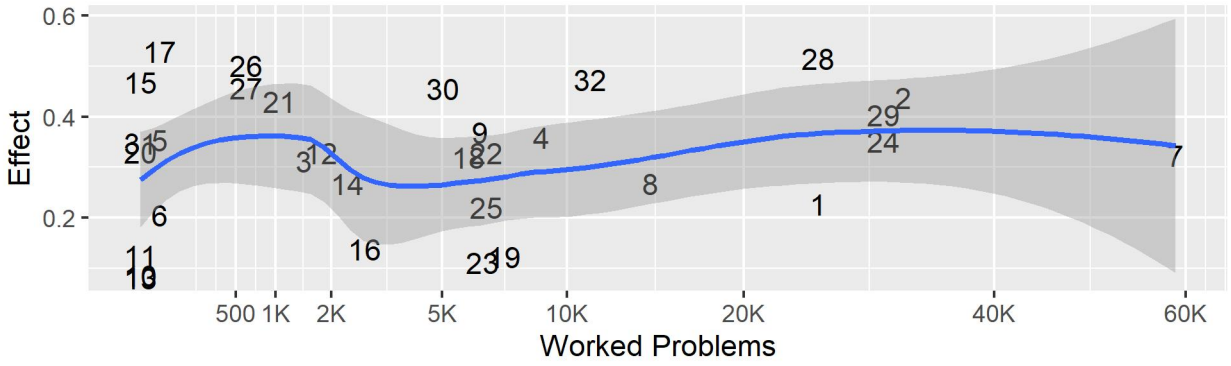

Figure 5: Estimated effects on multiple-choice TerraNova items plotted against the number of related ASSISTments problems that students in the treatment arm worked over the course of the study. The X-axis is plotted on the square-root scale, and a non-parametric loess fit is added for interpretation.

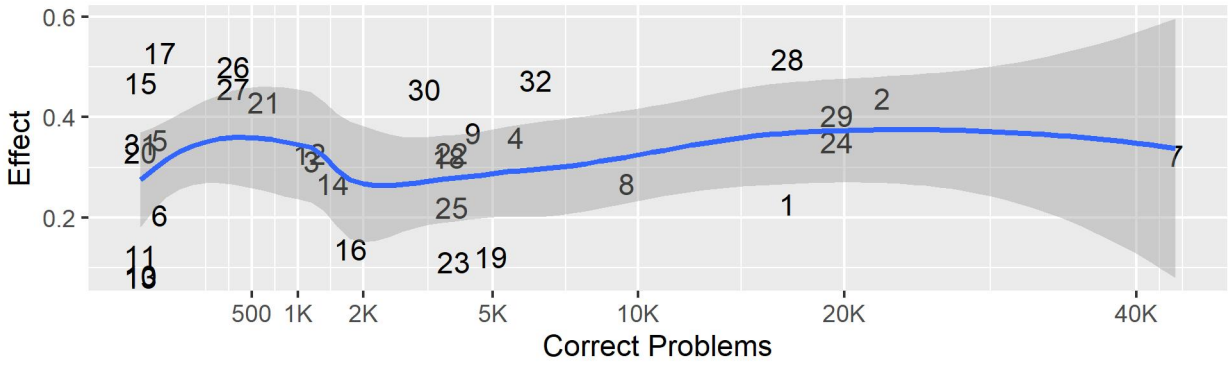

Figure 6: Estimated effects on multiple-choice TerraNova items plotted against the number of related ASSISTments problems that students in the treatment arm worked correctly over the course of the study. The X-axis is plotted on the square-root scale, and a non-parametric loess fit is added for interpretation.

ity involved in defining and determining item attributes and features, which is exacerbated by the fact that standardized tests generally cannot be made publicly available. Lastly, since student ITS work over the course of a study is necessarily post-treatment assignment, careful causal modeling (such as principal stratification [24]) may be necessary. Examining heterogeneity between item-specific treatment effects may play a larger role in helping to generate hypotheses about ITS effectiveness than in confirming hypotheses.

Despite those difficulties, the analysis here uncovered important information about the CTA1 and ASSISTments effects. First, the discovery that the effects vary between items is notable in itself. In our analysis of CTA1 we noticed that some of the largest effects—and differences between first and second-year effects— were for posttest items involving manipulating algebraic expressions and interpreting graphs. In our analysis of ASSISTments, we discovered a large difference between negative effects on open-ended questions and positive effects on multiple choice questions, and also that the largest effects were on problems requiring students to plug numbers into algebraic expressions.

We hope that this research will serve as a proof-of-concept and spur further work delving deeper into data we already have.

7. REFERENCES

[1] A. Agresti. *Categorical data analysis*, volume 482. John Wiley & Sons, 2003.

[2] J. R. Anderson, A. T. Corbett, K. R. Koedinger, and R. Pelletier. Cognitive tutors: Lessons learned. *The journal of the learning sciences*, 4(2):167–207, 1995.

[3] 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.

[4] H. S. Bloom, S. W. Raudenbush, M. J. Weiss, and K. Porter. Using multisite experiments to study cross-site variation in treatment effects: A hybrid approach with fixed intercepts and a random treatment coefficient. *Journal of Research on Educational Effectiveness*, 10(4):817–842, 2017.

[5] G. Casella. An introduction to empirical bayes data analysis. *The American Statistician*, 39(2):83–87, 1985.

[6] CTB/McGraw-Hill. Acuity algebra proficiency technical report. Monterey, CA, 2007.

[7] M. Escueta, V. Quan, A. Nickow, and P. Oreopoulos. Education technology: An evidence-based review. *NBER Working Paper*, (w23744), 2017.

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

[9] A. Gelman, J. Hill, and M. Yajima. Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, 5(2):189–211, 2012.

[10] N. T. Heffernan and C. L. Heffernan. The assistments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education*, 24(4):470–497, 2014.

[11] 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.

[12] R. Karam, J. F. Pane, B. A. Griffin, A. Robyn, A. Phillips, and L. Daugherty. Examining the implementation of technology-based blended algebra i curriculum at scale. *Educational Technology Research and Development*, 65(2):399–425, 2017.

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

[14] J. M. Montgomery, B. Nyhan, and M. Torres. How conditioning on posttreatment variables can ruin your experiment and what to do about it. *American Journal of Political Science*, 62(3):760–775, 2018.

[15] C. N. Morris. Parametric empirical bayes inference: theory and applications. *Journal of the American statistical Association*, 78(381):47–55, 1983.

[16] National Governors Association Center for Best Practices, Council of Chief State School Officers. Common core state standards: Mathematics, 2010.

[17] 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.

[18] 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.

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

[20] G. Rasch. *Probabilistic models for some intelligence and attainment tests.* ERIC, 1993.

[21] S. W. Raudenbush and A. S. Bryk. *Hierarchical linear models: Applications and data analysis methods*, volume 1. sage, 2002.

[22] J. Roschelle, M. Feng, R. F. Murphy, and C. A. Mason. Online mathematics homework increases student achievement. *AERA open*, 2(4):2332858416673968, 2016.

[23] A. Sales, T. Patikorn, and N. T. Heffernan. Bayesian partial pooling to improve inference across a/b tests in edm. In *Proceeding of the Educational Data Mining Conference*, 2018.

[24] A. Sales, A. Wilks, and J. Pane. Student usage predicts treatment effect heterogeneity in the cognitive tutor algebra i program. In *Proceedings of the 9th International Conference on Educational Data Mining. International Educational Data Mining Society*, pages 207–214, 2016.

[25] W. J. van der Linden and R. K. Hambleton. *Handbook of modern item response theory*. Springer Science & Business Media, 2013.

APPENDIX

A. A SIMULATION STUDY OF MULTIPLE COMPARISONS

We ran a small simulation study testing [9]'s assertion that multiplicity corrections are unnecessary when estimating different effects from BLUPs in a multilevel model. [9] stated their case in terms of fully Bayesian models, whereas we used an empirical Bayesian approach that may differ somewhat.

In our simulation, in each simulation run, we generated data on $Nexpr$ experiments, where $Nexpr$ was a parameter we varied. In each experiment, there were $n = 500$ simulated subjects, half assigned to treatment and half to control. They were given "outcome" data $Y \sim N(0, 1)$, with no treatment effect.

We analyzed the experiment data in two ways. First, we estimated a p-value for each experiment separately, using t-tests. This is the conventional approach. Then, we we estimated a multilevel model:

$$Y_{ij} = \beta_0 + \gamma_{1j} Expr_j + \gamma_{2j} Trt_i + \epsilon_{ij}$$

where β_0 is an intercept, γ_{1j} are random intercepts for experiment, γ_{2j} is the treatment effect for experiment j, and ϵ_{ij} is a normally-distributed error term. $\boldsymbol{\gamma} \sim MVN\left(\{0, \gamma_{20}\}, \Sigma\right)$ where γ_{20} is the average effect across all experiments. The number of experiments in each simulation run, $Nexpr$, was varied from 5 to 40, in increments of 5. In each case, we estimated the familywise error rate, the probability of at least one statistically significant effect estimate (at $\alpha = 0.05$) across the $Nexpr$ experiments.

The results are in Figure 7. As expected, the familywise error rate increased rapidly when effects were estimated and tested separately in each of the $Nexpr$ experiments. When effects were estimated jointly in a multilevel model, in a way analogous to the method described in Section 3, the familywise error rate remained roughly constant as $Nexpr$ increased. However, the familywise error rate in the multilevel modeling approach was slightly elevated, ranging from roughly 0.05 to 0.075.

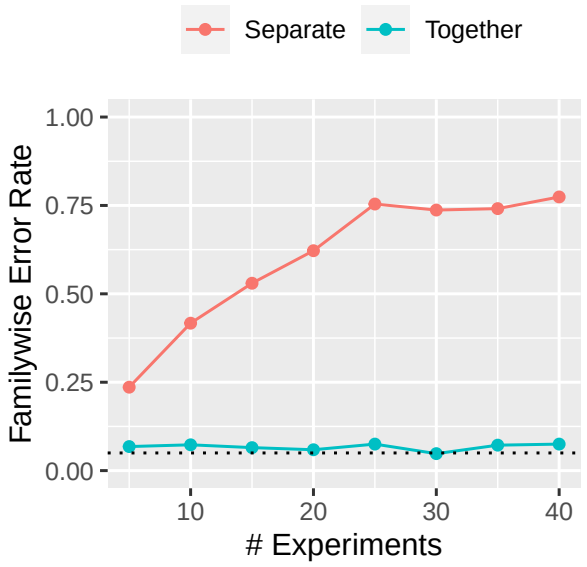

Figure 7: United we stand: results from a simulation of familywise error rate using separate t-tests for each experiment or using multilevel modeling.