Problem-Solving Behavior and EdTech Effectiveness: A Model for Exploratory Causal Analysis

Adam C. Sales
Worcester Polytechnic Institute
asales@wpi.edu

Kirk P. Vanacore
Worcester Polytechnic Institute
kpvanacore@wpi.edu

Hyeon-Ah Kang
University of Texas, Austin
hkang@austin.utexas.edu

Tiffany A. Whittaker
University of Texas, Austin
t.whittaker@austin.utexas.edu

ABSTRACT

The gold-standard evaluation of an educational technology product is a randomized study comparing students randomized to use a computer-based learning platform (CBLP) to students assigned to a "business as usual" condition, such as pencil-and-paper work, and estimating average treatment effects. However, not everyone uses the same CBLP in the same way—indeed, an individual may engage with CBLP in multiple different ways over the course of a study—and the platform's effectiveness may depend on how students are using it. This paper introduces a model that serves two aims: classifying different modes of problem-solving or engagement among CBLP users, and estimating varying program effectiveness with varying usage patterns. The model uses mixed-type problem-level variables, such as time spent, the number of errors committed, and the number of hints requested, to cluster each problem attempt by each student into one of a number of categories, using a model-based, probabilistic, latent profile model. Students differ from each other based on their probabilities of working on problems in each of the identified modes. Finally, the model uses a fully latent principal stratification approach to estimate varying treatment effects as a function of those probabilities. In this paper, we describe the model and estimation in detail and illustrate application using data from two large randomized field trials, one evaluating Cognitive Tutor Algebra I, and the other evaluating ASSISTments.

Keywords

Educational Technology Effectiveness, Student Behavior, Principal Stratification, Latent Profile Analysis

1. INTRODUCTION

One of the major advantages of computer-based learning platforms (CBLPs), compared with more traditional forms of instruction, is their flexibility. Learners with different and/or evolving needs or styles can use CBLP in the way that suits them best. For instance, a student who is unsure how to solve a particular math problem may have the option of requesting a hint or another form of just-in-time help. The same student may also attempt the problem multiple times, possibly making multiple errors and receiving immediate feedback, before figuring it out and entering the correct answer. Of course, students differ in how often and when they struggle to solve problems— while most students will find some problems easy and other problems hard, students will vary both in terms of which particular problems pose difficulty, and the ratio of difficult to easy problems. By offering struggling students multiple types of just-in-time help and feedback, CBLPs can accommodate this type of diversity better than pencil-and-paper programs.

On the other hand, the flexibility offered by CBLPs can complicate our understanding of their effectiveness. In the past two decades, several math CBLPs have been evaluated in large randomized controlled trials (RCTs), in which students or classrooms or schools were randomized between "business as usual" (BAU) mathematics instruction and computer-based learning. At the end of a year of instruction, the students complete a posttest gauging their learning. These randomized field trials are widely considered to be the most rigorous and reliable form of evidence for educational effectiveness for two principal reasons: first, if there is a statistically significant difference in average test scores across conditions, the randomization of treatment assignment allows researchers to rule out alternative, non-causal explanations such as confounding. Second, because field trials take place in real-life conditions across a diverse group of students and over a relatively long period of time, effectiveness results from field trials arguably reflect the type of effectiveness that educators and policymakers truly care about, and are less likely to reflect idiosyncrasies due to short-term effects, experimental conditions, or study participants.

The central goal of RCTs, both by design and by statistical necessity, is to estimate an average treatment effect—the effect on posttest scores of randomizing students to the CBLP instead of BAU, averaged across all of the students in the experiment [8]. This is arguably the most policy-relevant quantity to estimate if a CBLP is to be adopted by entire schools or districts. On the other hand, it is somewhat unsatisfying scientifically: if students are indeed using CBLPs in very different ways, shouldn't the effect of

A. C. Sales, K. P. Vanacore, H.-A. Kang, and T. A. Whittaker. Problem-solving behavior and edtech effectiveness: A model for exploratory causal analysis. In B. Paaßen and C. D. Epp, editors, *Proceedings of the 17th International Conference on Educational Data Mining*, pages 385–395, Atlanta, Georgia, USA, July 2024. International Educational Data Mining Society.

https://doi.org/10.5281/zenodo.12729842

the program vary considerably? What does it mean to say that computer-assisted instruction is effective, if "computer-assisted instruction" can mean so many different things to different students, even if they are all using the same CBLP? The flexibility of CBLPs—one of their principal assets—poses a significant challenge in the science of their effectiveness.

This paper proposes a modeling framework to address this challenge, first by describing and summarizing variability in student problem-solving styles or strategies when using CBLP, and then by estimating varying treatment effects associated with these different strategies or styles. The model is exploratory in the sense that it is not driven by particular hypotheses about student choices or behaviors within the system, but rather tries to discover patterns in the data along with their association with effectiveness.

The first part of the model, summarizing student problem-level log data, uses a finite mixture model—specifically, a latent profile model [16]—which we interpret as a form of model-based cluster analysis or unsupervised learning [13]. Mixture models are an important part of the educational data mining toolkit, and have a long history in the research community (e.g. [30, 31, 10, 14, 32, 20]). In this paper, we focus on one of the models described in [9], a multilevel latent profile model that incorporates multiple problem-level measurements for each student, along with problem- and student-level random effects and student-level covariate information. Briefly, it uses problem-level data to identify clusters of worked problem instances, which we will interpret as problem-solving modes, and estimates each student's probability of working a problem in each mode. In other words, the model imagines that each time a student begins a new problem, they flip a personalized biased coin (or roll a biased die) to decide how to work that problem.

The second component of our model is causal and is based on the "fully-latent principal stratification" framework (e.g. [25, 11, 33]). This framework considers a latent variable summarizing some aspect of program implementation as a measure of *potential* implementation—that is, a measurement of how students *would* implement a program if randomized to the treatment condition. Potential implementation is defined, though not observed, for students randomized to the control condition; hence, it can be imputed. Principal stratification estimators compare students randomized to the treatment condition, who implemented the program in a particular way, to students in the control condition who would have implemented the program in the same way, had they been given the opportunity.

We will introduce the combination of these two components, the latent profile and principal stratification models, in the context of an analysis of similar datasets from two different RCTs. One RCT, first described in [19], contrasts the Cognitive Tutor Algebra I curriculum [2], produced by Carnegie Learning and since superseded by Mathia, with traditional pencil-and-paper algebra instruction. The second, first described in [3], is part of a larger RCT contrasting four different CBLPs. We focus on the contrast between the ASSISTmenents CBLP [6] with immediate feedback and just-in-time tutoring, and a pared-down version of ASSISTments, designed to mimic pencil-and-paper work, in which students received feedback on problem correctness only after completing their assignment, and had no access to just-in-time tutoring.

The following section describes these two RCTs, and their associated datasets, in more detail. Section 3 describes the latent profile model of [9], and section 4 describes the causal principal stratification model. Section 5 describes the results of fitting the models to the two datasets, and Section 6 concludes.

2. DATASETS: COGNITIVE TUTOR ALGEBRA I AND ASSISTMENTS FIELD TRIALS

We will describe our model and illustrate it using similar datasets from two different RCTs, evaluating the Cognitive Tutor Algebra I curriculum and the ASSISTments online homework platform. In this section we will briefly describe both CBLPs and associated datasets.

2.1 Cognitive Tutor Algebra I

The Cognitive Tutor [2], originally developed at Carnegie Mellon University, then developed and run by Carnegie Learning, and now superseded by Carnegie Learning's Mathia program, was one of the first widely successful intelligent tutoring systems. The data for our study comes from a randomized effectiveness study sponsored by the US Department of Education and conducted between 2007 and 2009 by the RAND Corporation [19].

The original study included data from roughly 25,000 students across 73 high schools and 74 middle schools in seven states and across two school years (i.e. two cohorts of students). Schools were paired based on a range of factors including level (middle or high school), size, district, and prior achievement, and then randomized to use the CTA1 curriculum or continue with business as usual for the next two years. At the end of each academic year, all students taking algebra 1 in the treatment schools and matched controls took a standardized algebra 1 posttest. In the high school stratum, [19] estimated treatment effects of -0.1 standard deviations (95% CI -0.3–0.1) in year 1 and 0.21 standard deviations (95% CI 0.01–0.41) in year 2.

To reduce the considerable computational burden of our method, we analyzed a relatively small slice of that dataset—the first-year high school cohort. We further subsetted the log data from the group assigned to the treatment condition, focusing on a particular unit of the curriculum, *equation solver level 1*, and one day of student usage for each student. This unit was one of the most frequently assigned units, and its problems showed relatively similar measurement properties. If a student worked on the *equation solver level 1* unit across multiple days, we chose the one day that the student worked on the greatest number of problems. Finally, we dropped data from treatment schools in which fewer than 25% of students worked on problems from the *equation solver level 1* unit, along with their matched controls. All in all, we used data from 2,593 students in the treatment group and 2,130 in the control group. We used data from 107,933 worked problems from students in the treatment

group, drawn from problems testing 18 different knowledge components. We used three measurements for each worked problem: the total amount of time spent, the number of errors committed, and the number of hints requested. Finally, we have baseline covariate data for students in both treatment and control arms: the student's grade (parsed as 9th grade or higher), race (white/Asian, Black/multiracial, or Hispanic/American Indian/Alaskan Native), sex, free or reduced-price lunch status (FRL), English language learner status (ELL) and a pretest score. We imputed missing covariate data with a random forest algorithm [29] and included indicators for missing pretest or FRL. We also had school and randomization pair IDs.

2.2 ASSISTments

The ASSISTments system [6] is a free online homework system in which teachers can assign problems from popular open curricula for students to work online. Students receive immediate feedback on errors and have access to hints or explanations; teachers receive reports on student performance that they can review before teaching. ASSISTments has been shown to be effective in two large RCTs [23, 4], when compared with pencil-and-paper work.

The data for the current study [17] came from a third large RCT that compares four different CBLPs, ASSISTments, two gamified conditions, and an active control condition consisting of ASSISTments but with immediate feedback and just-in-time support disabled. The study consisted of data from 1,850 students in a single school district who were individually randomized between the four conditions. At the end of the school year, the students took an online 10-item posttest. The impact analysis [3] compared the active control condition to the three others, and found effects of 0.338 problems (95% CI: 0.05–0.63) and 0.56 problems (95% CI 0.24–0.87) for the two gamified conditions, and 0.24 problems (95% CI -0.07–0.55) for ASSISTments.

For this study, we used data from ASSISTments and the active control, including data 747 students (381 in treatment and 366 in control). For the treatment students, we collected data from all 68,524 problems worked. To compare with the CTA1 data, we aggregated log data to the problem level and calculated the same three measures: total time spent and numbers of hints and errors. We also had covariate data for all 747 students, including scores on a 10-item pretest, prior (5th grade) standardized mathematics test scores, baseline measures of math anxiety and math self-efficacy, time spent on the pretest, race (White, Asian, Hispanic, or Other), ELL status, and sex, along with school IDs.

3. PROBLEM-SOLVING MODES: A MULTIDIMENSIONAL, MULTILEVEL MIXTURE MODEL OF STUDENT LOG DATA

3.1 Assumptions for a Measurement Model

While using CBLP, each student i works on a potentially different set, number, and/or sequence of problems $t = 1, \ldots, T_i$. The total number of problems worked by student i, T_i, varies between students. For simplicity, let denote $T = \max(T_i : i = 1, \ldots, N)$ and we use it as a generic notation for the number of problems.

We assumed that each problem t was worked by student i in one of $M \geq 2$ modes, denoted as $S_{it} = 1, \ldots, M$, driving the student's interaction with the CBLP. The value of S can vary between students and across worked problems. As a student works on problem t in mode $S_{it} = m$, a set of observable indicating variables, $\boldsymbol{V}_{it} = (v_{ijt} : j = 1, \ldots, J)$ (e.g. interaction time), will exteriorize the latent state. Thus, the collection of indicator variables over time, $\boldsymbol{V}_i = (\boldsymbol{V}_{it} : t = 1, \ldots, T)$, forms a multivariate cross-sectional time series. If time-invariant baseline covariates are available for students, $\boldsymbol{X}_i = (x_{ik} : k = 1, \ldots, K)$, they can be included when modeling the density or mass function of S_{it}.

We assume that the observed problem-level variables $\boldsymbol{V}_{it}$ (i.e., the indicators at problem t) are conditionally independent given the student's latent state, S_{it}, i.e.

$$v_{ijt} \perp v_{ij't} | S_{it} \text{ for } j \neq j' \tag{1}$$

This local independence assumption is foundational for mixture modeling—the hope is that a latent variable can be found that explains the observed dependence between different measurements taken on the same student working on the same problem. The probability that each individual student works a problem in mode m is denoted as

$$\pi_{im} \equiv P(S_{it} = m \,|\, \eta_i,\, \boldsymbol{X}_i = \boldsymbol{x}_i) \tag{2}$$

Conditional on π, we assume $S_{it} \perp S_{it'}$ for $t \neq t'$. This is a strong assumption, and likely false, but quite useful in simplifying an otherwise quite complex model. [9] describes two other classes of models for the same data structure, a latent transition model and hidden Markov models that do not make this assumption; descriptions of usage patterns in [9] broadly agreed between the latent profile model, which assumes independence across problems, and the other two modeling frameworks. Finally, we assume that the problems are weakly invariant across the students and times (i.e., measurement invariance).

3.2 Random-Effect Latent Class Model

We identify latent states, or modes $m = 1, \ldots, M$ with latent class modeling (LCM; often, the term "latent class model" is applied to models of categorical data and latent profile model refer to models of continuous data; our data includes both types, so we use both terms interchangeably.) The LCM includes two sub-models: (i) a measurement model that describes the conditional probability of indicators given a latent state, $P(\boldsymbol{V}_{it} \,|\, S_{it} = m)$, and (ii) a structural model that describes the probability of a latent state, $P(S_{it} = m)$ $(m = 1, \ldots, M)$.

The measurement model describes the probability of $\boldsymbol{v}_{it}$ given the latent state S_{it}. A measurement model for a variable j is parameterized as

$$P_j(v_{ijt} \,|\, S_{it} = m,\, \delta_{jt}) = f_j(\boldsymbol{\psi}_{jm},\, \delta_{jt}), \tag{3}$$

where f_j takes a functional form specific to the indicator j, $\boldsymbol{\psi}_{jm}$ contains parameters of f_j, and δ_{jt} is a random problem effect that models the idiosyncratic effect of problem t on the indicator j. The functional form of f_j is defined according to the type of the variable.

For the log data in this study, we specified a measurement model for each indicator as follows. The total time spent by

student i on problem t, v_{i1t}, was modeled by a log-normal density:

$$\log v_{i1t} \,|\, S_{it} = m,\, \delta_{1t} \sim \mathcal{N}\left(\delta_{1t} + \mu_{1m},\, \sigma^2_{1m}\right) \tag{4}$$

so that both the mean and standard deviation of the distribution vary with the latent states.

The count variables—the number of errors, v_{i2t}, and the number of hints requested, v_{i3t}—were treated as ordinal variables, with categories zero, one, and more than one, and modeled via ordinal logistic regression. For example, the probability distribution of the number of errors, v_{i2t}, was modeled as

$$P(v_{i2t} = h \,|\, S_{it} = m,\, \delta_{2t}) = \begin{cases} 1 - \text{logit}^{-1}\left(\mu_{2m} + \delta_{2t} - c_{21}\right) & h = 0 \\ \text{logit}^{-1}\left(\mu_{2m} + \delta_{2t} - c_{21}\right) & \\ \quad -\text{logit}^{-1}\left(\mu_{2m} + \delta_{2t} - c_{22}\right) & h = 1 \\ \text{logit}^{-1}\left(\mu_{2m} + \delta_{2t} - c_{22}\right) & h > 1 \end{cases} \tag{5}$$

with the location parameter μ_{2m} varying across the latent states, and the step parameters satisfying $c_{22} > c_{21} > 0$. As above, δ_{2t} is the problem-level random effect for errors ($j = 2$). The inverse-logit function is $\text{logit}^{-1}(x) = 1 - \exp(-x)^{-1}$. The number of hints was modeled analogously with μ_{3m}, δ_{3t}, c_{31}, and c_{32} each replacing μ_{2m}, δ_{2t}, c_{21} and c_{22}.

For binary variables, common practice is to use the Bernoulli distribution with a probit or logit link. The ordinal variables are typically modeled by cumulative probability functions such as proportional-odds models, adjacent-categories, or continuation-ratio logit models [1]. The count and continuous variables can be modeled by each Poisson regression and Gaussian model.

The random-effect term, δ_{jt}, in (3), (4), and (5) is parameterized for each j such that $\delta_{jt} \sim \mathcal{N}(0, \sigma^2_{\delta_j})$ and $\sigma^2_{\delta_j} = \text{Var}(\delta_{jt};\, t = 1, \ldots, T)$ models the variance across the problems in indicator j. The random-effect terms from the three indicators were jointly modeled by a trivariate normal distribution: $\boldsymbol{\delta}_t = (\delta_{1t}, \delta_{2t}, \delta_{3t})^\top \sim \mathcal{N}_3(\mathbf{0},\, \boldsymbol{\Sigma}_{\boldsymbol{\delta}})$. These random effects weaken the assumption of measurement invariance between problems, by allowing the base rate of time spent, hints, or errors to vary between problems. In the CTA1 dataset, for the sake of parsimony, we grouped the problems by homogeneous skill sets—termed *knowledge components* [22]—so that there were only 18 values of $\boldsymbol{\delta}_t$. Problems with the same set of knowledge components were assumed to be similar in terms of difficulty, time spent, or hints requested. In the ASSISTments dataset, in which the set of problems worked was more homogeneous, we allowed δ to take a different value for each individual problem.

The structural model describes the probability that a student exhibits a certain latent state at any moment. In the current setting, the random-effect LCM can be formulated as follows. Let η_i model the idiosyncratic effect of student i on the state probability. If covariates are available, the probability of a latent state m can be modeled as

$$\pi_{im} \equiv P(S_{it} = m \,|\, \eta_i,\, \boldsymbol{X}_i = \boldsymbol{x}_i) = \frac{\exp\left(\beta_{m0} + \boldsymbol{\beta}_m^\top \boldsymbol{x}_i + \eta_i\right)}{\sum_{l=1}^{M} \exp\left(\beta_{l0} + \boldsymbol{\beta}_l^\top \boldsymbol{x}_i + \eta_i\right)}, \tag{6}$$

where β_{m0} determines the conditional probability of the latent state m when $\boldsymbol{X}_i = \mathbf{0}$ and $\eta_i = 0$, and $\boldsymbol{\beta}_m$ models the covariate effects on the logit. To identify the model parameters, we assume that $\eta_i \sim \mathcal{N}(0,\, \sigma^2_\eta)$ with $\sigma^2_\eta = \text{Var}(\eta_i;\, i = 1, \ldots, N)$ modeling the magnitude of extra variance from the individuals (i.e., beyond covariate effects).

Integrating the two sub-models, the random-effect LCM becomes a finite mixture model:

$$P(\boldsymbol{V} \,|\, \boldsymbol{X},\, \boldsymbol{\eta},\, \boldsymbol{\delta}) = \prod_{i=1}^{N} \prod_{t=1}^{T} \sum_{S_t \in \mathcal{S}} P(S_t \,|\, \eta_i,\, \boldsymbol{X}_i) P(\boldsymbol{V}_{it} \,|\, S_t,\, \boldsymbol{\delta}_t), \tag{7}$$

where $\boldsymbol{V}$, $\boldsymbol{X}$, and $\boldsymbol{\eta}$ each denote an array of variables for all students, and $\boldsymbol{\delta} = (\boldsymbol{\delta}_t : t = 1, \ldots, T)$ where $\boldsymbol{\delta}_t = (\delta_{jt} : j = 1, \ldots, J)$.

4. CAUSAL FRAMEWORK: FULLY-LATENT PRINCIPAL STRATIFICATION (FLPS)

The model described in the previous section, and also in [9], results in a description of $M \geq 2$ modes in which students solve practice problems, as well as a student-level parameter vector π_{im}, $m = 1, \ldots, M$, the probability that student i works on a problem in mode m. In this section, we address how $\boldsymbol{\pi}_i$ relates to program effectiveness—that is, do students who are more likely to solve problems in a particular mode tend to experience smaller or larger treatment effects? In our applied examples, $M = 2$, so $\boldsymbol{\pi}_i = \{\pi_{i1}, \pi_{i2}\} = \{\pi_{i1}, 1 - \pi_{i1}\}$. Hence, we can consider the parameter $\boldsymbol{\pi}_i$ to be unidimensional, $\pi_i = Pr(S_{it} = 1 | \eta_i, \boldsymbol{X}_i)$.

Let Y_i be student i's posttest score, and let $Z_i = 1$ if student i is randomized to the treatment condition, with $Z_i = 0$ if i is randomized to control. Then, following [15, 24], let Y_i^T be the posttest score i would achieve if assigned to the treatment condition, and Y_i^C be the score i would achieve if assigned to control, so $Y_i = Z_i Y_i^T + (1 - Z_i) Y_i^T$. Finally, define the treatment effect for student i as $\tau_i = Y_i^T - Y_i^C$. Since only one of $\{Y_i^C, Y_i^T\}$ is ever observed for each i, τ_i is unobserved; however, under some circumstances we can estimate averages or expectations of τ.

Given π_0, a particular probability of working problems in mode $M = 1$, our goal is to estimate

$$E[\tau | \pi = \pi_0, Z = 1] = E[Y^T | \pi = \pi_0, Z = 1] - E[Y^C | \pi = \pi_0, Z = 1] \tag{8}$$

the effect experienced by students assigned to the treatment condition with $\pi = \pi_0$. Estimating the first term in (8) is relatively straightforward, since for students assigned to treatment, $Y = Y^T$ and π is identified based on their observed problem-level measurements $\boldsymbol{V}$. In contrast, the second term, $E[Y^C | \pi = \pi_0, Z = 1]$, barely even makes sense. First of all, Y^C is never observed when $Z = 1$. Second of all, students assigned to control—for whom Y^C is observed—cannot request hints or make multiple errors, so the problem-solving modes captured in $\boldsymbol{S}$, and hence in π, are irrelevant. In other words, even if we can identify a set of students with $Z = 1$ and $\pi = \pi_0$, in order to estimate their average treatment effect we need to identify a suitable comparison group.

The answer to this question, following the principal stratification literature [5, 18, 27], is to slightly re-define π_i as

$$\pi_i \equiv P(S_{it} = 1 \,|\, \eta_i,\, \boldsymbol{X}_i = \boldsymbol{x}_i, Z_i = 1) \tag{9}$$

That is, if student i *were* assigned to the treatment condition, what *would be* the probability that they work problems in mode 1? Stated this way, π is a potential value—students' potential of working problems in a particular mode, if given the opportunity. Since π is a latent variable, never directly observed, our analysis falls under the "fully-latent principal stratification" framework of [12].

While Y^T and π are observed and estimated, respectively, for students assigned to the treatment condition, they are unobserved for students assigned to the control condition, as are problem-level indicators $\boldsymbol{V}$ used to estimate π. However, as potential values, they are well-defined; we may consider them as missing data to be imputed.

At this point, randomization of treatment assignment becomes crucial. Because Z_i is randomized, it is independent of π_i, Y_i^C, and Y_i^T, as well as covariates $\boldsymbol{X}_i$:

$$\{\pi_i, Y_i^C, Y_i^T, \boldsymbol{X}_i\} \perp Z_i \tag{10}$$

This implies two useful facts: first, the distribution of π_i (which is random, due to its dependence on student random intercept ν_i) conditional on covariates $\boldsymbol{X}$ is the same in both treatment groups:

$$p(\pi_i|\boldsymbol{X}, Z_i = 1) = p(\pi_i|\boldsymbol{X}, Z_i = 0) = p(\pi_i|\boldsymbol{X})$$

That means that we can estimate a model $p(\pi_i|\boldsymbol{X}, Z_i = 1)$ using observed data in the treatment group, and extend that model to the control group. Second, the distributions of Y^C and Y^T are independent of Z_i—in particular,

$$\begin{aligned} E[Y^C|\pi = \pi_0, Z_i = 1] &= E[Y^C|\pi = \pi_0, Z_i = 0] \\ &= E[Y|\pi = \pi_0, Z_i = 0] \end{aligned}$$

If we could identify the subset of the students randomized to the control condition with $\pi = \pi_0$, then we could use their observed outcomes Y to estimate the pesky 2^{nd} term of (8), $E[Y^C|\pi = \pi_0, Z = 1]$.

In practice, we estimate causal effects conditional on π using a regression model of posttest scores Y as a function of treatment status Z, imputed or estimated $\hat{\pi}$, and covariates $\boldsymbol{X}$. Specifically, we fit the model

$$Y_i = \gamma_0 + \omega\hat{\pi}_i + Z_i(\tau_0 + \tau_1\hat{\pi}_i) + \sum_k \gamma_k X_{ik} + \epsilon_i \tag{11}$$

In this model, subject i's treatment effect is modeled as linear in π: $\tau_0 + \tau_1\pi_i$, so τ_0 represents the average effect for subjects with $\pi = 0$ and τ_1 represents the change in treatment effects as π increases. Meanwhile, ω captures the correlation between π and control potential outcomes Y^C: the extent to which students who often work problems in mode 1 would score higher (or lower) than other students, in the absence of any treatment. In the CTA1 analysis, we included school random intercepts in the outcome submodel (11) to account for the school-level randomization.

Putting it all together, our empirical strategy is to (1) estimate models (3) and (6) using log data and covariates from students assigned to the treatment condition, (2) use those models to (probabilistically) impute π for students in the control group, and lastly to use outcomes, covariates, and estimated or imputed π for students randomized to either condition to estimate principal effects $\tau(\pi) = E[\tau|\pi]$.

To ensure appropriate likelihood inference and propagation of errors, we fit submodels (7) and (11) simultaneously using a "No U-Turn" Markov Chain Monte Carlo (MCMC) sampler [7] in Stan [28] which we called from `R` [21].

5. RESULTS

For the sake of simplicity and proof-of-concept, we focused on estimating models with $M = 2$ problem-solving modes. Future work will develop models with $M > 2$. As it is, fitting the models was extremely computationally intensive: on a local multicore server, the CAT1 model took roughly nine days to run, and the ASSISTments model took approximately six days.

5.1 Measurement Models

Tables 1 and 2 show the estimated parameters from fitting submodel (3), (4), and (5) to CTA1 and ASSISTments log data, respectively. While the specific parameter values differ between the two CBLPs, both models appear to be capturing the same qualitative phenomenon. Students solving problems in State 1 spend more time per problem (with a larger between-problem variance), request more hints and make more errors than students solving problems in State 2. Loosely speaking, it appears students in State 1 are struggling more with the material than students in State 2. Interestingly, it appears that State 1 is much more common in ASSISTments, where it accounts for half or more of worked problems, than in CTA1, where it only describes about 20% of problems solved.

5.2 Modeling π

The left-hand columns of Tables3 and 4 give the results of model (6), predicting $logit(1-\pi)$ as a function of covariates, fit to the CTA1 and ASSISTments datasets, respectively. Note that the models predict the probability of working problems in State 2, with less struggle, rather than in state 1. In the CTA1 dataset, students from underrepresented minorities, boys, and students with lower pretest scores are all more likely to work problems in state 1—that is, struggle more often—than their peers.

Surprisingly, in the ASSISTments dataset, the only statistically significant coefficient is on student sex, suggesting boys are less likely to work problems in state 1 one than girls, the opposite sign as the analogous coefficient in the CTA1 analysis. There appears to be little relationship between pretest measures and the frequency of working on problems in state 1.

5.3 Principal Effects

The right-hand columns of Tables 3 and 4 give estimated coefficients from outcome regressions (11) in the CTA1 and ASSISTments datasets. Of particular interest are the coefficients labeled "tzero" and "tone," corresponding to τ_0 and τ_1 in the model. These parameterize the principal effect function $\tau_0 + \tau_1\hat{\pi}_i$. These functions are also plotted, with

	Notation	State 1	State 2	Difference
Time (mean)	μ_{1m}	0.988 (0.118)	-0.503 (0.117)	1.491 (0.008)
Time (SD)	σ_m	0.838 (0.005)	0.757 (0.003)	0.081 (0.005)
Error	μ_{2m}	2.002 (3.529)	-2.181 (3.530)	4.183 (0.038)
Hint	μ_{3m}	6.098 (3.902)	-6.066 (3.900)	12.165 (3.073)
Probability		0.203 (0.073)	0.797 (0.073)	

Table 1: Estimated parameters for the CTA1 measurement model

par	Notation	State 1	State 2	Difference
Time (mean)	μ_{1m}	0.349*** (0.035)	-0.233*** (0.035)	0.581*** (0.007)
Time (SD)	σ_m	0.881*** (0.004)	0.532*** (0.003)	0.349*** (0.005)
Error	μ_{2m}	1.228 (3.486)	-1.156 (3.486)	2.384*** (0.032)
Hint	μ_{3m}	3.693 (3.490)	-3.595 (3.493)	7.288*** (0.280)
Probability		0.545* (0.224)	0.455* (0.224)	

Table 2: Estimated parameters for the ASSISTments measurement model

	Dependent variable:	
	$\text{logit}(1-\pi)$	Posttest
	(1)	(2)
(Intercept)	1.228^{***}	-0.164
	(0.156)	(0.228)
ω		-1.950^{***}
		(0.335)
τ_0		-0.171
		(0.112)
τ_1		0.558
		(0.333)
Grade 10+	0.045	-0.089^{*}
	(0.065)	(0.036)
Black	-0.125^{*}	-0.060
	(0.059)	(0.035)
Hispanic	-0.217^{**}	-0.035
	(0.071)	(0.044)
Male	-0.098^{**}	0.011
	(0.033)	(0.019)
FRL	-0.052	-0.020
	(0.040)	(0.023)
ns(pretest)1	0.710^{***}	0.624^{***}
	(0.098)	(0.060)
ns(pretest)2	1.111^{***}	1.441^{***}
	(0.287)	(0.176)
ns(pretest)3	0.772^{*}	2.221^{***}
	(0.302)	(0.165)
ESL	-0.252	0.023
	(0.144)	(0.099)
No Pretest	0.062	0.173^{***}
	(0.054)	(0.030)
No FRL	-0.033	-0.092^{*}
	(0.072)	(0.040)
Note:	*p<0.05; **p<0.01; ***p<0.001	

Table 3: CTA1 regression results for models of π (6) and of posttest scores (11) as functions of covariates and (in the latter case) treatment assignment and π. Fixed effects for school randomization pair and random intercepts for school were included in the model but are omitted from the table.

	Dependent variable:	
	$\text{logit}(1-\pi)$	Posttest
	(1)	(2)
(Intercept)	0.641	-2.287
	(1.732)	(1.366)
λ		-0.044
		(0.431)
τ_0		0.382
		(0.316)
τ_1		-0.225
		(0.516)
Pretest	-0.009	0.316^{***}
	(0.060)	(0.042)
Gr. 5 State Test	-0.004	0.011^{***}
	(0.003)	(0.002)
Math Anx.	0.028	-0.012
	(0.024)	(0.017)
Math Self Eff.	-0.006	0.047^{*}
	(0.027)	(0.020)
No State Test	-0.111	-0.341
	(0.324)	(0.244)
log(Pretest Time)	0.221	-0.167
	(0.150)	(0.111)
Hispanic	-0.020	-0.137
	(0.386)	(0.291)
Asian	0.143	0.821^{***}
	(0.328)	(0.243)
Other Race	-0.171	0.315
	(0.368)	(0.295)
ESOL	0.120	0.571
	(0.404)	(0.336)
Male	0.786^{***}	-0.273
	(0.226)	(0.163)
Note:	*p<0.05; **p<0.01; ***p<0.001	

Table 4: ASSISTments regression results for models of π (6) and of posttest scores (11) as functions of covariates and (in the latter case) treatment assignment and π. Fixed effects for school were included in the model but were omitted from the table.

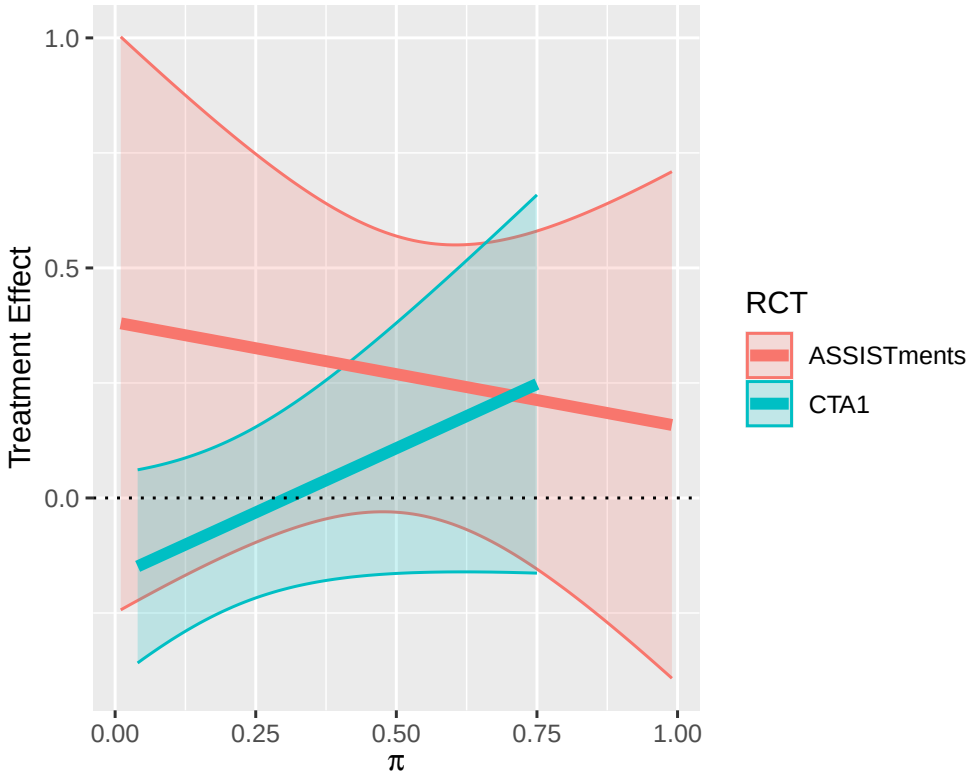

Figure 1: Estimated effect curves as a function of π for CTA1 and ASSIStments

confidence bands, in Figure 1. In CTA1, students who were more likely to work problems in State 1, and therefore struggle more often, appeared to experience greater benefits from the program. This might suggest that the struggle they were experiencing was productive (e.g. [34]) or that CTA1 was more beneficial to struggling students than for already-high-achieving students for other reasons. The results indicate that the data are also consistent with no effects and no heterogeneity, or slightly negative slopes as well, so speculation about the interpretation of the findings should be even more tempered. Conclusions must be even weaker for ASSISTments, where there is a very wide range of lines consistent with the data, including positive and negative slopes, as well as no effect at all.

6. DISCUSSION

In summary, this paper has introduced a novel method with broad applicability for bridging descriptive modeling of student usage with effectiveness studies in computer-based education. The potential impact of this method extends to enhancing our understanding of educational technology effectiveness, thereby contributing significantly to the field.

The most important limitation of the study is its dependence on a highly-parametric model. Unfortunately, at this stage the possible impacts of different types of model misspecification are unknown. However, there is also cause for optimism: [9], where our measurement model originated, included results from two other models of the same dataset, and all three models gave similar measurement results. That paper also includes other information on model selection and validity. We are currently working on a moment-based estimator in the mold of [26] that will rely less heavily on distributional assumptions. A moment-based estimator will also run much faster, addressing the second major limitation of this work.

Beyond those concerns, we hope to extend the framework to include models with more than two latent states, in order to provide deeper insights into the complexities of student behavior and learning outcomes. Lastly, incorporating different types of measurement models, such as those accounting for autocorrelation between a student's latent states or models analyzing more granular clickstream data, could offer a more nuanced understanding of student engagement and learning processes.

By addressing these research needs, future studies can further advance the field of educational technology effectiveness and contribute to the development of more sophisticated and impactful instructional strategies. This research has the potential to not only enhance educational practices but also improve learning outcomes for students across diverse educational settings.

7. ACKNOWLEDGMENTS

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D210036. The opinions expressed are those of the authors and do not represent views of the Institute or

the U.S. Department of Education.

The authors wish to thank Sooyong Lee, Erin Ottmar, John Pane, Neil Heffernan, and three anonymous reviewers for advice, comments, and help with the data.

8. REPLICATION MATERIALS

Replication data for the ASSISTments analysis can be obtained by following the instructions at `https://osf.io/r3nf2/`. Unfortunately, the CTA1 dataset is not public.

All analysis code, along with instructions for replicating the ASSISTments analysis, are available at `https://github.com/adamSales/LCA.PS.EDM`

APPENDIX

The CTA1 Pretest Model

The model for the analysis of CTA1 data included a natural spline for pretest scores, the coefficients of which (reported in Table 3) are nearly uninterpretable. Figure 2 plots the estimated conditional relationships between pretest scores and $logit(1-\pi)$ and Posttest scores, along with a 90% confidence band. The Y-axis records the contribution to the linear predictor of each model due to pretest.

Model Code

The following is the Stan code for the ASSISTments model. The code for the CTA model also includes school random effects and can be found at `https://github.com/adamSales/LCA.PS.EDM`.

```
data {
  int<lower=1> nworked; //number of worked items
  int<lower=1> nprob;      // number of items
  int<lower=1> nstud;      // number of students
  int<lower=1> ncov; // number of person-level covariates
  int<lower=0> hint[nworked];
  int<lower=0> err[nworked];
  real ltime[nworked];
  int<lower=1,upper=nprob> prob[nworked];
  int<lower=1,upper=nstud> stud[nworked];
  matrix[nstud,ncov] X;
  real Y[nstud];
  vector[nstud] Z;
  vector[3] zeros;
}
parameters {
  real meanTime[2];
  real<lower=0> sigTime[2];
  real effHint[2];
  real effErr[2];
  vector[3] probEff[nprob]; // hint, err, time
  corr_matrix[3] OmegaProb;
  vector<lower=0>[3] sigProb;
  real alpha;
  vector[nstud] studEff;
  real<lower=0> sigStud;
  vector[ncov] beta;
  ordered[2] cHint;
  ordered[2] cErr;
  real gamma0;
  real tzero;
  real lambda;
  real tone;
  vector[ncov] gamma;
  real<lower=0> sigY;
}
transformed parameters {
  cov_matrix[3] SigmaProb=
    quad_form_diag(OmegaProb, sigProb);
}
model{
 real yhat[nstud];
 vector[nstud] nu=inv_logit(alpha+X*beta+studEff);
  for(i in 1:nstud)
    yhat[i]=gamma0+tzero*Z[i]+lambda*nu[i]+
      tone*Z[i]*nu[i]+X[i,]*gamma;

// priors
 meanTime~normal(0,5);
 sigTime~normal(0,5);
 effHint~normal(0,5);
 effErr~normal(0,5);
 sigProb~normal(0,1);
 to_vector(beta)~normal(0,1);
tzero~std_normal();
tone~std_normal();
lambda~std_normal();
gamma~normal(0,5);
sigProb~normal(0,3);
sigStud~normal(0,3);
sigY~normal(0,5);

 studEff~normal(0,sigStud);
 probEff~multi_normal(zeros,SigmaProb);

 for(w in 1:nworked)
   target += log_sum_exp(
    log(nu[stud[w]])+
    ordered_logistic_lpmf(hint[w]|probEff[prob[w]][1]+
      effHint[1],cHint)+
    ordered_logistic_lpmf(err[w]|probEff[prob[w]][2]+
      effErr[1],cErr)+
    normal_lpdf(ltime[w]| probEff[prob[w]][3]+
      meanTime[1],sigTime[1]),
    log(1-nu[stud[w]])+
    ordered_logistic_lpmf(hint[w]|probEff[prob[w]][1]+
      effHint[2],cHint)+
    ordered_logistic_lpmf(err[w]|probEff[prob[w]][2]+
      effErr[2],cErr)+
    normal_lpdf(ltime[w]| probEff[prob[w]][3]+
      meanTime[2],sigTime[2])
   );

 Y~normal(yhat,sigY);
}
```

A. REFERENCES

[1] A. Agresti. *Categorical Data Analysis*. Wiley, New York, NY, 3rd. edition, 2012.

[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] L. E. Decker-Woodrow, C. A. Mason, J.-E. Lee,

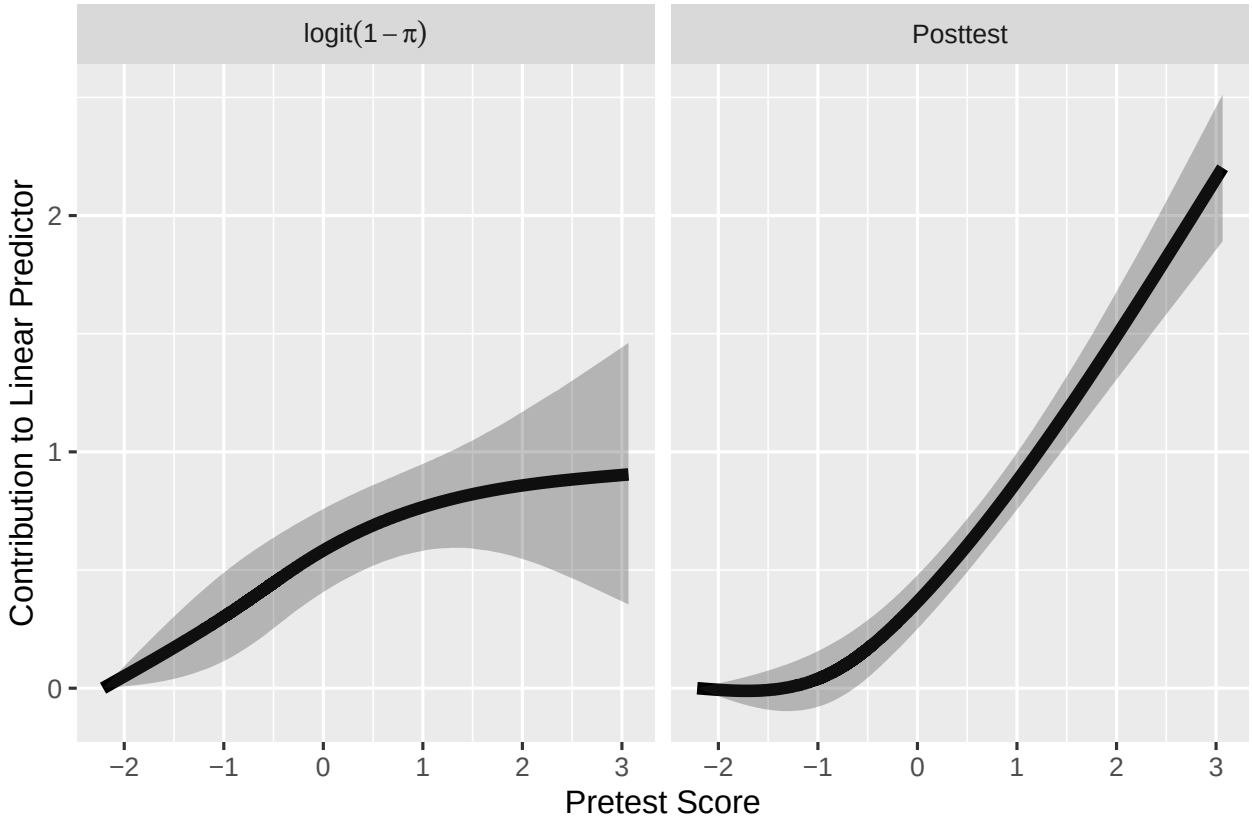

Figure 2: A plot of the estimated spline curves reported in Table 3

J. Y.-C. Chan, A. Sales, A. Liu, and S. Tu. The impacts of three educational technologies on algebraic understanding in the context of covid-19. *AERA open*, 9:23328584231165919, 2023.

[4] M. Feng, C. Huang, and K. Collins. Promising long term effects of assistments online math homework support. In *International Conference on Artificial Intelligence in Education*, pages 212–217. Springer, 2023.

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

[6] 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, Oct 2014.

[7] M. D. Hoffman and A. Gelman. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

[8] G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[9] H.-A. Kang, A. C. Sales, and T. A. Whittaker. Flow with an intelligent tutor: A latent variable modeling approach to tracking flow during artificial tutoring. *Behavior Research Methods*, 56(2):615–638, Feb 2024.

[10] T. Le Quy, G. Friege, and E. Ntoutsi. A review of clustering models in educational data science toward fairness-aware learning. *Educational Data Science: Essentials, Approaches, and Tendencies: Proactive Education based on Empirical Big Data Evidence*, pages 43–94, 2023.

[11] S. Lee, S. Adam, H.-A. Kang, and T. A. Whittaker. Fully latent principal stratification: Combining ps with model-based measurement models. In *The Annual Meeting of the Psychometric Society*, pages 287–298. Springer, 2022.

[12] S. Lee, S. Adam, H.-A. Kang, and T. A. Whittaker. Fully latent principal stratification: Combining ps with model-based measurement models. In *The Annual Meeting of the Psychometric Society*, pages 287–298. Springer, 2022.

[13] F. Liu, D. Yang, Y. Liu, Q. Zhang, S. Chen, W. Li, J. Ren, X. Tian, and X. Wang. Use of latent profile analysis and k-means clustering to identify student anxiety profiles. *BMC psychiatry*, 22(1):1–11, 2022.

[14] R. Maqsood, P. Ceravolo, C. Romero, and S. Ventura. Modeling and predicting students' engagement behaviors using mixture markov models. *Knowledge and Information Systems*, 64(5):1349–1384, 2022.

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

[16] D. Oberski. Mixture models: Latent profile and latent class analysis. *Modern statistical methods for HCI*, pages 275–287, 2016.

[17] E. Ottmar, J.-E. Lee, K. Vanacore, S. Pradhan, L. Decker-Woodrow, and C. A. Mason. Data from the efficacy study of from here to there! a dynamic technology for improving algebraic understanding. *Journal of Open Psychology Data*, 11(1):5, Apr 2023.

[18] 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.

[19] J. F. Pane, B. A. Griffin, D. F. McCaffrey, and R. Karam. Effectiveness of Cognitive Tutor Algebra I at scale. *Educ. Eval. Policy Anal.*, 36(2):127–144, 2014.

[20] J. Park, R. Yu, F. Rodriguez, R. Baker, P. Smyth, and M. Warschauer. Understanding student procrastination via mixture models. *International Educational Data Mining Society*, 2018.

[21] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.

[22] 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.

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

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

[25] A. C. Sales and J. F. Pane. Student log-data from a randomized evaluation of educational technology: A causal case study. *Journal of Research on Educational Effectiveness*, 14(1):241–269, Jan 2021. arXiv:1808.02528 [stat].

[26] A. C. Sales, K. P. Vanacore, and E. R. Ottmar. Geepers: Principal stratification using principal scores and stacked estimating equations. *arXiv preprint arXiv:2212.10406*, 2022.

[27] A. C. Sales, A. Wilks, and J. F. Pane. Student usage predicts treatment effect heterogeneity in the cognitive tutor algebra i program. In *Proceedings of the 12th International Conference on Educational Data Mining*. www.educationaldatamining.org, 2016.

[28] Stan Development Team. RStan: the R interface to Stan, 2016. R package version 2.14.1.

[29] D. J. Stekhoven and P. Buehlmann. Missforest - non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

[30] M. Streeter. Mixture modeling of individual learning curves. *International Educational Data Mining Society*, 2015.

[31] M. Tanai, J. Kim, and J. H. Chang. Model-based clustering analysis of student data. In *Convergence and Hybrid Information Technology: 5th International Conference, ICHIT 2011, Daejeon, Korea, September 22-24, 2011. Proceedings 5*, pages 669–676. Springer, 2011.

[32] K. Vanacore, E. Ottmar, A. Liu, and A. Sales. Remote monitoring of implementation fidelity using log-file data from multiple online learning platforms. *Journal of Research on Technology in Education*, pages 1–21, 2024.

[33] K. Vanacore, A. Sales, A. Liu, and E. Ottmar. Benefit of gamification for persistent learners: Propensity to replay problems moderates algebra-game effectiveness. In *Proceedings of the Tenth ACM Conference on Learning@ Scale*, pages 164–173, 2023.

[34] H. K. Warshauer. Productive struggle in middle school mathematics classrooms. *Journal of Mathematics Teacher Education*, 18:375–400, 2015.