Statistical Consequences of using Multi-armed Bandits to Conduct Adaptive Educational Experiments

Anna N. Rafferty
Department of Computer Science
Carleton College
arafferty@carleton.edu

Huiji Ying
Department of Computer Science
Carleton College
yingh@carleton.edu

Joseph Jay Williams
Department of Computer Science
University of Toronto
williams@cs.toronto.edu

Randomized experiments can provide key insights for improving educational technologies, but many students may experience conditions associated with inferior learning outcomes in these experiments. Multi-armed bandit (MAB) algorithms can address this issue by accumulating evidence from the experiment as it runs and modifying the experimental design to assign more helpful conditions to a greater proportion of future students. Using simulations, we explore the statistical impact of using MAB algorithms for experiment design, focusing on the tradeoff between acquiring statistically reliable information from the experiment and benefits to students. We consider how temporal biases in patterns of student behavior may impact the results of MAB experiments, and model data from ten previous educational experiments to demonstrate potential impacts of MAB assignment. Results suggest that MAB experiments can lead to much higher average benefits to students than traditional experimental designs, although at least twice as many participants are needed for acceptable statistical power. Using an optimistic prior distribution for the MAB algorithm mitigates the loss in power to some extent, without significantly reducing benefits to students. Additionally, longer experiments with MAB assignment still assign fewer students to a less effective condition than typical practice of a shorter experiment followed by choosing one condition for all future students. Yet, MAB assignment does increase false positive rates, especially if there are temporal biases in when students enter the experiment. Caution must thus be used when interpreting results from MAB assignment in cases where students can choose when to participate in the experiment. Overall, in scenarios where student characteristics do not vary over time, MAB experimental designs can be beneficial for students and effective for reliably determining which of two differing conditions is better given large sample sizes.

Keywords: multi-armed bandit algorithms, adaptive experimentation, experimental design, simulation, statistical hypothesis testing, educational experiments

1. Introduction

Educational technologies can lower the barriers to conducting randomized controlled experiments. These experiments can be used to answer research questions in the learning sciences and education, as well as to address practical questions of interest to curriculum designers. For example, an experiment could be used to investigate a question like whether video or text explanations are associated with better performance on later problems. Typically, the experimental design would assign half of students to view video explanations and half to view text explanations. This experimental design fulfills the goal of collecting as much information as possible for the researcher: given no prior information about the conditions, an experimental design that splits participants equally among conditions maximizes statistical power. This provides the researcher the best chance to identify which condition is most effective, potentially enabling the development of future technologies that are more effective for all learners. However, it is indifferent to benefits for learners within the current experiment. Even if one of the conditions is clearly less effective than the other, half of students will experience that condition.

Multi-armed bandit (MAB) algorithms offer a potential alternative that could benefit learners in the experiment by considering the utility of different versions of content. MABs select a version for each user by optimizing expected *reward*. Reward is specific to the problem the MAB is applied to; in the context of an experiment, the reward is the outcome that is being used to define the effectiveness of the conditions. For example, in the experiment comparing how text versus video explanations affect performance on later problems, the reward could be defined as the score on the next problem after viewing the explanation. The MAB algorithm would then select condition assignments to maximize the proportion of students who got the next problem right. MAB algorithms are designed to solve online decision problems, where decisions are made sequentially and information about an option is acquired only by choosing that particular option (in contrast to supervised learning). Traditionally, MABs have been used for applications like selecting online ads (Tang et al., 2013), but they have also been used in education to choose what version of a system to give to each learner (Liu et al., 2014; Williams et al., 2016). Since different learners interact with the system at different times, the success (or failure) of a learner in a particular version of the system can be used to inform what version of the system to give to the next learner. If the version of the system is viewed as an experimental condition, then the algorithm will direct more students to more effective conditions over time. Because MAB algorithms make decisions sequentially, they are particularly relevant to decision making about alternative pedagogies in educational technologies, where students may access materials asynchronously. Experiments using MAB assignment have been conducted within course quizzes with the aim of increasing benefits to students (Williams et al., 2018).

However, maximizing benefits to students is not directly aligned with gaining information about differences between conditions (Erraqabi et al., 2017; Liu et al., 2014). While many MAB algorithms make probabilistic guarantees about selecting options that maximize reward over time, this is not equivalent to accurately estimating the effectiveness of each condition: because there are dependencies between previous results and future assignment choices, data collected by the algorithm can result in measurement error, in which the expected estimate of each condition's effectiveness is not equal to the true effectiveness (Erraqabi et al., 2017). Additionally, the uneven division of students across conditions that allows MABs to increase benefits to students will likely decrease statistical power to detect effects, limiting the inferences that can be drawn from experiments. Work like Liu et al. (2014) and Erraqabi et al. (2017) has suggested ways to adjust the tradeoff between benefits to research and to students, examining measurement accuracy, but has not systematically explored how MAB assignment impacts inferential statistics, including its impact on false positive rates.

In this paper, we investigate the tradeoff between benefits to students and scientific gain. First, we explore the impact of MAB versus uniform assignment in simulations mirroring typical experiments, using one common regret-minimizing MAB algorithm, Thompson Sampling. Simulations have frequently been used to explore the behavior of MAB algorithms (e.g., Radlinski et al., 2008), including the exploration of possible benefits in experimental contexts (Kuleshov and Precup, 2014). We investigate how likely MAB experiments are to detect effects when they are present and to avoid detecting an effect when it is not present. By exploring different prior distributions in MAB assignment, we demonstrate how optimistic priors can improve power while maintaining student benefits. Because our initial simulations found a decrease in power, we next explore whether MAB assignment is still beneficial to students if larger sample sizes (i.e., more students) are required, and show that for the sample sizes required to attain high power, MAB is more beneficial for students in the experiment than a shorter traditional experiment followed by choosing a single condition for the remaining students. To better understand how MAB experiments might work in practical contexts, we then consider what happens when experiments are embedded in activities like homework where when more- (or less-) able students may tend to engage with activities earlier than others. We show that such biases can have profound impacts on both how often one erroneously detects a difference between conditions when there is no difference and how often one fails to detect an actual difference between conditions, with less-able students engaging first increasing false positive rates. Finally, we model previously collected educational data to illustrate the impact of bandit assignment in real-world experiments, demonstrating improved student benefits and less drastic decreases in power than anticipated. We end by discussing the practical consequences of these results for using MABs within education research.

2. Related work

Balancing benefits to educational researchers and benefits to student participants is a specific example of the more general problem of balancing community benefits and scientific benefits of experiments, which has been addressed in the statistical literature. Especially in medical experiments, randomized trials may be ended early if results indicate a large benefit of one condition over another. Planned interim analyses controlling false positive rates are often part of these experiment designs (e.g., Demets and Lan, 1994; Jennison and Turnbull, 2005; Chow et al., 2007). While ending an experiment early can negatively impact scientific gains, such as by overestimating effect sizes (Bassler et al., 2010), the procedure is commonly used to more ethically conduct studies impacting participants' health. Our work shares this concern with the ethics of experimental design but addresses it by continuously modifying the allocation of participants to conditions and exploring the impact on statistical inference.

Within the clinical trial literature, response-adaptive randomization procedures are techniques that dynamically change the experimental design over time (see Hu and Rosenberger, 2006, for an overview). Examples like doubly adaptive biased coin designs (Eisele and Woodroofe, 1995) seek to allocate participants to conditions so as to meet some target allocation criterion, such as maximizing power or minimizing the total number of harmful patient events. Most analyses of these types of procedures have focused on large samples and have been primarily focused on whether a target allocation is achieved by a particular sequential sampling procedure (e.g., Hu and Rosenberger, 2003, is an example of this type of analysis). A few papers have addressed the question of whether the changing allocation leads to bias in the measurements about the conditions (Atkinson, 2014) and recently, have proposed alternatives to the maximum likelihood estimator as a way to compensate for that bias (Bowden and Trippa, 2017). Kuleshov and Precup (2014) suggested using MABs for clinical trials, demonstrating positive benefits to patients in simulation and showing that the best treatment could still typically be identified. Within the literature on clinical trials, there is limited work on temporal trends in when participants engage with an experiment; Duan and Hu (2009) address cases where changes are made to an experiment that result in differences in treatment effectiveness over time, but they assume that eventually, expected effectiveness for patients at a particular time point converges to a fixed value. This varies from the case we consider, where the overall effectiveness over a population has a fixed value but members of the population are organized in time so that the expected effectiveness is steadily increasing or decreasing.

Outside of the medical literature, there also exist variations of optimal experiment design that change the experimental design over time. For instance, Cavagnaro et al. (2010) aim to design experimental conditions for individuals to best differentiate among scientific hypotheses. While this can indirectly benefit possible participants by reducing sample size requirements, it does not directly address increasing the utility of those who do participate in an experiment.

Within educational experiments, there has been some prior exploration of using MABs for condition assignment (Liu et al., 2014; Williams et al., 2016). Liu et al. (2014) show that MABs under-sample some conditions due to a high probability that those conditions lead to lower rewards (i.e., less learning for students), translating to less gain for researchers in measuring the relative effectiveness of all conditions. They propose a heuristic modification to one popular MAB algorithm, UCB (Upper Confidence Bound), that leads to more sampling of suboptimal conditions; later work introduces an algorithm that explicitly optimizes both measurement accuracy and average reward, giving concrete bounds and a parameter weighting these two objectives (Erraqabi et al., 2017). Our work is interested in some of these same ideas and complements their focus on measurement accuracy by mapping out the consequences of using MABs both for detecting effects that are present and falsely detecting effects in cases with no underlying effect. Our simulations can provide more context for researchers to understand how measurement errors impact statistical inferences, especially in the case where conditions do not actually differ in effectiveness, which is not explored in Erraqabi et al. (2017).

The prior work on measurement errors in MAB assignment for educational experiments focused on UCB, whereas in this paper, we explore a different MAB algorithm, Thompson sampling. Thompson sampling may be more interpretable to researchers due to employing weighted randomization. Williams et al. (2018) proposed using Thompson sampling for MAB experiment design as a way to make experimentation a more accessible tool for teachers to deploy in their classrooms. However, that work focused on meeting the needs of teachers and students, ignoring the statistical impact of MAB assignment. If MAB assignment is to be used within experiments conducted by researchers, it is necessary for researchers to understand how MAB assignment impacts inferential statistics. MAB assignment violates the assumptions of traditional inferential statistics because assignment to conditions for future students is dependent on the assignments and outcomes of past students, and this paper demonstrates the impact of this violation.

MAB algorithms have also been used in educational applications for purposes other than condition assignment in experiments, particularly sequencing educational content. Clement et al. (2015) developed a MAB-based algorithm that selects a problem for a student that is likely to be in her zone of proximal development, providing personalized problem selection with only limited domain knowledge in the form of a pre-requisite graph. A variation on this algorithm that integrates a more complex student model has also been explored in simulation, showing positive impacts when there is large diversity in student learning patterns (Mu et al., 2018). MABs have also been used to improve problem selection in a system that first automatically estimates problem difficulty from offline data (Segal et al., 2018), demonstrating the benefit of online learning as more data become available. Finally, contextual MABs, which make selections based on user-specific information, have been used in a variety of applications to make recommendations to users or personalize their experiences in a system (e.g., Li et al., 2010; Tang et al., 2014). Within education, Xu et al. (2016) used contextual MABs to recommend course sequences to students and showed in simulations using prior data that personalized course sequences could lead to higher student GPAs. While we do not focus on personalization in this paper, we briefly discuss how personalization and experiment design could be brought together.

3. Multi-armed bandits for educational experiments

Multi-armed bandit (MAB) problems have been explored in computer science and statistics as a way of modeling online decision-making scenarios where one must repeatedly choose among a fixed set of options and can only learn about the value of those options by choosing them. We first provide an overview of the formal setup of a MAB problem and of Thompson sampling, the MAB algorithm we use in later simulations; we then describe how the allocation of students to conditions in an educational experiment can be viewed as a multi-armed bandit problem.

3.1. Multi-armed bandit problems and Thompson sampling

In MAB scenarios, agents solve an online decision problem in which they learn to maximize a reward over time based on experience. At each timestep, the agent chooses one of a discrete set of actions A, and receives reward r generated by an unknown, stochastic function; the reward r is typically related only to the chosen action and is independently and identically distributed (IID) given the action. The agent's *policy* for choosing actions must balance exploiting information already collected with exploring actions to collect additional information. Most commonly, the policy aims to minimize expected regret. Expected regret is the difference between using the agent's policy and always choosing the best action. Formally, if the action with highest expected reward is a^*, the agent's choice of action at time t is a_t, and the reward from action a at time t is a random variable denoted $X_{a,t}$, then the expected regret through time T is $E[\sum_{t=0}^{T}(X_{a^*,t} - X_{a_t,t})]$, with the expectation averaging over both the stochastic reward function and the impact of the reward on action choices. An introduction to the multi-armed bandit problem and regret minimization can be found in Lai and Robbins (1985).

Thompson sampling is one algorithm for choosing actions in a MAB problem. The bound on regret for Thompson sampling grows logarithmically with the number of samples (Agrawal and Goyal, 2012); this bound is close to the optimal bound on regret growth over time (Lai and Robbins, 1985). Additionally, Thompson sampling has been shown to perform well in practice (Chapelle and Li, 2011). Thompson sampling corresponds to weighted randomization based on the expected value of the reward. It maintains a distribution over rewards for each action, based on the reward observed each time the action has been chosen so far. Let these distributions be $D_1, \ldots, D_n$, where n is the number of actions. At each timestep, it chooses an action by sampling a value r_i from each distribution D_i and selecting the action with highest sampled value: $\text{argmax}_i\, r_i$. After choosing action i and observing the reward, distribution D_i is updated to reflect the new information.

For example, consider the case where the rewards are binary and there are two actions. The Thompson sampling algorithm could be applied to the MAB problem by modeling the outcome from each action choice as a Bernoulli random variable and placing a (conjugate) Beta prior on each action. If equal priors of Beta(s, f) are placed on each action, then to choose the first action, the algorithm draws $s_1^{(1)} \sim$ Beta(s, f) and $s_2^{(1)} \sim$ Beta(s, f). Action 1 is chosen if $s_1^{(1)} \geq s_2^{(1)}$, and action 2 is chosen otherwise. Assume action 1 is chosen, and that the action was a success (reward of 1 rather than 0). Then to choose the second action, the algorithm draws $s_1^{(2)} \sim$ Beta$(s+1, f)$ and $s_2^{(2)} \sim$ Beta(s, f): the distribution for action 1 was updated because we observed the result of choosing action 1, but the distribution for action 2 was not updated. The algorithm again chooses the action that had the larger sample ($\text{argmax}_i\, s_i^{(2)}$), observes the reward for the chosen action, and updates the corresponding distribution. Over time, this process will result in choosing the action with higher reward more frequently than the action with lower average reward.

3.2. Modeling educational experiments as MAB problems

Educational experimentation can be viewed as a MAB problem by treating condition assignments for students as action choices, with the dependent outcome serving as the reward. For example, in an experiment comparing video versus text hints, there would be two possible actions: assign a student to the video hint condition and assign a student to the text hint condition. The reward (outcome) could be defined to be 1 if the attempt after the hint was correct and 0 otherwise. The MAB algorithm sequentially assigns students to conditions, based on the rewards (outcomes) for previous students; more students can thus be assigned to better conditions. While some MAB algorithms optimize objectives other than regret (e.g., Kaufmann et al., 2016), we focus on minimizing regret because this corresponds to maximizing student outcomes. We prioritize maximizing the benefits to students and then explore the consequences of this choice on the information gained from the analysis of results.

Assigning conditions using an algorithm like Thompson sampling that minimizes regret means that later students will tend to be more frequently assigned to conditions that have higher expected outcomes. For instance, in the video versus text hint experiment, assuming there is no initial information for the algorithm to know whether video or text hints are better, the algorithm will choose an action uniformly at random to assign the first student to a condition. If that student is correct on her next attempt, then the algorithm will have gained some information about the effectiveness of her condition. After several students have been assigned to conditions, the algorithm will have more information that it can use to determine which condition is more effective, and it will tend to assign more students to that condition, exploiting the information it has learned. Thus, if there is an underlying difference in the outcome measure across conditions, the algorithm will tend to produce an unequal assignment of students to conditions. At all timesteps, Thompson sampling will have a non-zero probability of choosing each condition (action), so as to maintain some exploration, but in general, this probability will not be equal across conditions.

Within the simulations that follow, we apply Thompson sampling to model educational experiments with binary outcomes (e.g., whether a student completes an activity) and real-valued outcomes (e.g., time to finish a problem). While non-exhaustive, these categories cover most experiments, especially if cases like discrete scores on a post-test are treated as real-valued outcomes. Note that the outcome from a student maps to the reward that Thompson sampling seeks to maximize; throughout the remainder of this paper, we use "outcome" and "reward" interchangeably.

For binary outcomes (rewards), we use the Beta-Bernoulli model described above: a Beta distribution is maintained for each condition, and after a student is assigned to a condition, the Beta distribution for the chosen condition is updated based on the outcome for that student. Assuming Beta(s, f) prior, after observing n successes and r failures for a single condition, the distribution for that condition will be Beta$(s + n, f + r)$.

For real-valued outcomes, we assume the outcomes (rewards) are normally distributed. Thompson sampling again maintains a distribution for each condition, where here, the likelihood for the outcome of a single trial is assumed to be $\mathcal{N}(\mu, \sigma)$, where μ and σ are unknown. To make implementation efficient, a Normal-Gamma prior is used for each condition, which is conjugate to the normal distribution parameterized via the mean μ and precision $\lambda = \sigma^{-1}$. Specifically, given prior NormalGamma$(\mu, \kappa, \alpha, \beta)$, after observing n rewards $x_1, \ldots, x_n$ with mean $\bar{x}$, the posterior is Normal-Gamma with parameters $\frac{\kappa\cdot\mu+n\cdot\bar{x}}{\kappa+n}$, $\kappa+n$, $\alpha+\frac{n}{2}$, and $\beta+\frac{1}{2}\sum_{i=1}^{n}(x_i-\bar{x})^2+\frac{\kappa n(\bar{x}-\mu)^2}{2(\kappa+n)}$.

To summarize, educational experiment design can be modeled as a MAB problem by viewing each condition as an action, where choosing condition a at timestep t means assigning the tth student to condition a. The outcome measure in the experiment is then the reward for the MAB algorithm. This model is consistent with previous work using MABs for educational experimentation (Erraqabi et al., 2017; Williams et al., 2018).

4. Statistical consequences of MAB-assigned conditions

While using MAB assignment for experimentation could assign more students to better conditions, this benefit comes from conditions with unequal sample sizes and from making condition assignment dependent on previous students' results. We now turn to investigating how this impacts drawing statistical conclusions from the data via simulations comparing condition assignment using Thompson sampling versus typical uniform random assignment.

In the simulations, we consider several experimental outcome measures and effect sizes and examined the impact on both the expected benefits to students and on the results when analyzing the data, such as the power when conditions were in fact different and the false positive rate when the conditions were equally effective. We also examined how likely the direction of effects is to be erroneously reversed in the results ("Type S errors," Gelman and Carlin, 2014), which are perhaps even more serious than falsely detecting an effect that is not present. The results of these simulations thus address both how much more beneficial MAB assignment is for students in the experiment and the cost of this benefit in terms of analyzing the data from the experiment.

4.1. Simulation methods

Each simulation corresponds to an instantiation of an experiment with simulated participants. Across simulations, we varied method of condition assignment (MAB versus uniformly at random), reward type, true effect size, and number of participants (sample size). Table 1 summarizes these factors, and more details are given below. Each set of parameters was used for 500

Table 1: Factors varied in the simulations.

Condition assignment method	Uniformly at random MAB (Thompson sampling)
Reward (outcome) types	Binary Normally distributed (real-valued)
Sample sizes	Non-zero effect size: $n \in \{0.5m, m, 2m, 4m\}$, where m is the number of participants required to achieved expected power of 0.8 with uniform random assignment to condition Zero effect size: Same sample sizes as for non-zero effect size, using the values of m calculated for the non-zero effect size simulation with the same parameters.
Effect sizes	Small, moderate, and large effect sizes for each type of reward. Binary rewards: Cohen's $w \in \{.1, .2, .3\}$ Normally distributed rewards: Cohen's $d \in \{.2, .5, .8\}$
Prior distributions	(Only for MAB assignment. All conditions in one simulation have identical priors, and all prior distributions are equally strong.) Prior above: Mean of the prior distributions is above both condition means. Prior between: Mean of the prior distributions is between the two condition means when condition means differ and is equal to both condition means when the condition means do not differ. Prior below: Mean of the prior distributions is below both condition means.

simulations. Each run of a simulation involved generating an initial simulated participant based on the condition parameters, assigning that participant to a condition using either a MAB policy or uniform random policy, observing the result, and then repeating the same procedure. For the uniform random policy, the results of each previous simulated participant did not affect the condition assignment of the next simulated participant, while for the MAB policy, the previous results were used by Thompson sampling to influence the probability of each condition for the next participant.

Reward Models and Types: As described in Section 3.2., we considered both binary rewards (e.g., whether a student turns in her homework) and real-valued rewards (e.g., student's rating of her proficiency in a particular skill). For the simulations, a Bernoulli distribution was used to generate the binary rewards, and a normal distribution generated the real-valued rewards. For the Thompson sampling implementation, conjugate priors were used for Bernoulli and normal likelihoods; see Section 3.2. for details.

Effect sizes: Three non-zero effect sizes were used for each reward type, corresponding to common thresholds for small, moderate, and large effects: binary reward thresholds were 0.1, 0.3, and 0.5 for Cohen's w (Cohen, 1988), and normally distributed reward thresholds were 0.2, 0.5, and 0.8 for Cohen's d (Cohen, 1988). For binary rewards, the average probability of success across conditions was fixed to 0.5, resulting in conditions with 45% and 55% success rates for $w = 0.1$, 35% and 65% success for $w = 0.3$, and 25% and 75% success for $w = 0.5$. For normally distributed rewards, the means of the conditions were set to -0.5 and 0.5, and then the desired effect size was used to compute the variance for each condition; within a simulation, the two conditions always had the same variance.

An effect size of zero was also used, corresponding to the case where the conditions did not in fact differ from one another. For zero effect size, binary reward simulations had condition means equal to 0.5, and the normally distributed reward simulations had condition means set to zero, with the same variances as for the non-zero effect sizes.

Sample sizes: For each effect size, we computed m, the sample size that achieves 0.8 power with equally sized conditions. That is, given that there is a difference between conditions, there is an 80% chance that the statistical hypothesis test will detect this effect, assuming equally balanced conditions. These are standard thresholds for experimental design in psychology. Simulations with $0.5m$ (lowest power), m, $2m$, and $4m$ (highest power) simulated students (steps) were conducted. When effect size was zero, the same sample sizes were included as for all non-zero effect size simulations.

Prior distributions: In the MAB simulations, the same prior was placed on both conditions, corresponding to no prior preference for one condition over another. While the priors were weak, they still could influence results, and thus three variations were compared: *Prior between* placed the prior mean between the means of the two conditions. For binary rewards, this was a Beta$(1, 1)$ prior. For normally distributed rewards, the prior was NG$(0, 1, 1, 1)$ (i.e., prior on the mean has mean zero). *Prior above* placed the prior mean above both conditions (binary rewards: Beta$(1.5, 0.5)$ prior; normally distributed rewards: NG$(1, 1, 1, 1)$ prior). *Prior below* placed the prior mean below both conditions (binary rewards: Beta$(0.5, 1.5)$ prior; normally distributed rewards: NG$(-1, 1, 1, 1)$ prior). When effect size was zero, the same priors were used. Intuitively, *prior between* is not systematically biased compared to the conditions, while *prior above* is overly optimistic about the conditions and *prior below* is pessimistic. The strength of the priors was held constant across the variations in order to isolate the impact of the relation between the expected effectiveness of the conditions as indicated by the prior and their actual effectiveness.

4.1.1. Data analysis

Data analyses examined (a) rewards per student, (b) statistical power to detect effects, (c) the rate of detecting an effect in the incorrect direction (Type S error rate), and (d) the rate of incorrectly rejecting the null hypothesis (Type I error rate).

To determine the impact of the different factors on these outcomes, we analyzed each simulation individually as if it were an actual experiment and then aggregated results across simulations. First, for each simulation, we tested whether there was a significant difference between conditions based on the collected data. Since the reward for each student corresponds to the outcome measure for that student, we compared conditions based on rewards. For normally distributed rewards, we compared conditions using a t-test[1], and for binary rewards, we performed a χ-squared contingency test. In both cases, significance is set at $\alpha = 0.05$, following typical norms in education and psychology experiments.

For each set of parameter values, 500 repetitions of that simulation were performed. When the two conditions differ in effectiveness, the statistical power for a particular simulation is the proportion of those repetitions where a significant difference between conditions was found.

[1] Because of the unequal sample sizes, some researchers might use Welch's t-test (Welch, 1938) rather than a traditional Student t-test, although we note that the true variances across conditions are identical. Results are very similar if Welch's test is used to test for significance.

When the conditions do not actually differ in effectiveness, the false positive rate (Type I error rate) is again the proportion where a significant difference between conditions was found. Type S errors can only occur when the conditions differ, and the Type S error rate is the proportion of simulations where a significant difference was found and the condition that is in actuality more effective is measured as significantly worse than the other condition.

To determine which parameters impacted the outcome of the simulations, we used linear and logistic regressions. Linear regression was used to predict the average reward per student (i.e., average outcome measure for a student) based on the simulation set up, while logistic regression was used to predict whether a significant difference would be found between conditions, since this measure is binary. The regressions included predictors for assignment type, reward type, effect size, and sample size, with type of prior included in the analyses that were specifically examining the impact of prior distributions. Each run of a simulation is included as a data point in the regression, meaning the 500 runs for each set of parameter values provide multiple samples to aid in determining which factors reliably impact simulation outcomes.

In order to show that the statistical consequences of MAB assignment are not due simply to having adopted a particular p-value or more generally adopting frequentist analyses, we also considered a more Bayesian approach in which we fit a distribution over the mean of each condition for each simulation with binary rewards and over the mean and variance for each simulation with normally distributed rewards; this analysis treats the values of the means (and variances) as random variables. We adopt a Beta$(1, 1)$ prior for the binary rewards, and a Normal-Gamma$(0, 1, 1, 1)$, indicating no a priori beliefs about which condition is better. Posteriors are then computed by using Bayes' rule to combine the dataset collected by the simulation and these priors. For each simulation, we computed the difference in expected value of the posterior condition means, and additionally, approximated the probability that the true mean of one condition is larger than the other by drawing 1000 samples from each condition.

4.2. Results

4.2.1. Results when conditions differ in effectiveness

When conditions have different benefits for students, the best outcome from the perspective of the researcher is to detect that the difference is reliable, and the best outcome from the perspective of the students is to assign more students to the better condition. We explore whether the first outcome is achieved by examining experimental power and the second by examining average rewards, which correspond to student outcomes. MAB assignment without an optimistic or pessimistic prior (*prior between*) decreased power from an expected 0.80 to 0.57 for binary rewards (Figure 1a) and 0.49 for normally-distributed rewards (Figure 1c). Doubling the sample size raised power closer to the desired 0.80 (0.79 and 0.69 respectively). As Figures 1b/d show, there are diminishing returns of more students: evidence for the superiority of one condition leads to assigning fewer and fewer students to the alternative, which is the condition that needs to be measured to improve power.

This unequal allocation of participants to conditions drives the decrease in power. As shown in Figure 2a-b, MAB assignment results in many simulations with the vast majority of participants in one condition, while uniform assignment (Figure 2c) has a much more even distribution. For instance, in 93% of simulated experiments with normally-distributed rewards and large effect size, at least two-thirds of participants were assigned to the better condition when MAB assignment was used (Figure 2b), while this happened only 1% of the time with uniform ran-

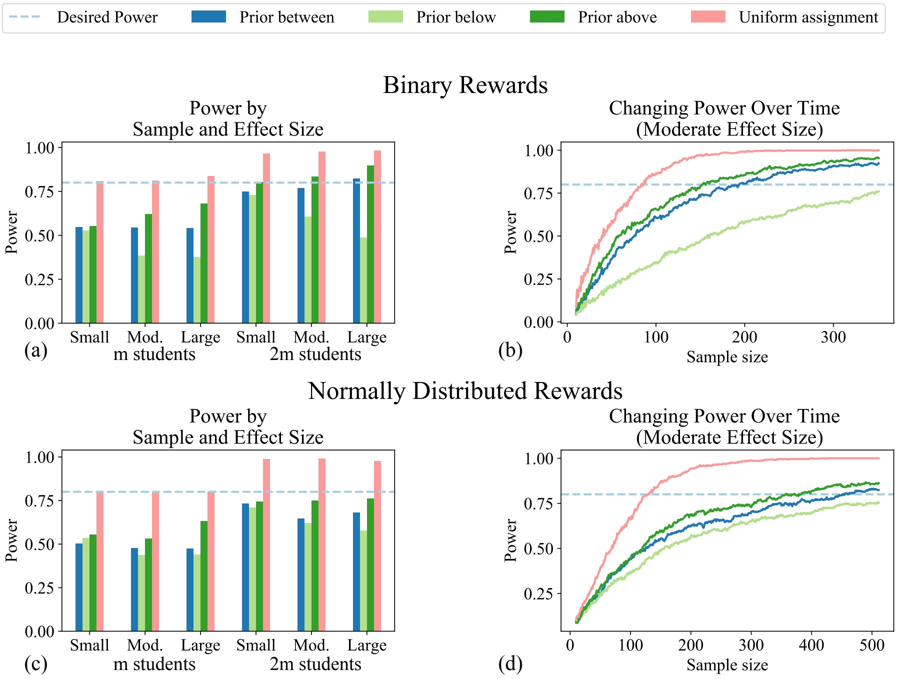

Figure 1: Power by assignment and outcome measure. Decrease in power is similar across effect sizes. (a/c): Power with m participants (expected power $= 0.8$) and $2m$ participants, by reward type. (b/d) Power over time by reward type, with a total of $4m$ participants.

dom assignment.

As described in Section 3.1.1, we used logistic regression to determine what factors impacted power (i.e., whether a simulation found a significant difference between conditions). This analysis confirms that MAB assignment led to lower power than uniform assignment, $t(23995) = 39.5$, $p < .0001$. Normally distributed rewards also lowered power, driven by the MAB simulations, $t(23995) = 8.59$, $p < .0001$. These rewards have a larger range than binary and thus a single student outcome can cause a large change in the expected value of a condition, causing condition probabilities to shift sharply; this can result in almost all probability on a single condition after a small number of participants. The lack of participants in one condition leads to a less precise estimate of that condition, and the statistical test does not find a reliable difference.

While decreased power will be of concern to researchers, Type S errors, in which a significant finding is in the opposite direction of the true effect, are potentially much more damaging. Overall, Type S errors were rare ($< 0.15\%$), and no difference by assignment type was detected: 0.13% of MAB simulations had a Type S error, and 0.00% of uniform simulations had a Type S error.

MAB assignment overestimated effect sizes for normally distributed rewards, with greater overestimation for smaller effect sizes: for small actual effect size of 0.2, *prior-between* simulations found an average effect size of 0.30 with sample size m, while simulations with the large effect size of 0.8 found average effect size 0.93. Overestimation of effect sizes is known to occur in low-powered studies (see, e.g., Button et al. 2013), but here, the overestimation is not due to filtering out nonsignificant simulations. Instead, it is a consequence of the measurement errors

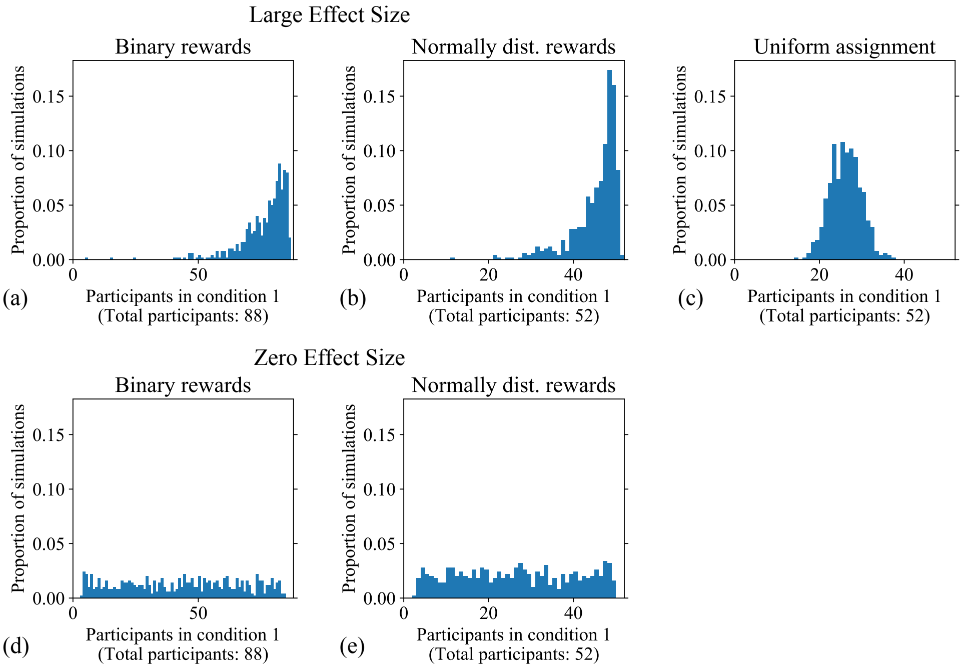

Figure 2: Histograms of the number of participants assigned to condition 1 with *prior between* for MAB assignment. Number of participants is that needed to attain 0.8 power with equally sized conditions. (a-b) When effect sizes are non-zero, most MAB simulations assign the vast majority of participants to the better condition. (c) Uniform sampling has the same assignment pattern in all cases (binary rewards case shown here), tending to assign about half of participants to condition 1, and extremely unbalanced simulations almost never occur. (d-e) When effect sizes are zero, the distribution of sample sizes for condition 1 is roughly uniform for MAB sampling.

in the condition means that have been shown to occur with MAB sampling (see Section 2).

This overestimation of effect sizes for normally distributed rewards suggests that the statistical consequences of MAB assignment are due to not only to unbalanced sample sizes but to the way in which the adaptive assignment affects which data are collected. This is illustrated by the results of the Bayesian analysis in which we examined the posterior distributions over conditions based on the data collected in one simulation, where results show the same trends as in the frequentist analyses. As shown in Figure 3a, the distribution of the probability that the mean of condition 1 is greater than the mean of condition 2 is more concentrated at higher values for uniform assignment than for MAB assignment: that is, simulations that used uniform assignment tended to have higher probabilities that the better condition was (on average) more effective than the worse condition. Uniform assignment also produced more accurate differences in expected value of the means (Figure 3b). The longer right tail of this distribution for MAB assignment demonstrates a tendency towards overestimation.

MAB assignment does obtain greater rewards than uniform: the expected reward for a single student is close to the success rate of the more effective condition for binary rewards and approaches the mean of the better condition for normally distributed rewards a bit more slowly (Figure 4). Thus, while decreased power means more participants (students) will be necessary to detect an effect, a large proportion of students will be assigned to the better condition while the experiment is running.

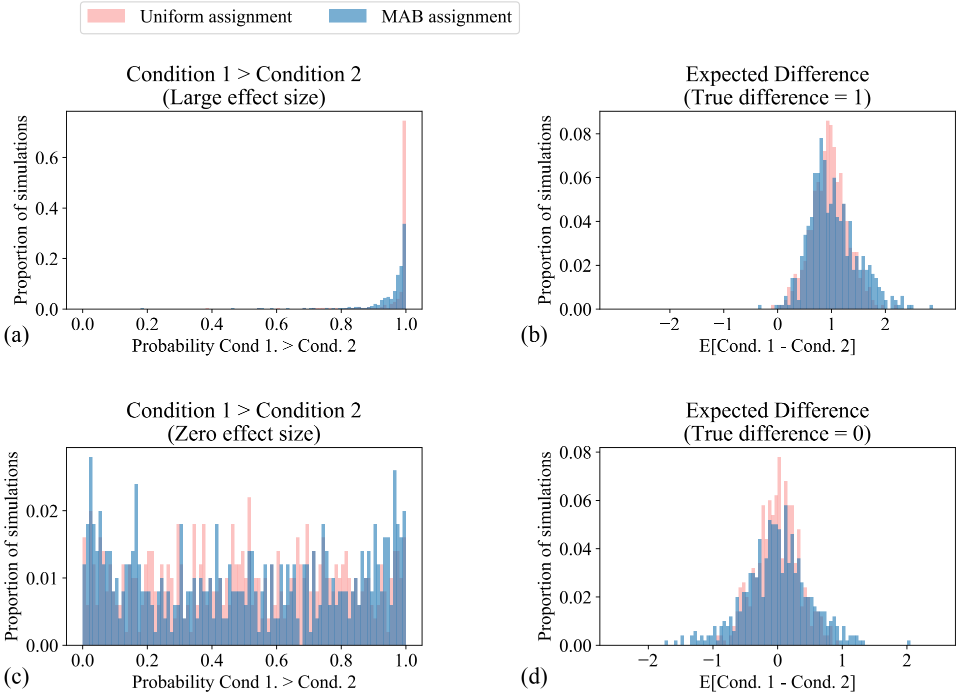

Figure 3: Histograms of condition differences across simulations, assuming a distribution over condition means. Rewards are normally distributed; binary rewards result in similar trends. (a) Distribution of the probability that condition 1 is more effective than condition 2 (large effect size; small and moderate effect sizes show similar trends). (b) Difference in expected value between posterior condition means (large effect size). (c) Distribution of the probability that condition 1 is more effective than condition 2 given no underlying difference between the two conditions. (d) Difference in expected value between the condition means (zero effect size), showing longer tails for MAB assignment.

The choice of prior in the MAB simulations impacted power (coefficient *prior below* $= -0.72$, $t(35994) = 24.7$, $p < .0001$; coefficient *prior between* $= -0.30$, $t(35994) = 10.3$, $p < .0001$). An optimistic prior (*prior above*) led to higher power and more accurate effect sizes than the other two priors: *prior above* found a significant effect in 68% of simulations, compared to 62% for *prior between*, and 53% for *prior below*. This is partially driven by very unbalanced simulations: 8% of *prior below* and 2% of *prior between* simulations assigned at least 99% of participants to a single condition, compared to 1% of *prior above* and none of the uniform assignment simulations. With the optimistic prior, the first few samples tend to decrease the algorithm's expectations, since the samples are likely below the prior mean. This leads to more equal sampling across conditions initially even though all priors have the same strength (i.e., correspond to the same number of fictional samples), and this more equal sampling leads provides better evidence to estimate the means.

Differences in the prior have a similar impact on Type S error rates. Type S errors increase slightly for the less optimistic *prior between* (marginally significant: $t(35994) = 1.91, p = .056$) and *prior below* ($t(35994) = 2.81$, $p = .0049$). However, Type S errors are still extremely rare (0.21% for the prior with the highest rate).

Despite the impact of the choice of prior on power and Type S error rates, average reward is

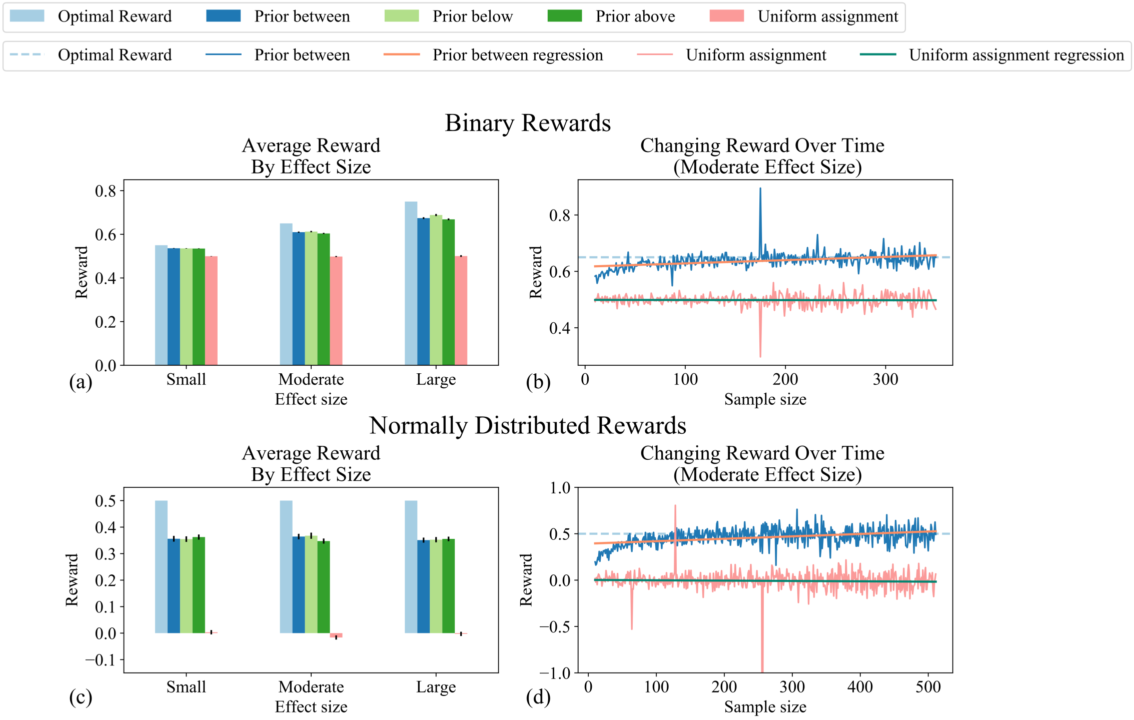

Figure 4: Benefits to students based on sampling type, compared to assigning all students to the best condition (light blue bar): MAB assignment obtains higher rewards, with increasing reward over time. Differences in average reward per step with small, moderate, and large effect sizes are shown for binary rewards in (a) and for normally distributed rewards in (b). For the binary rewards in (a), the average reward of the best (and worst) arm varied across the effect sizes and the average reward was always .5; for the normally distributed rewards in (b), the arms always had expected rewards of 0.5 and -0.5, leading uniform assignment to have an average reward per step very close to zero. Error bars represent one standard error. (b) and (d) show trend of reward growth as the experiment runs (i.e., as the sample size in the experiment increases).

only modestly decreased with more optimistic priors (Figure 4). Using a more optimistic prior slightly delays when the algorithm will primarily exploit what it has learned from the collected data, but it still concentrates most of its action choices on the more effective condition.

:

4.2.2. Results when conditions are equally effective

When there is no underlying difference between conditions, there is no biased sampling pattern that can improve rewards, and we confirmed via linear regression that assignment type did not impact rewards. Instead, the primary concern is to avoid falsely rejecting the null hypothesis (i.e., a type I error). We aggregated across prior types for MAB assignment because these did not have a statistically significant impact.[2] As shown in Figure 5, MAB assignment increased the Type I error rate: 5.0% of simulations using uniform assignment found a significant difference between conditions, while 9.4% of simulations using MAB assignment found a difference. Note that in both cases, we use $\alpha = 0.05$ in a standard statistical test, demonstrating that a

[2] Results are very similar if only one prior type, rather than all three, are included for MAB assignment.

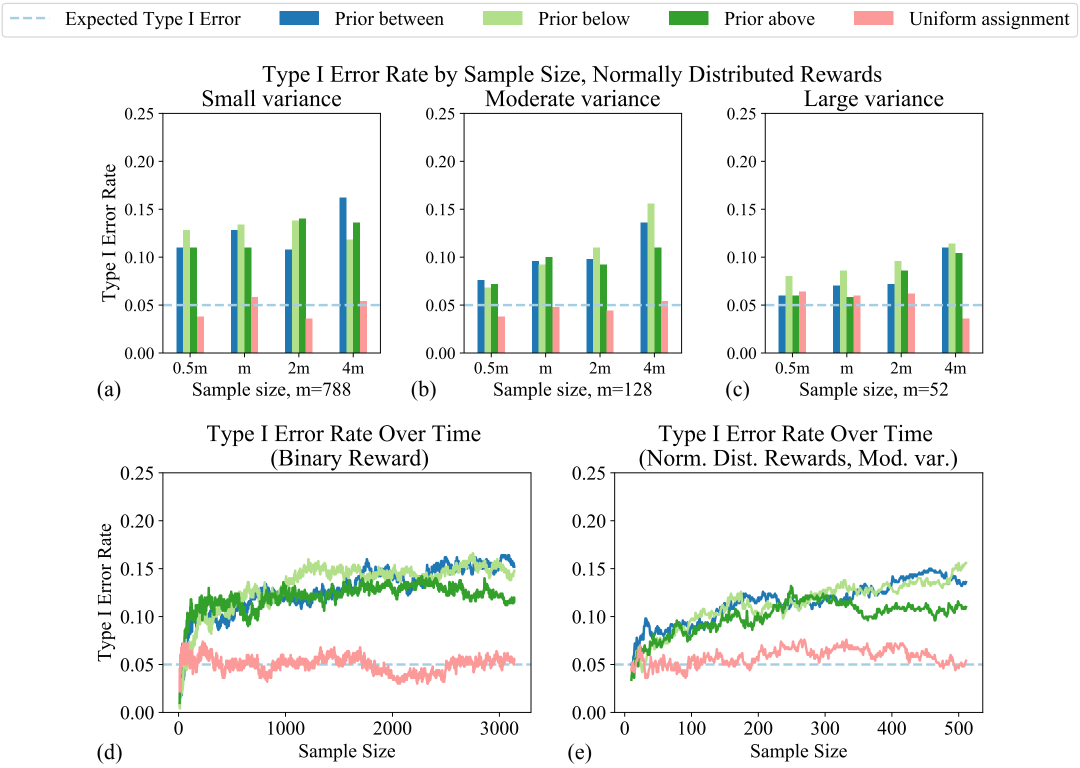

Figure 5: MAB assignment increases Type I error rate over uniform assignment. (a-c) Type I error rates for normally distributed rewards by variance of conditions. (d-e) Type I error rates by sample size for binary (d) and normally distributed rewards (e).

researcher who uses standard statistical tests and parameters to analyze results collected via a MAB-assigned experiment design will underestimate their false positive rate. Quantifying the magnitude of this increase is important for adjusting α for MAB experiments: lower α decreases Type I errors (and power).

Logistic regression found that all predictors were statistically reliable, but assignment type had the largest impact, $t(23995) = 14.8$, $p < .0001$. There was a slightly higher Type I error rate for normally distributed rewards than for binary rewards, $t(23995) = 5.20$, $p < .0001$, primarily due to insufficient exploration with small variances. While sample size has a reliable impact, the practical impact is likely limited: thousands of additional students are needed to meaningfully increase false positives.

Modeling the means of the conditions as random variables also demonstrates differences in the posterior distributions between uniform and MAB assignment. As shown in Figure 3c, when there is no difference between conditions, the probability that the mean of condition 1 is greater than the mean of condition 2 skews closer to the extremes (0 and 1) for MAB assignment than for uniform assignment. These extremes indicate evidence for a difference between conditions, consistent with the increased Type I error rate. Additionally, the expected value of the difference between the means tends to be more extreme for MAB assignment than for uniform assignment, as shown by the longer tails in Figure 3d. This shows MAB assignment collects data that is more likely than data collected using uniform assignment to increase beliefs in a difference between conditions even when there is no underlying difference.

5. Choosing a policy for future students with MAB versus a traditional experiment

MAB experiments reduce power because they put more students in one condition than another. This reduction in power means researchers must plan for larger sample sizes to detect effects. If uniform assignment was used for the experiment, a smaller sample size could be used, and then the information from that experiment used to decide what condition should be used for all future students. Across a longer timescale with larger numbers of students, will MAB assignment place more students in the less effective condition than choosing one condition for future students based on a small initial experiment? In this section, we compare these two scenarios, running simulations with very large sample sizes. The MAB simulations use MAB assignment throughout, while the uniform assignment simulations switch to assigning all students to one condition after a pre-determined sample size. We consider both always switching to the condition that has been more effective so far and switching to the better condition only if it is significantly better than the alternative. The latter mirrors the commonly encountered situation where a technology will be revised only if a researcher or curriculum designer has reliable evidence that the proposed new version is more effective. These simulations address whether the large sample sizes needed for sufficient power offset the benefits to students of MAB assignment that we saw in the previous section. Because these simulations address whether the improvements in outcomes for students could be compelling enough to warrant larger experiments, we consider only cases where the conditions in the experiment actually differ and thus there is the possibility of improved outcomes (rewards) for students.

5.1. Methods

Simulation lengths are again based off of m, the expected number of participants given the true effect size to achieve 0.8 power using uniform random assignment. We include results from simulations of length $2m$, $6m$, and $11m$, where the first m participants are the experimental phase for the two uniform random assignment approaches (described below). All simulations use a moderate effect size.

Methods for the MAB simulations match those in the previous section, with only non-zero effect sizes included and *prior between* used for all simulations.

We consider two variations of choosing a condition based on an initial period of uniform random assignment. In both, uniform random assignment is used for the first m students. For the *revise always* simulations, the means of the two conditions are compared after the first m students, and remaining students are assigned to whichever condition has larger mean. This *revise always* approach is somewhat similar to Thompson sampling, as both incorporate evidence of effectiveness even if that evidence is not sufficient to demonstrate a reliable difference. In the MAB literature, the *revise always* approach corresponds to an ϵ-first algorithm, where all exploration occurs in some fixed number of initial trials; as noted by Scott (2010), this approach tends to be much less efficient than other MAB algorithms. However, in fixed horizon problems, the amount of exploration can be chosen intelligently (as discussed in, e.g., Langford and Zhang 2008), often leading to good performance. Our approach differs from that in Langford and Zhang (2008) because we choose the amount of exploration based on typical experiment design patterns, rather than trying to directly maximize reward. While *revise always* is a simple MAB algorithm, we discuss it separately from the Thompson Sampling MAB approach due to the fact that in education, *revise always* is more likely to arise as a variation on typical uniform assignment; in subsequent sections, MAB assignment thus continues to refer to MAB assignment via Thompson sampling.

For the *revise if different* simulations, a t-test (normally distributed rewards) or a χ^2-test (binary rewards) is conducted after the first m participants. If the p-value of the test is less than 0.05, the remaining simulated students are assigned to the condition with higher measured mean. If the p-value of the test is greater than or equal to 0.05, one of the two conditions is chosen uniformly at random, and the remaining students are assigned to that condition. This random choice for which condition to give to the remaining students is intended to represent a return to whatever the original version of the technology was, and we assume there is no *a priori* bias in which condition is the experimental condition and which the control.

5.2. Results

Across all sample sizes, the simulations using MAB assignment had higher average reward per student than either of the two variations that began with uniform random assignment (Figure 6a/b). For the initial experimentation period consisting of the number of students to achieve 0.8 power with uniform random sampling, both *revise if different* and *revise always* sample uniformly at random and thus had average rewards very close to the average of the expected values of both conditions (0 for normally distributed rewards and 0.5 for binary rewards), while (as demonstrated in the previous section), MAB assignment preferentially assigns students to the more effective condition and thus attained average reward of 0.37 for normally distributed rewards and 0.61 for binary rewards. At this stage, the comparison is equivalent to comparing MAB assignment and uniform random assignment, and as documented in the previous section, this difference in rewards is reliable.

After $2m$ simulated students, all three assignment methods have higher average rewards, again as shown in Figure 6a/b: MAB assignment has highest average rewards per student (0.42 for normally distributed rewards and 0.63 for binary rewards), followed by the *revise always* assignment method (0.26 for normally distributed rewards and 0.57 for binary rewards), and then the *revise if different* assignment method (0.20 for normally distributed rewards and 0.56 for binary rewards). These differences were reliable for normally distributed rewards, as shown by linear regression to predict average reward per student based on assignment method (coefficient for *revise always* $= -0.16$, $t(1497) = 15.6$, $p < .001$; coefficient for *revise if different* $= -0.22$, $t(1497) = 21.5$, $p < .001$). However, for the binary rewards case, given the small magnitude of the differences across average reward per step, the fit of a logistic regression was not significantly better than a constant model ($\chi^2 = 4.6$, $p = .1$, *n.s.*; coefficient for *revise always* $= -0.21$, $t(1497) = 1.65$, $p = .099$; coefficient for *revise if different* $= -0.26$, $t(1497) = 2.01$, $p = .044$).

These higher average rewards per student are a reflection of the fact that on average, MAB assignment directs fewer of the $2m$ students to the worse condition compared to the other assignment methods (Figure 6c/d): for normally distributed rewards, an average of 19 out of 256 students were assigned to the worse condition by MAB assignment (7.5% of all simulated students), while *revise always* assigned an average of 64 students (25%) to this condition and *revise if different* assigned an average of 77 students (30%) to this condition. For binary rewards, MAB assignment assigned an average of 15 out of 176 students (8.4%) to the worse condition, compared to 45 students (25%) for *revise always* and 52 students (30%) for *revise if different*. Thus, across both types of rewards, MAB assignments lower the total number of students who experi-

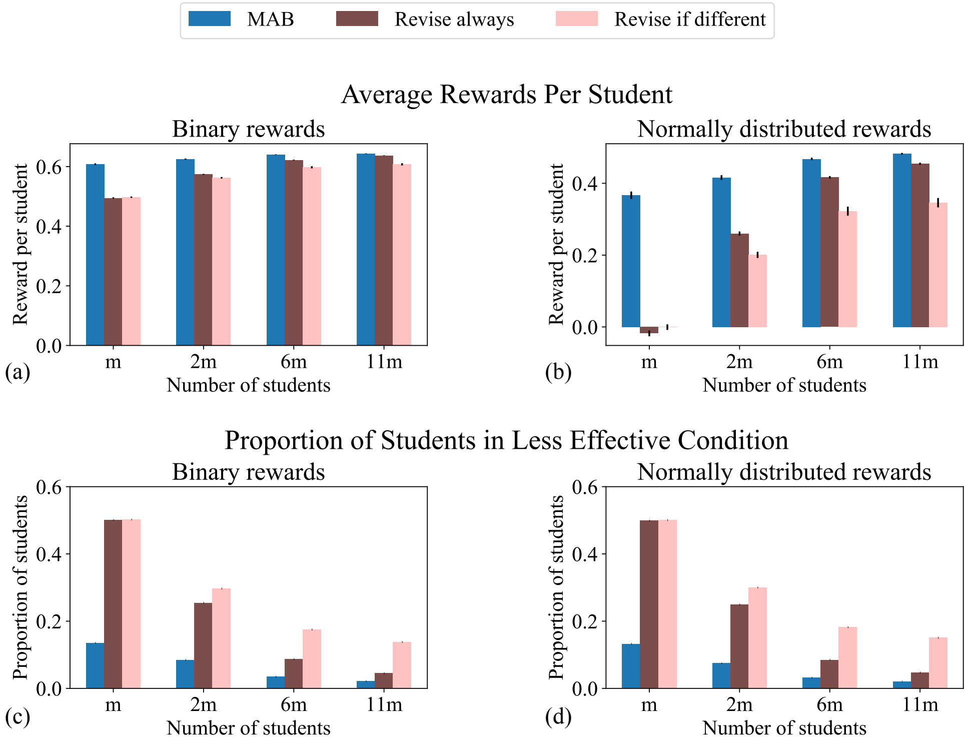

Figure 6: Differences in reward and condition assignment from MAB assignment versus uniform assignment followed by committing to a single condition. (a-b) Average reward per step is higher for MAB assignment, and remains reliably higher for normally distributed rewards even as the number of students increases. Error bars reflect one standard error. (c-d) Fewer students overall are assigned to the less effective condition by MAB assignment.

ence the less effective condition. After $2m$ simulated students, the power for MAB assignment is much higher for both normally distributed and binary rewards, as documented in Section 4.2.

The same trends persist as the number of simulated students increases: MAB assignment consistently has the highest average rewards per student. This effect is reliable at both $6m$ and $11m$ simulated students in the normally distributed rewards case, although not for binary rewards. This difference in average rewards is due to the fact that MAB is more effective than the other assignment methods at reducing the number of students assigned to the less effective condition: MAB assigns 3.3% of students to this condition given a sample size of $6m$ and 2.1% of students given a sample size of $11m$, compared to 8.6% of $6m$ students and 4.6% of $11m$ students for *revise always*, and 18% of $6m$ students and 14% of $11m$ students for *revise if different*. In addition to showing that MAB assignment via Thompson sampling is better than the explored alternatives, these results show that the most typical pattern in educational experiments (*revise if different*) is in fact the most poorly performing.

With these larger sample sizes, MAB assignment exceeds 0.8 power: at $6m$ students, measured power was 0.89 for MAB assignment with normally distributed rewards and 0.97 with binary rewards, and at $11m$ students measured power increased to 0.96 and 0.98 respectively.

These results demonstrate that MAB assignment can improve student experiences even though longer experiments are required to achieve sufficient power, even compared to running an appropriately powered experiment and incorporating its results into the curricula to benefit all future students. While a researcher will need to choose some number of students to include in the experiment so as to be able to analyze their data, the results at both $6m$ and $11m$ students demonstrate that even much large sample sizes are not hurting students: fewer overall students are assigned to the less effective condition by MAB assignment than by the experimental designs that commit to one condition based on the data from a smaller experiment. This occurs because both of the designs that begin with uniform assignment assign more students to the less effective condition during the initial experimentation period than MAB assignment does over even 11 times as many students.

These results also show the benefits to students of treating experimental condition assignment as an online decision-making problem where the assignment for a student is based on all previous students. The alternatives to MAB assignment that we explored can be viewed as batch versions of this decision problem: rather than always assigning the tth student to a condition based on students $1, \ldots, t-1$, each student $1, \ldots, m$ is assigned without any reference to previous students, and the remaining students are assigned based only on the first m students. This limits the adaptiveness of condition assignment: MAB assignment can quickly change the relative probability of conditions based on evidence that one is better than another, and can also reverse that change if new evidence suggests a condition is not as good as it initially appeared. The alternatives that we considered here can only change their proportions once (after m students), and can erroneously commit to the condition that is less effective.

Improved power and fewer students experiencing the less effective condition mean that MAB assignment may be an effective design for researchers who are primarily concerned with detecting an effect when one is present. However, we note that which of the three assignment methods is used will also influence the measured mean of the less effective condition. Even when using MAB assignment for $11m$ participants, the estimate of the mean for the worse condition is still lower than its true value: MAB assignment with $11m$ participants results in a lower average mean for the worse condition than uniform random assignment with m participants (linear regression to predict mean of worse condition, using reward type and sampling type as predictors; coefficient for MAB assignment $= -0.15$, $t(2997) = 15.3$, $p < .001$). This translates to an average value of -0.74 rather than -0.5 for normally distributed rewards, and an average value of 0.28 rather than 0.35 for binary rewards. In contrast, the average mean of the better condition based on data collected by MAB assignment is very close to the actual mean. This phenomenon matches what was observed in the previous simulations and illustrates the measurement error found in previous work (described in Section 2).

Thus, longer experiments can mitigate decreased power from MAB assignment while still being more beneficial to students than switching to one condition after a traditional length experiment. However, these longer experiments do not mitigate measurement errors and will not address increased false positives when the conditions are equally effective; instead, strategies like de-biasing the collected data must be used. We briefly discuss possible strategies at the conclusion of the paper.

6. Temporal biases in distribution of students

By changing the proportion of students assigned to conditions as time goes by, MAB assignment introduces challenges if the student population is not uniformly distributed over time. For example, consider an experiment about which version of an online review maximizes scores on the associated homework. Some students will complete the review and homework soon after their release, while others will procrastinate. If procrastination tends to be associated with lower (or higher) scores on the homework, this may influence experimental results, even if the influence of procrastination is not related to condition. Because MAB assignment bases condition assignments on previous results, this biased pattern of which students engage with the experiment at a particular time may lead to the larger biases in estimates of condition means. For instance, if the procrastinating students have lower scores, then more low-scoring students may be assigned to the seemingly better condition, as most of the algorithm's exploration will take place at the beginning of the experiment. This skewed assignment may depress the estimate of the effectiveness of the better condition. When assigning students to conditions uniformly at random, this type of noise will average out: procrastination will still be associated with scores, but procrastination levels will be similarly distributed in each condition.

In the following simulations, we investigate the effects of this type of relationship between students' learning outcomes and the time step at which they are likely to participate in an experiment in the context of MAB assignment. If statistical impacts are similar to the impact of MAB assignment in the previous section, then it suggests that researchers need not explicitly consider such biases when using MAB experimental design; however, if statistical impacts are heightened or changed, then it may be necessary to test for whether a bias is present in a particular experimental context and if so, change the MAB algorithm to account for this bias.

6.1. Methods

All simulations used the same parameters as the previous *prior between* MAB simulations, with moderate effect size (when conditions differ), and sample sizes equal to $0.5m$, m, $2m$, and $4m$ (as before). Only MAB simulations are included because student order does not impact uniform assignment. We varied whether students with higher outcome measures tend to engage with the experiment earlier (*higher-earlier*) or later (*higher-later*), as well as the magnitude of the tendency for early and late students to differ. This bias is implemented as follows:

1. Rewards $r_{c1}^{(t)}$ and $r_{c2}^{(t)}$ are generated for each student t from the distributions of each condition, where $r_{c1}^{(t)}$ is the reward if the student is assigned to condition 1 and $r_{c2}^{(t)}$ the reward if assigned to condition 2.

2. Students are placed in sorted order based on $r_{c1}^{(t)} + r_{c2}^{(t)}$; sorting is ascending for *higher-earlier* and descending for *higher-later*. The sorted order is divided into quartiles.

3. Each quartile q_i is assigned a probability using a softmax function: $p(q_i) \propto \exp(\beta \cdot i)$, where β controls the degree of bias. Larger values of β place increasingly large probability on the final quartile.

4. The actual order that students will enter the experiment is calculated by repeatedly sampling without replacement using these quartile probabilities and uniform probability within each quartile, meaning that students in quartiles with higher probabilities will tend to enter the experiment prior to students in quartiles with lower probabilities.

We consider $\beta \in [0, 0.5]$, giving a maximum spread of 0.10 for the first quartile and 0.46 for the fourth quartile. Intuitively, as β increases, the students go from a random ordering to one that is more and more sorted by their rewards.

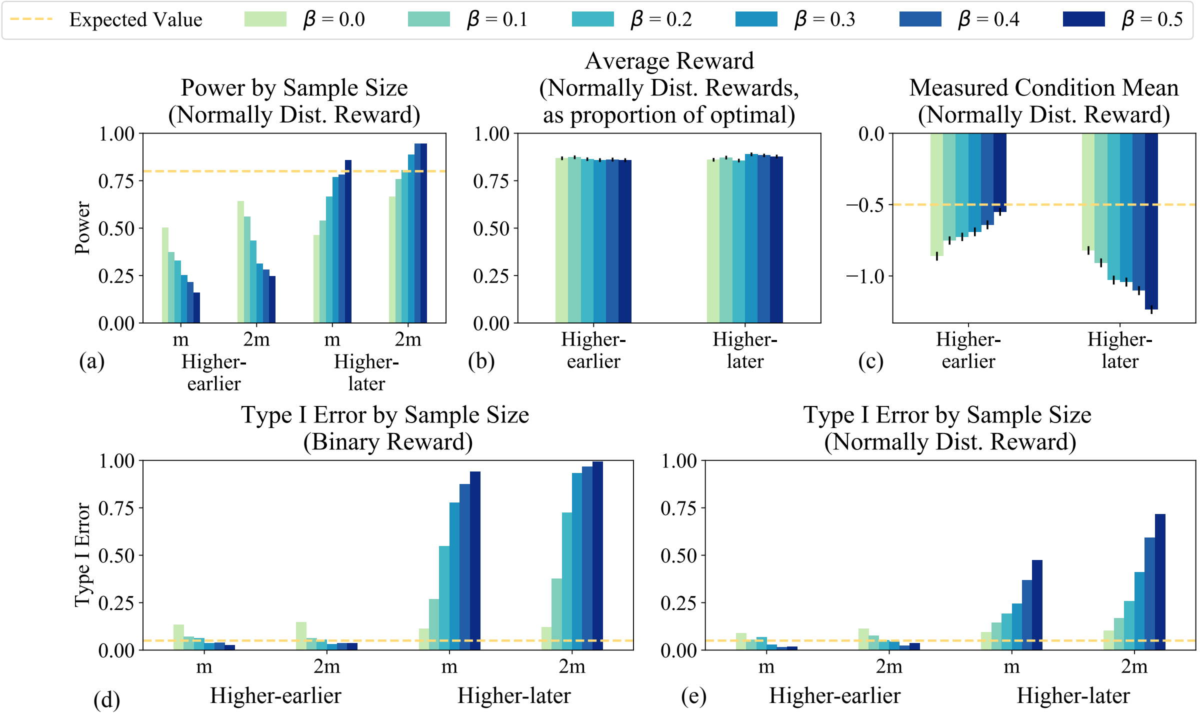

Figure 7: Results from simulations with bias in students' time to enter the experiment when conditions differ (a-c) and do not differ (d-e). (a) Power for normally distributed rewards for varying levels of bias (β). (b) Average reward per student, compared to mean of better condition: bias has no significant impact. Error bars reflect one standard error. (c) Measured average mean of the worse condition: when earlier students have higher scores, measurement is more accurate. (d-e) Type I error rates by reward type.

6.2. Results

6.2.1. Conditions that differ in effectiveness

As shown in Figure 7a, power is decreased when earlier students tend to have higher scores (*higher-earlier*), and this decrease is more pronounced as the magnitude of the bias (β) increases. In contrast, when later students tend to have higher scores (*higher-later*), power is increased and increases further with greater bias. This increased power is actually due to *less accurate* measurement: when later students have higher scores, the effectiveness of the worse condition is underestimated and the effectiveness of the better condition slightly overestimated, making the difference between conditions appear larger than it is (see Figure 7c for estimates of the worse condition). Logistic regression confirms both that β impacts power and that there is an interaction between the way β impacts power and whether it is earlier or later students who tend to have higher scores (coefficient for $\beta = -3.57$, $t(47994) = 39.8$, $p < .0001$; coefficient for interaction with *higher-later* $= 6.79$, $t(47994) = 50.9$, $p < .0001$). All Type S error rates were quite low ($< 0.5\%$) and not reliably impacted by the bias-related predictors.

Despite large differences in power, all simulations showed relatively similar average reward per student (Figure 7b).[3] The lack of an impact on reward occurs because while all types of

[3] For binary rewards, there was a significant main effect of β (coefficient $= -0.00838$, $t(23995) = 2.73$,

simulations are able to exploit the better condition, the times and amounts they explore differ. Early on, when most exploration happens, *higher-later*'s samples of both conditions tend to be relatively low, leading to lower estimated means for both conditions. Later, the algorithm tends to sample more from the better condition, leaving the estimate of the worse condition even lower than its actual performance (see Figure 7c). This larger difference means that a significant effect is detected even with the small number of participants in one condition, as the estimated effect size is larger than the true effect size. In contrast, in the *higher-earlier* case, the difference between means is smaller (and variances are similar), so the estimated effect size is more similar to the actual effect size and there are insufficient participants in one condition to determine if the effect is statistically reliable.

6.2.2. Conditions that are equally effective

As shown in Figure 7d, the false positive rate (Type I error rate) substantially increases with the amount of bias (β) for *higher-later*, reaching as high as 95%. Type I error rate decreases for *higher-earlier* (coefficient for $\beta = -2.66$, $t(47994) = 15.2$, $p < .0001$; coefficient for interaction between β and *higher-later* $= 7.79$, $t(47994) = 38.7$, $p < .0001$). This occurs because when *later* students have higher scores, condition estimates tend to increase over time. If by chance one condition appears better early on, and thus is sampled more, its mean estimate will increase because later samples tend to have higher value, leading to it being sampled even more frequently. Because more of the high-performing students will be assigned to that condition, there will appear to be a difference between conditions when in fact the difference is an artifact of the combination of when students enter the experiment and MAB assignment. In contrast, when *earlier* students have higher scores, further sampling decreases estimates, so chance differences early on will tend to lessen, and the algorithm will switch conditions more often. In this case, the algorithm essentially tends to always be optimistic about the future value of both conditions (as both are getting worse).

Figure 8 shows how the differences in estimated condition means vary across simulation, based on direction and level of bias. Since in actuality, there is no underlying difference between conditions, we see *higher-later* simulations (top row) becoming more accurate as β increases: the distribution of differences is increasingly centered around zero. For *higher-later* simulations (bottom row), as β increases, the distribution becomes bimodal: the estimated difference tends to be different from zero, which is reflected in the Type I errors.

Larger sample sizes also increase false positive rates, although the impact of larger sample sizes is smaller than the impact of the amount of bias. The increase in false positives is especially large for *higher-later*, as there is more opportunity for a chance difference between conditions to set up the reinforcing cycle of uneven sampling (coefficient 0.0035, $t(47994) = 38.8$, $p < .0001$).

7. MAB-ASSIGNMENT IN EDUCATIONAL EXPERIMENTS

The simulations in previous sections provide an understanding of how statistical power and benefits to students vary across types of outcomes and effect sizes. However, real experiments

$p < .01$) and the interaction between β and *higher-later* (coefficient $= 0.0136$, $t(23995) = 3.01$, $p < .01$) on average reward per student, but the size of these impacts was quite small.

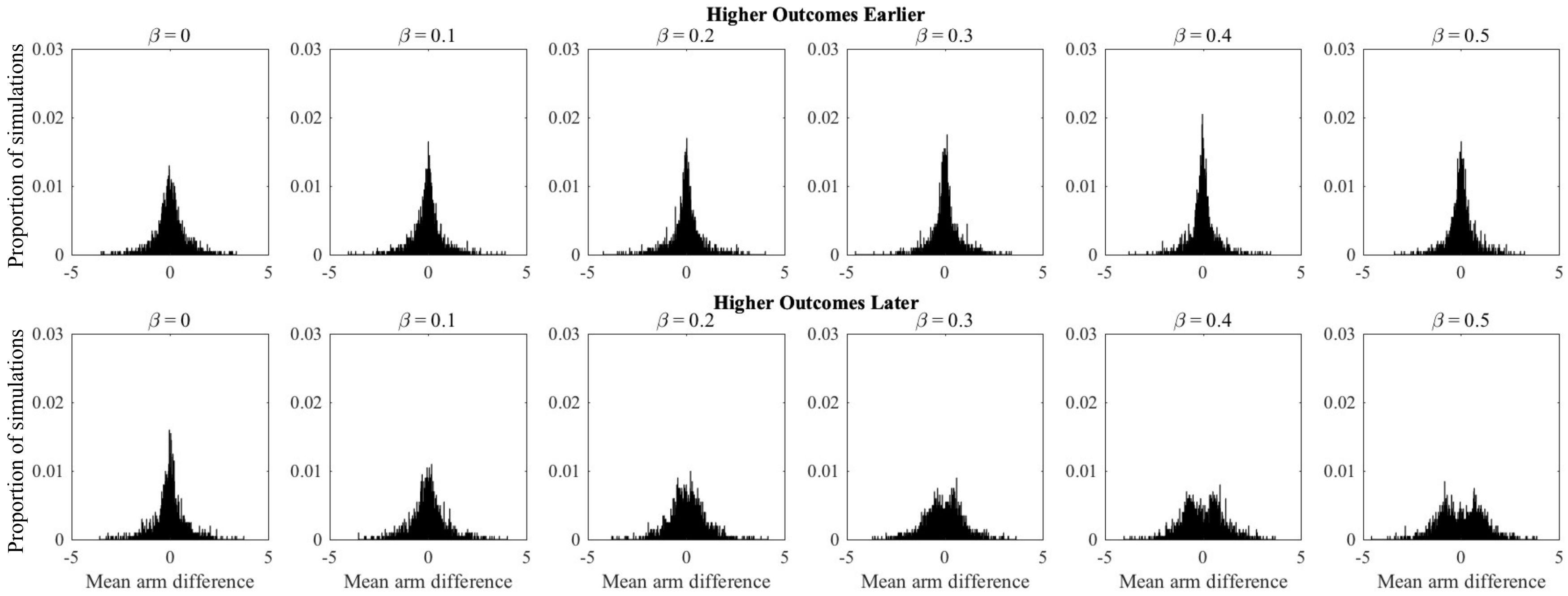

Figure 8: Histograms of measured differences between condition means for normal rewards simulations based on the amount of bias (β) and type, given that the condition means are actually equal. When *earlier* students have higher rewards, larger β is associated with smaller average difference between conditions; when *later* students have higher rewards, larger β is associated with more variance in the difference, and more simulations with differences farther from zero.

may differ from these simulations. All simulations assumed that outcomes were correctly modeled using a Bernoulli or normal distribution, and that for real-valued rewards, the variance of each condition was identical. In a real experiment, these assumptions may not hold, and the experiment may have a smaller or larger effect size relative to sample size as compared to the simulations. To understand how the effects found in the simulations might translate to real experiments, we analyzed MAB assignment in the context of significant/marginal results from a collection of twenty-two randomized controlled experiments run in ASSISTments (Selent et al., 2016). Our goal is to illustrate how MAB-assignment might impact a selection of typical educational experiments, both if the assumptions about the reward (outcome) models were correct and if these assumptions were incorrect. To address the first goal, we performed simulations in which we estimated reward models from previously collected data, and to address the second goal, we performed simulations of MAB-assignment that directly used the previously collected data and thus were consistent with a typical situation in which a researcher would not know with certainty that an outcome measure had a particular distribution prior to running the experiment.

7.1. Methods

Each of the 22 experiments included three outcomes (Selent et al., 2016): whether a student *completed* the assignment (solved three consecutive problems correctly), the *problem count* for a student to complete the assignment (only for students who completed the assignment), and the *log problem count* (base-10 logarithm of the problem count). Selent et al. (2016) note that they consider *log problem count* because *problem count* is positively skewed in some cases, with a long tail of students with large numbers of problem counts. We consider both *problem count* and *log problem count* as separate outcomes, both of which are real-valued and which we will model as normally distributed. While clearly both of these outcomes cannot be normally distributed within the same dataset, we model them in this way to test the robustness of MAB-assignment to mismatches between the assumed and actual reward model.

We analyzed a total of ten data sets: the four experiments with lowest p-values for each of the *problem count* and *log problem count* outcomes, and the two with lowest p-values for the *completed* outcome.[4] *Problem count* and *log problem count* were treated as normally distributed, and *completed* was binary. Because solving fewer problems to complete an assignment is desirable, the negation of *problem count* and *log problem count* were used as rewards.[5]

Two types of simulations were conducted: simulations using *parameters* based on the experiments and simulations using the actual measured *outcomes*. *Parameter* simulations used measured means (and variances) from the experiment, using these parameters to generate new samples. For example, in one experiment the average for *log problem count* in condition 1 was 1.21 (variance 0.012), and in condition 2 it was 1.12 (variance 0.011). Each time condition 1 was chosen, a new value was sampled from $\mathcal{N}(1.21, 0.012)$, and its negation was used as the reward. This approach permits assigning an unlimited number of simulated students to either condition, rather than only the actual number in the experiment. However, it also assumes the reward model is correct and the parameters measured experimentally were accurate. We compared MAB and uniform assignment, setting sample size to the number of students in the original experiment.

[4] p-values were used as a filter rather than effect sizes to exclude experiments with small sample sizes; only two experiments were analyzed for *completed* because only two reached even marginal significance for this outcome ($p < .1$).

[5] This is simplified, as problem count could be decreased by decreasing homework completion. Since this option is not possible in the simulations, we ignore it here.

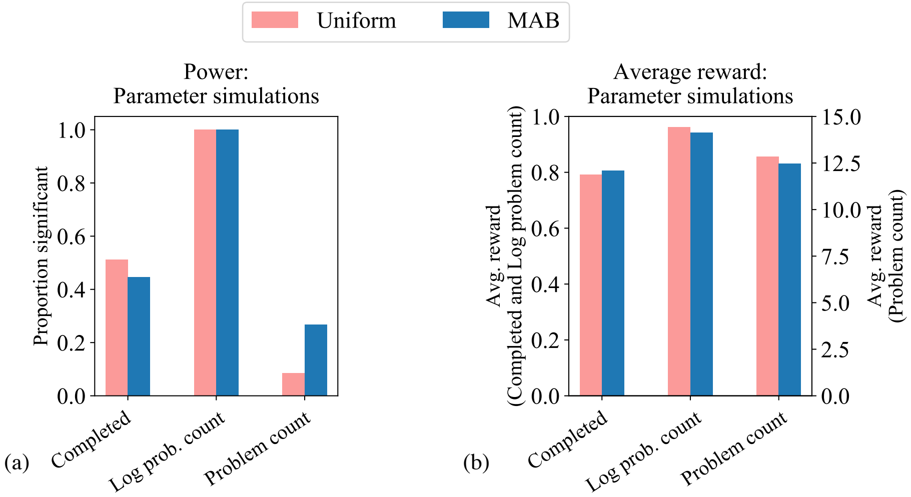

Figure 9: Results for *parameter* simulations, averaged across the educational experiments for each outcome measure. (a) Average power. (b) Average reward or cost per step. Higher bars are better for *completed*; lower bars (i.e., lower cost) are better for the other measures. Error bars indicate one standard error.

Outcome simulations relax the assumptions above by using the data as actually collected. Each time a condition is chosen, a student assigned to that condition in the data set is chosen randomly (without replacement) and their measured outcome used for the reward. These simulations terminate when no more students are available in a chosen condition. These simulations address the robustness of the MAB algorithm to cases where the reward distribution does not match the model, which will be the case for at least some of the *problem count* and *log problem count* simulations.

7.2. Results

7.2.1. Parameter simulations

As shown in Figure 9b, MAB assignment in the *parameter* simulations resulted in small improvements on average reward per student across all outcome measures, $t(9989) = 5.10$, $p < .0001$, with effect sizes ranging from $d = 0.07$ to $d = 13$ for individual experiments (median $d = 0.70$).

Impacts on power differed by assignment type as shown in Figure 9a. For *log problem count*, both MAB and uniform assignment always found a significant effect. MAB assignment did decrease power for the *completed* measure, consistent with the previous simulations. Counterintuitively, MAB assignment increases power for the *problem count* measure. This is due to the high variability for *problem count* and the fact that variability differed across conditions: MAB assignment can oversample a highly variable condition and gain a more confident estimate. While the figures summarize multiple experiments for each measure, these trends also held within the individual experiments. Across all experiments, Type S error rates were small, averaging 0.3% for uniform assignment and 0.4% for MAB.

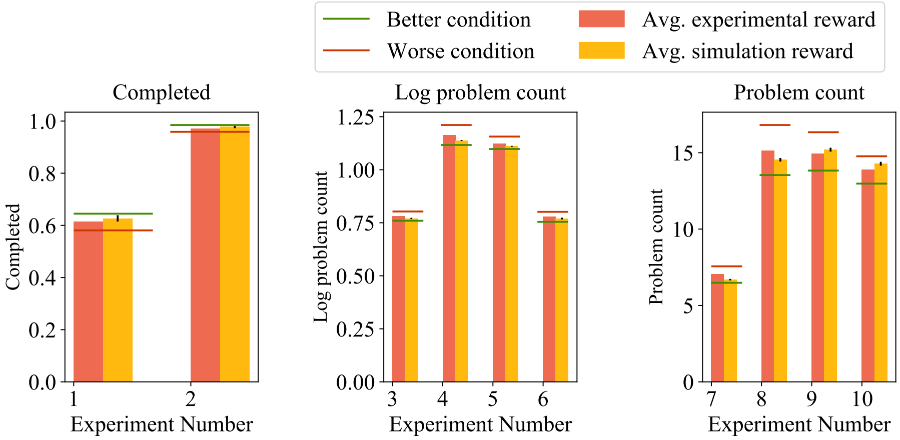

Figure 10: Observed rewards in ten previously conducted experiments and average reward using MAB assignment in the *outcome* simulations. “Better” and “worse” conditions are the average observed values for each condition in the experiment. Error bars indicate one standard error.

7.2.2. Outcome simulations

In the outcome simulations, we compared the actual average outcomes in each of ten experiments with the outcomes achieved using MAB assignment to conditions. As shown in Figure 10, MAB assignment achieved small improvements on the average value of each outcome measure for eight out of ten experiments. While the overall improvements are small, MAB assignment achieves outcomes that are almost as good as the better condition, which is the maximum possible. The two cases where MAB assignment did not improve on experimental rewards were for *problem count*, which had high variability. For *problem count*, the normal distribution may not be a good model for how the rewards are generated, and the lack of improvement may point to MAB assignment not performing as well when this mismatch is present. However, the log transform of problem count also did not necessarily result in a normally distributed outcome measure, and we did see improvements in reward for the experiments that used *log problem count* as an outcome measure. In these experiments, the “experimental” condition was not always better. Because MAB assignment adapts to the reward signal, it is indifferent to which condition is more effective and improved reward regardless of whether the control was better than the experimental condition.

Across all nine experiments where a significant effect was originally detected, 65% of simulations found a significant difference between conditions; 27% of simulations for the experiment with a marginally significant effect detected a difference between conditions. While these rates are lower than the desired power of 0.8, power analysis was not used to determine the original sample sizes (and indeed, power for uniform assignment in the *parameter* simulations was 0.55 and 0.44, respectively), and these simulations included an average of only 67% of the students in the original experiments, as the simulations terminated when MAB assignment chose a condition for which no new student remained. For all experiments, average Type S error rate was 0.1%.

Overall, while power is lowered and reward increased, the experimental modeling finds these effects to be less extreme than in the previous simulations. Power was actually increased for *problem count* due to high variability of this measure across students. The lack of a large increase in rewards is partially due to small differences between conditions, especially for the *completed* outcome, and because we used fixed prior means rather than individualizing them for each outcome, which tended to be highly optimistic for the *problem count* outcome. Additionally, the results in the *outcome* simulations suggest that a mismatch between reward model and distribution of the reward in the data, as was certainly present for some of the simulations using the real-valued rewards, may have only small impacts when using MAB assignment in real experiments.

8. Discussion

Randomized experiments can identify more effective educational strategies, but there are practical and ethical concerns about assigning students to suboptimal conditions. To examine the potential of MABs to more rapidly use data from online experiments to help students, this paper explored some consequences of using Thompson sampling, a common MAB algorithm, versus traditional uniform random assignment. Our simulations demonstrate that MAB assignment can benefit students. However, this comes at the cost of reliability: while reversing the sign of an effect is rare, power is decreased, and false positive rate (Type I error rate) is increased due both to the uneven distribution of students across conditions and measurement errors from adaptively changing the design based on previous results. An optimistic prior did lessen the impact on power, but twice as many students are still needed to achieve 0.8 power. Yet, the impacts on power may not be as important to researchers if they have the ability to run experiments with large sample sizes: even if many more students are needed for an experiment, it is still likely to lead to better outcomes for those students than a smaller experiment followed by choosing a single condition. Biases in the temporal ordering of students led to very high Type I error rates when students with higher rewards came later, indicating the need to monitor for such biases in real experiments. Overall, results suggest the benefits to students of MAB assignment using standard regret-minimizing algorithms must be weighed against the decreased ability to make reliable generalizations. Analyses of existing experiments demonstrated these effects in real-world data, although decreases in power were much smaller than in previous simulations.

One reason for the smaller decreases in power when modeling the real-world data may have been the choice of prior distribution. The real-valued rewards used very optimistic priors, especially for the problem count. This tends to lead to even more initial exploration than in our simulations of a more moderate optimistic prior. Researchers who wish to use MABs in their experimental designs may benefit from also employing this extreme optimism. Alternatively, or in addition, researchers could use stronger priors. This work did not directly explore the impact of the strength of the prior on results, but stronger priors that are the same across conditions will tend to maintain similar posterior distributions across conditions for a longer period of time. This will lead to more equal sampling across conditions, and assuming no temporal bias, this will tend to lead to more accurate measurement and correct statistical inferences.

In general, we took a frequentist approach to data analysis. Our goal was to illustrate the problems with using the types of statistical analyses that are most commonly employed for these types of experiments. Bayesian data analysis could be used to analyze the results, and we provided a limited demonstration of this in Section 3.1.1, showing that the results were not simply due to a particular choice of p-value. While we did not use Bayesian data analysis in the remainder of the paper, we note that doing so would not address the temporal bias issues that we raised nor would it mitigate the underlying measurement errors that occur with MAB assignment.

One motivation for educational experimentation with MAB assignment was to encourage teachers to allow experiments in their classes as their students would directly benefit (e.g., Williams et al. 2018). Our results suggest that this should be approached cautiously: while MAB assignment can increase the proportion of students in more effective conditions, researchers may be dissatisfied by the amount of information gained from the experiment. One possible solution is to use experiments with MAB assignment as a filter for which manipulations to replicate using typical methodologies. Power could be increased by increasing the α for statistical tests, and larger sample sizes are likely achievable in online settings. Such use will require more nuanced communication with teachers about the methodology and goals of a program of research but has the potential to address the needs of both teachers and researchers.

8.1. Limitations

There are several limitations to this work. First, we focused only on experiments with two conditions. We chose this focus to illustrate that challenges in statistical analysis occur even in very simple experimental settings. Similar issues will likely occur with larger numbers of conditions, especially when considering pairwise differences between conditions, although the exact pattern of differences in condition means will impact results. The ASSISTments experiments that we modeled in Section 7 demonstrate that two condition experiments are common.

Second, when considering alternative prior distributions, we kept the strength of all prior distributions the same, varying only the relation between the mean of the prior and the means of the actions. Increasing the strength of the prior (while maintaining symmetry across actions) would lead to lower rewards but higher power, as the algorithm would tend to more evenly distribute students across conditions. We focus on the relation of the mean of the prior compared to the actions in order to illustrate the subtleties of how the prior impacts results, but we encourage researchers to consider both the strength and location of the prior when deploying MAB algorithms for experimental design.

Finally, we focused on a regret-minimizing algorithm, rather than MAB algorithms for identifying the best condition (Audibert and Bubeck, 2010) or trading off rewards and measurement accuracy (Erraqabi et al., 2017). While exploring the statistical consequences of other objectives is important future work, our goal is to illustrate how standard MAB algorithms impact conclusions for researchers who may be excited by the potential benefits to students. We hope this will lead to careful consideration of what safeguards are needed to achieve both research and pedagogical aims, and that our focus on statistical significance demonstrates that MAB assignment can lead to erroneous generalizations in addition to measurement error.

8.2. Future work

One promising area for future work is to explore how to correct for the bias in the estimates of condition means. Correcting this bias would decrease false positive rates and provide more accurate measures of effect sizes. Bowden and Trippa (2017) develop unbiased estimators for adaptive experimental designs used in medical trials, and similar techniques may be effective for MAB designs. More generally, inverse probability weighting is a statistical technique that can be used in cases where parts of a population are oversampled or when the probability of assignment to a condition is uneven (see Mansournia and Altman (2016) for a short overview). We plan to explore how to adapt these types of techniques to more accurately assess condition effectiveness when data are collected via MAB assignment. Such techniques are likely to be more easily adapted to data collected via MAB algorithms like Thompson sampling that perform weighted randomization for each student, rather than MAB algorithms that choose one condition deterministically based on previous data. Exploring how to apply these methods and what the consequences are to statistical inference, especially likely continued decreased power and potential instability due to extreme assignment probabilities, is an important direction to enable adoption of MAB experimental designs.

Exploring alternative distributions for modeling reward is another area that is likely to facilitate deployment of MAB experimental designs. Many outcome measures, including those considered in our analyses of existing experiments, are not normally distributed, even if they tend to be treated as normally distributed in statistical hypothesis testing. For instance, the number of problems completed cannot be negative, and could exhibit positive skew. In our analyses of existing experiments, we treated this outcome as normally distributed, but an alternative path would have been to choose a different likelihood that more accurately models the outcome. For instance, one could use a negative binomial distribution. After choosing an alternative likelihood, a prior that is conjugate is likely to be the easiest approach in order to apply Thompson sampling. Practitioners must consider whether this approach fits their needs in terms of accurately reflecting the assumed form of the posterior distribution. Further simulations should be used to explore how deviations between the model of the outcome distributions and the actual outcome distributions impacts condition assignments, focusing especially on what types of discrepancies are associated with greater imbalances in the number of students in each condition.

In our simulations, we considered biased patterns of student engagement but did not examine detecting or correcting for such bias. There are a number of modifications that could be considered. First, one could assume that outcomes are non-stationary and use an algorithm that assumes rewards change over time (Besbes et al., 2014). While this would lead to more exploration than in the current simulations, there are still likely to be large measurement errors that lead to erroneous conclusions. To the extent that more-able students will tend to perform better regardless of condition, both conditions will tend to experience similar changes over time (i.e., both increasing or both decreasing). This means that further exploration at a particular point in time will still tend to provide information favoring the condition that was better previously, with continued limited sampling of the other condition.

Another way of addressing patterns of student engagement would be to model temporal bias explicitly. Modeling temporal bias in real online experiments would demonstrate whether our assumptions about student engagement are accurate and if not, provide a fuller picture for what kinds of biases should be accounted for in online experiments. For example, we assumed that which condition will be more effective for a student is unrelated to when she will engage with the experiment. However, it could be that students who tend to procrastinate are also those who spend less time on homework, which could have differential impacts on questions like what type of hints are most effective if some hints require more time to examine than others.

Modeling temporal bias raises the question of whether it would be fruitful to model other individual student characteristics when assigning students to conditions and determining the efficacy of those conditions. Individual characteristics could be incorporated via contextual bandit algorithms, which assume there is a vector of additional information ("context variables") at each time step and that information is used to determine which condition is best for this particular instance; for example, a contextual bandits version of Thompson sampling uses regression to model the relation between the additional information (predictors) and the mean of each choice (Agrawal and Goyal, 2013). Lan and Baraniuk (2016) used contextual bandits to provide personalized study suggestions for students, choosing, for instance, if a particular student would be most helped by viewing a video or solving a problem. While we are optimistic about the potential of contextual bandits for personalizing educational technologies, using contextual bandits for experimentation would be addressing the question of "what works best for what type of students" (or in what type of situations). Here, we have tackled the simpler research question of "what works best," and demonstrated that statistical challenges arise even in this setting. This focus corresponds to concentrating on main effects rather than interactions. Contextual bandits might be able to address temporal bias if the researcher included time of engagement as a contextual variable, but our work makes clear that if this bias occurs and is not accounted for in the experimental design, it can lead to erroneous conclusions.

8.3. Conclusion

Online educational technologies offer opportunities for easily conducting experiments within real pedagogical contexts, but as experiments become more ubiquitous, it is vital that they meet the needs of students, teachers, and researchers. Our work demonstrates the consequences of using MAB assignment to mitigate costs to students, exploring different contexts that may arise based on the intervention and educational setting (e.g., varying sample sizes and effect sizes), and points to the fact that for many experiments, standard MAB algorithms will limit statistical power and increase false positives due both to unbalanced assignment across conditions and to adaptively changing assignment policy based on past results. While increasing sample sizes mitigates the loss in power without negative consequences for students compared to typical experimental practices, larger sample sizes do not decrease the false positive rate. Researchers must be mindful of these consequences when deciding if the increased benefits to students justify the limitations to statistical inference, and may wish to follow up promising experiments using MAB assignment with experiments using traditional uniform assignment.

9. Acknowledgments

We thank Louie Hoang for his work on initial software development for this project. This work was partially funded by Office of Naval Research grant number #N00014-18-1-2755 to Joseph J. Williams.

References

Agrawal, S. and Goyal, N. 2012. Analysis of Thompson sampling for the multi-armed bandit problem. In *Proceedings of the 25th Annual Conference on Learning Theory*, S. Mannor, N. Srebro, and R. C. Williamson, Eds. Vol. 23. PMLR, Edinburgh, Scotland, 39.1–39.26.

Agrawal, S. and Goyal, N. 2013. Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the 30th International Conference on International Conference on Machine Learning*, S. Dasgupta and D. McAllester, Eds. Vol. 28. JMLR, 127–135.

Atkinson, A. C. 2014. Selecting a biased-coin design. *Statistical Science 29,* 1, 144–163.

Audibert, J.-Y. and Bubeck, S. 2010. Best arm identification in multi-armed bandits. In *Proceedings of the 23rd Annual Conference on Learning Theory*. 41–53.

Bassler, D., Briel, M., Montori, V. M., Lane, M., Glasziou, P., Zhou, Q., Heels-Ansdell, D., Walter, S. D., Guyatt, G. H., Group, S.-. S., et al. 2010. Stopping randomized trials early for benefit and estimation of treatment effects: Systematic review and meta-regression analysis. *JAMA 303,* 12, 1180–1187.

Besbes, O., Gur, Y., and Zeevi, A. 2014. Stochastic multi-armed-bandit problem with non-stationary rewards. In *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds. Curran Associates, Inc., 199–207.

Bowden, J. and Trippa, L. 2017. Unbiased estimation for response adaptive clinical trials. *Statistical Methods in Medical Research 26,* 5, 2376–2388.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., and Munafò, M. R. 2013. Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience 14,* 5, 365–376.

Cavagnaro, D. R., Myung, J. I., Pitt, M. A., and Kujala, J. V. 2010. Adaptive design optimization: A mutual information-based approach to model discrimination in cognitive science. *Neural Computation 22,* 4, 887–905.

Chapelle, O. and Li, L. 2011. An empirical evaluation of Thompson sampling. In *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2249–2257.

Chow, S.-C., Wang, H., and Shao, J. 2007. *Sample size calculations in clinical research.* CRC Press, Boca Raton, FL.

Clement, B., Roy, D., Oudeyer, P.-Y., and Lopes, M. 2015. Multi-armed bandits for intelligent tutoring systems. *Journal of Educational Data Mining 7*, 20–48.

Cohen, J. 1988. *Statistical power analysis for the behavioral sciences*, 2 ed. Lawrence Erlbaum Associates, Mahwah, NJ.

Demets, D. L. and Lan, K. 1994. Interim analysis: the alpha spending function approach. *Statistics in Medicine 13,* 13-14, 1341–1352.

Duan, L. and Hu, F. 2009. Doubly adaptive biased coin designs with heterogeneous responses. *Journal of Statistical Planning and Inference 139,* 9, 3220–3230.

Eisele, J. R. and Woodroofe, M. B. 1995. Central limit theorems for doubly adaptive biased coin designs. *The Annals of Statistics 23,* 1, 234–254.

Erraqabi, A., Lazaric, A., Valko, M., Brunskill, E., and Liu, Y.-E. 2017. Trading off rewards and errors in multi-armed bandits. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, A. Singh and J. Zhu, Eds. Vol. 54. PMLR, 709–717.

Gelman, A. and Carlin, J. 2014. Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science 9,* 6, 641–651.

Hu, F. and Rosenberger, W. F. 2003. Optimality, variability, power: evaluating response-adaptive randomization procedures for treatment comparisons. *Journal of the American Statistical Association 98,* 463, 671–678.

Hu, F. and Rosenberger, W. F. 2006. *The theory of response-adaptive randomization in clinical trials.* Vol. 525. John Wiley & Sons, Hoboken, NJ.

Jennison, C. and Turnbull, B. W. 2005. Meta-analyses and adaptive group sequential designs in the clinical development process. *Journal of Biopharmaceutical Statistics 15,* 4, 537–558.

Kaufmann, E., Cappé, O., and Garivier, A. 2016. On the complexity of best arm identification in multi-armed bandit models. *Journal of Machine Learning Research 17,* 1, 1–42.

KULESHOV, V. AND PRECUP, D. 2014. Algorithms for multi-armed bandit problems. *arXiv preprint arXiv:1402.6028.*

LAI, T. L. AND ROBBINS, H. 1985. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics 6,* 1, 4–22.

LAN, A. S. AND BARANIUK, R. G. 2016. A contextual bandits framework for personalized learning action selection. In *Proceedings of the Ninth International Conference on Educational Data Mining*, T. Barnes, M. Chi, and M. Feng, Eds. 424–429.

LANGFORD, J. AND ZHANG, T. 2008. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in Neural Information Processing Systems 21*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds. Curran Associates, Inc., 817–824.

LI, L., CHU, W., LANGFORD, J., AND SCHAPIRE, R. E. 2010. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*. ACM, 661–670.

LIU, Y.-E., MANDEL, T., BRUNSKILL, E., AND POPOVIC, Z. 2014. Trading off scientific knowledge and user learning with multi-armed bandits. In *Proceedings of the 7th International Conference on Educational Data Mining*, J. Stamper, Z. Pardos, M. Mavrikis, and B. McLaren, Eds. 161–168.

MANSOURNIA, M. A. AND ALTMAN, D. G. 2016. Inverse probability weighting. *British Medical Journal 352*, i189.

MU, T., WANG, S., ANDERSEN, E., AND BRUNSKILL, E. 2018. Combining adaptivity with progression ordering for intelligent tutoring systems. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*. ACM, 15:1–15:4.

RADLINSKI, F., KLEINBERG, R., AND JOACHIMS, T. 2008. Learning diverse rankings with multi-armed bandits. In *Proceedings of the 25th International Conference on Machine Learning*, A. McCallum and S. Roweis, Eds. ACM, 784–791.

SCOTT, S. L. 2010. A modern Bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry 26,* 6, 639–658.

SEGAL, A., DAVID, Y. B., WILLIAMS, J. J., GAL, K., AND SHALOM, Y. 2018. Combining difficulty ranking with multi-armed bandits to sequence educational content. In *Proceedings of the 19th International Conference on Artificial Intelligence in Education*, C. Penstein Rosé, R. Martnez-Maldonado, U. Hoppe, R. Luckin, M. Mavrikis, K. Porayska-Pomsta, B. McLaren, and B. du Boulay, Eds. Springer, 317–321.

SELENT, D., PATIKORN, T., AND HEFFERNAN, N. 2016. ASSISTments dataset from multiple randomized controlled experiments. In *Proceedings of the Third ACM Conference on Learning at Scale*. ACM, 181–184.

TANG, L., JIANG, Y., LI, L., AND LI, T. 2014. Ensemble contextual bandits for personalized recommendation. In *Proceedings of the 8th ACM Conference on Recommender Systems*. ACM, 73–80.

TANG, L., ROSALES, R., SINGH, A., AND AGARWAL, D. 2013. Automatic ad format selection via contextual bandits. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*. ACM, 1587–1594.

WELCH, B. L. 1938. The significance of the difference between two means when the population variances are unequal. *Biometrika 29,* 3/4, 350–362.

WILLIAMS, J. J., KIM, J., RAFFERTY, A., MALDONADO, S., GAJOS, K. Z., LASECKI, W. S., AND HEFFERNAN, N. 2016. Axis: Generating explanations at scale with learnersourcing and machine learning. In *Proceedings of the Third ACM Conference on Learning at Scale*. ACM, 379–388.

Williams, J. J., Rafferty, A. N., Tingley, D., Ang, A., Lasecki, W. S., and Kim, J. 2018. Enhancing online problems through instructor-centered tools for randomized experiments. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 207:1–207:12.

Xu, J., Xing, T., and Van Der Schaar, M. 2016. Personalized course sequence recommendations. *IEEE Transactions on Signal Processing 64,* 20, 5340–5352.