Adversarial bandits for drawing generalizable conclusions in non-adversarial experiments: an empirical study

Yang Zhi-Han
Carleton College
yangz2@carleton.edu

Shiyue Zhang
Carleton College
zhangs@carleton.edu

Anna N. Rafferty
Carleton College
arafferty@carleton.edu

ABSTRACT

Online educational technologies facilitate pedagogical experimentation, but typical experimental designs assign a fixed proportion of students to each condition, even if early results suggest some are ineffective. Experimental designs using multi-armed bandit (MAB) algorithms vary the probability of condition assignment for a new student based on prior results, placing more students in more effective conditions. While stochastic MAB algorithms have been used for educational experiments, they collect data that decreases power and increases false positive rates [22]. Instead, we propose using adversarial MAB algorithms, which are less exploitative and thus may exhibit more robustness. Through simulations involving data from 20+ educational experiments [29], we show data collected using adversarial MAB algorithms does not have the statistical downsides of that from stochastic MAB algorithms. Further, we explore how differences in condition variability (e.g., performance gaps between students being narrowed by an intervention) impact MAB versus uniform experimental design. Data from stochastic MAB algorithms systematically reduce power when the better arm is less variable, while increasing it when the better arm is more variable; data from the adversarial MAB algorithms results in the same statistical power as uniform assignment. Overall, these results demonstrate that adversarial MAB algorithms are a viable "off-the-shelf" solution for researchers who want to preserve the statistical power of standard experimental designs while also benefiting student participants.

Keywords

MAB, bandits, experimental design, hypothesis testing

1. INTRODUCTION

Digital educational technologies offer unique opportunities to conduct pedagogical experiments and learn how to improve student outcomes. For example, experimenters can compare worked examples versus tutoring [19] or vary an avatar's dialect [9]. When an intervention's impact is measured soon after the intervention (e.g., via response times as in [32] or later problem correctness as in [18]), real-time data could be used to direct more students to more effective conditions. Multi-armed bandit (MAB) algorithms have been proposed as a way to conduct such adaptive experiments [15] and used for optimizing A/B comparisons (e.g. [16, 25, 26]).

While MAB assignment tends to improve outcomes for participants, it poses problems for drawing conclusions from the data. Prior research has shown systematic measurement errors, increases in false positive rate (FPR), and decreases in power when stochastic MAB algorithms are used for A/B comparisons (e.g., [22]). Potential benefits to student participants could thus be outweighed by harm to the research: lower power decreases the probability that effective interventions will be detected and deployed outside the trial, and higher FPR may lead to deploying unhelpful interventions at significant cost. While developing new algorithms and analysis approaches for these challenges is an active area of research (e.g., [5, 8]), these approaches are not yet an "off-the-shelf" solution for MAB-based experimental design.

In this paper, we consider the impact of using MAB algorithms that make performance guarantees in the adversarial case. These algorithms make weaker assumptions about their environment [4], decreasing the degree to which they can assign more students to a perceived better condition. Yet, we hypothesize that these relaxed assumptions may also decrease the *negative* consequences for drawing conclusions from the data, resulting in collected data that is more robust to the realities of educational experiments. Adversarial MAB algorithms could thus be used by researchers who want some of the benefits of condition-assignment via MAB algorithms but where their primary focus remains on drawing generalizable conclusions. These algorithms could also be *more* effective than uniform random assignment in some cases, as they can be sensitive to condition variability, and allocating more participants to an extremely variable condition can result in a better measurement of its effectiveness.

Using simulations, we first explore how 22 previously conducted experiments [29] might have been impacted if conditions had been assigned with a stochastic bandit algorithm (Thompson sampling [30]) or with one of three adversarial MAB algorithms in the Exp family [4], rather than with uniform random assignment.[1] These experiments were all conducted in ASSISTments homework assignments [11],

[1]All code: http://tiny.cc/MABExpDesign (OSF link).

Y. Zhi-Han, S. Zhang, and A. Rafferty. Adversarial bandits for drawing generalizable conclusions in non-adversarial experiments: an empirical study. In A. Mitrovic and N. Bosch, editors, *Proceedings of the 15th International Conference on Educational Data Mining*, pages 353–360, Durham, United Kingdom, July 2022. International Educational Data Mining Society.

https://doi.org/10.5281/zenodo.6853039

leading to a thorough investigation of how these algorithms might perform in real-world settings. We go beyond prior work that has often focused on binary outcomes for students (e.g. [35]) to examine real-valued outcomes that may follow idiosyncratic distribution patterns. We show that the Exp family of adversarial bandit algorithms largely avoids the measurement and hypothesis testing inaccuracies incurred by stochastic bandits, while still providing a small but reliable improvement in average student outcomes.

We then turn to exploring MAB algorithms in the context of experiments where conditions differ in variance, which occurs when interventions narrow (or widen) gaps among students. We find that the potential power advantage for adversarial MAB algorithms is not realized in practice, although there are some scenarios in which the more exploitative stochastic MAB algorithm does increase power.

Overall, we make the following contributions: (a) introducing the idea of adversarial and stochastic-adversarial MAB algorithms as an off-the-shelf solution for allocating students to conditions; (b) demonstrating that these MAB algorithms have minimal detrimental impacts on the performance of statistical hypothesis testing in a range of real-world educational experiments; and (c) illustrating that differences in variance across two conditions are not sufficient for condition assignment with adversarial MAB algorithms to increase power over uniform assignment. These results suggest that adversarial and stochastic-adversarial MAB algorithms offer a good solution for researchers who want to improve student participant's experiences without negatively impacting their own ability to learn from the experiment and improve experiences for the many students who are not participants.

2. RELATED WORK

Assigning participants to experimental conditions using MAB algorithms has been proposed as an alternative to uniform assignment (e.g., [12, 31, 34]). In clinical trials, a variety of methods that adapt based on previous results have been presented (e.g., [3,5,7], often to more quickly use data to benefit patients (e.g., [36]). More closely related to our work, applications of MAB algorithms to educational settings can help to more quickly identify pedagogically effective conditions (e.g., [15, 16, 33]) and lower barriers to teachers conducting experiments in their classrooms [35]. While MAB algorithms have also been used in education to assign students to educational interventions with the sole goal of producing the best outcomes for that particular student (e.g., [6, 13]), we focus here specifically on cases where there is a desire to extrapolate beyond individual students and draw generalizable conclusions, as in traditional scientific experiments.

While MAB algorithms are a natural choice for assigning participants to conditions to limit how many students are assigned to less effective conditions, these algorithms have consequences for researchers' ability to use the collected data to draw conclusions about the relative effectiveness of conditions. Traditional stochastic MAB algorithms minimize regret, but this can lead to systematically biased estimates of arm means (e.g., [5, 20, 37]). They also impact traditional statistical hypothesis testing: false positive rates, where there is no actual effect in the population but the test points to there being an effect, can be increased, and power, which measures how often a test will detect a difference when there is one, can be decreased (see e.g., [22], for specific documentation of these phenomena with Thompson sampling [1, 30], which we compare to in this paper). A variety of approaches have been taken to addressing these issues, including both statistical approaches to create unbiased estimators from the collected data (e.g., [5]) and algorithmic approaches that incorporate measurement into the algorithms' objective (e.g., by including estimation accuracy in the objective [8] or incorporating lower bounds on power [38]). Using such approaches off-the-shelf can be challenging: the power-constrained bandits algorithm [38] makes multiple decisions about one participant, rather than learning across participants, and researchers may wish to use their standard statistical estimators and tests rather than switching to a different paradigm. In this paper, we take a slightly different approach by exploring an alternative class of MAB algorithms and examining their performance in real-world scenarios.

3. BANDIT-DRIVEN EXPERIMENTS

MAB problems are a kind of reinforcement learning problem focused on maximizing immediate rewards. A classic example is allocating limited pulls to a set of slot machines. In these problems, agents must balance collecting new information about little explored arms and exploiting information from received rewards. Here, we model selecting better educational interventions as a MAB problem: Each intervention is an arm (action) the system can choose, and initially, rewards are unknown. After a student experiences an intervention, the system receives a stochastic reward (e.g., a measure of the student's understanding/efficiency) that influences the arm choice for the next student.

Stochastic bandit algorithms. These algorithms assume there is some stationary reward distribution underlying each arm. Here, we focus on Thompson sampling (TS; [2, 30]), which maintains an estimate of the reward distribution of each arm. From the reward that it gets from each choice, it updates this distribution. At each time step, the algorithm samples from the posterior reward distribution of each arm and chooses the arm that has the highest sampled value.

Adversarial bandit algorithms. Adversarial bandits make no statistical assumptions about the reward distribution of each arm, making them appropriate for reward distributions that are non-stationary or of unknown form. Because of the lack of assumptions, they are designed to explore more strongly and adapt more quickly to perceived changes in reward distributions. If the rewards in fact follow stationary distributions, this can lead to lower expected reward than stochastic bandit algorithms, but when reward distributions deviate from these assumptions, adversarial bandit algorithms have stronger performance guarantees than stochastic bandit algorithms. From the perspective of using MAB algorithms for experimental design, the extra exploration in adversarial algorithms could collect better data for drawing conclusions about differences between arms, albeit while lowering benefits for participants in the experiment.

In this work, we focus on the popular Exp3 family of adversarial bandit algorithms [4]. Exp3 balances exploration and exploitation via an exploration hyperparameter that influ-

ences both the probability of picking an arm uniformly at random and the strength of response to high rewards from low-probability arms. This hyperparameter allows an experimenter to adjust the amount of exploration, potentially increasing reward at the cost of collecting more biased data. Because this choice is difficult to optimize ahead of time, we also examine the performance of Exp3.1, which eliminates the hyperparameter and provides worst case performance guarantees regardless of the true reward distributions [4].

One algorithm that has performance guarantees in both stochastic and adversarial environments is Exp3++ [28]. While this algorithm achieves lower expected rewards than TS, it often improves upon the obtained reward of Exp3 while still employing enough exploration to perform well in adversarial environments. It also can be used off-the-shelf, with fixed values for the hyperparameters that probabilistically guarantee asymptotic performance [27].

We want to explore how well adversarial and stochastic MAB algorithms meet the needs of researchers for data collection in educational experiments and how they impact student experiences compared to traditional uniform assignment. We hypothesize that the adversarial bandit algorithms (Exp3 with a fixed value of 0.05 for the hyperparameter, Exp3.1, Exp3++) will have comparable performance for collecting research data to uniform random allocation, with benefits to students that are greater than uniform random allocation but less than those from a representative stochastic bandit algorithm (TS). Further, Exp3++ is likely to improve on the purely adversarial algorithms' performances in assigning more students to better arms.

4. EVALUATING ADVERSARIAL BANDITS: ASSISTMENTS EXPERIMENTS

The probabilistic asymptotic performance guarantees of MAB algorithms, both in stochastic and adversarial environments, suggest that student participants would benefit if these algorithms were used for experimental design. However, these guarantees do not speak to how biased the collected data will be, nor whether standard statistical hypothesis testing will be able to draw accurate conclusions from that data. Further, real educational experiments may have non-normally distributed outcome measures, impacting the collected data and performance of each algorithm. To explore how well the Exp3 variants meet the needs of researchers to draw accurate conclusions and the desire to place more students in a better condition, we conduct simulations that leverage previously collected datasets from educational experiments, conducted on the ASSISTMents platform [11]. We use these datasets as a case study for how to apply MAB algorithms to scenarios with varying, real-valued reward distributions, where some students reach mastery quickly and others never do so, and by repeatedly simulating the potential impact of each of the four bandit algorithms (TS, Exp3, Exp3.1, Exp3++), we can measure statistical power, false positive rate (FPR), and accuracy of arm measurement across algorithms.

4.1 Methods

4.1.1 Modeling real-world datasets

We focus on datasets from 22 randomized controlled experiments run inside the *SkillBuilder* interface of the ASSISTments online learning platform [29], which focuses primarily on 4th-12th grade math. These datasets included a total of 14,947 students in grades 5-12 (25% of students lacked a reported grade). Students had 200 different teachers and were drawn from 19 states (states deidentified in the data), and included a "guessed" gender based on name for 68% of students (of these, 53% were female; see [10] for information on gender methodology). No information on student race/ethnicity or SES was available, and experiments were IRB approved; see [29] for more dataset details.

In each experiment, students were placed into one of two conditions when completing homework. Each student must answer several *consecutive* problems correctly to complete the homework (typically three), and the number of problems P the student attempted before completion was recorded. Both completing homework and doing so in fewer problems are desirable, and we translate these measures into a reward signal for the MAB algorithms. To eliminate scaling issues, rewards are scaled to a fixed range as follows. Based on examination of the range of problems to completion across experiments, we cap the maximum number of problems at 30. If $P \geq 30$ or the student did not complete the homework, then we set P to 30. Because lower values of P are better, reward is then $r = 30 - P$. This reward is guaranteed to be in $[0, 30]$ and is then linearly interpolated into $[0, 1]$ for the MAB algorithms. We refer to the better condition as arm 1 and the worse condition as arm 2 on all datasets.

To conduct repeated trials with data from previously conducted experiments, our framework *resamples* an outcome associated with the chosen condition in the dataset when that condition is assigned to an incoming student. Within each trial, we fix the number of students n to the number in the original experiment ($n \in [129, 1797]$).

4.1.2 Simulation setup

To assess the ability of traditional hypothesis tests to draw accurate conclusions from the collected data, we measure (1) power – the proportion of the time an effect is detected when one exists – using scenarios where two conditions are *different* and (2) false positive rate (FPR) – the proportion of the time an effect is falsely detected when one does not exist – using scenarios where two conditions are the *same* in terms of expected reward. For (1), we focus on the seven ASSISTments datasets with the largest effect sizes measured in terms of Cohen's d ($0.16 \sim 0.51$), as larger effect sizes are more likely to reflect educationally relevant differences and may lead to larger differences across algorithms; we refer to these as ASSISTments-GES (greater effect size). We examine how the allocation methods impact average reward, which measures student outcomes, and statistical power and measured arm means, which are of large importance to researchers. For (2), we create modified datasets in which each condition's reward list contains all rewards from *both* arms. This creates two conditions with the same expected outcome while keeping their realistic reward distributions. We refer to these 22 modified datasets as ASSISTments-RC (reward combined). Because no allocation method could increase benefits to students given that the conditions do not differ, we focus here on examining FPR and measured arm means. For both (1) and (2), we follow Section 4.1.1 and run 1000 trials for each dataset-algorithm combination.

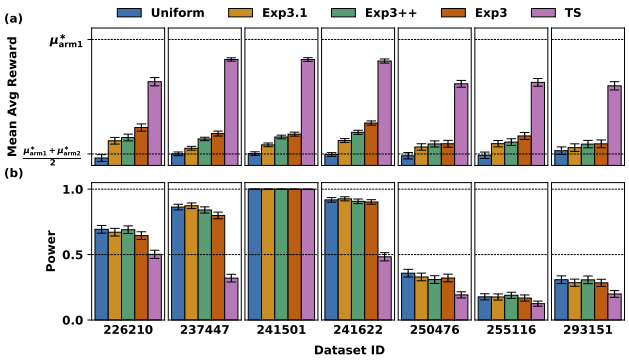

Figure 1: In terms of benefiting students as measured by mean average reward (a), all Exp allocations under-performed TS but out-performed uniform (except on dataset 293151). This was possible even when Exp3.1 and Exp3++ allocation achieved comparable power to uniform allocation, as indicated by overlapping error bars on all seven datasets (b). Error bars show $\pm 1.96 \times$ SE.

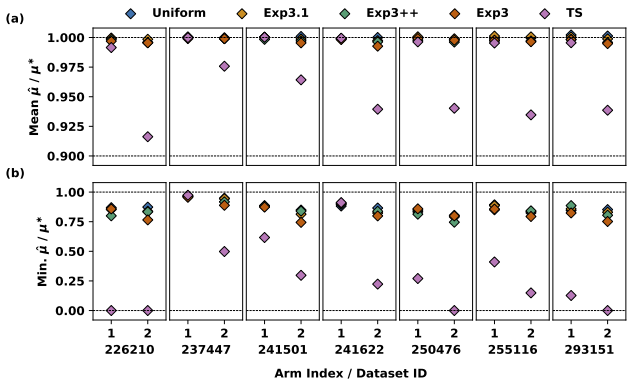

Figure 2: TS allocation led to a clear, systematic underestimation of arm 2 [worse arm] on all seven datasets (a), while Exp allocations resulted in arm mean estimates similar to those of Uniform (a). Error bars are not shown for clarity and are delegated to hypothesis testing. Also, the worst-case $\hat{\mu}$'s are much worse for TS allocation than for Exp (b).

4.1.3 Data collection and analysis

For each trial, we record all condition assignments and unscaled rewards. We then compute: the average reward per student, the average reward $\hat{\mu}$ for each condition (true expected reward denoted by μ^*), and the conclusion of a two-sided hypothesis test of whether the two conditions differ at a population level. Because capping P as discussed in Section 4.1.1 can lead to strong bimodality in reward distributions, which dissatisfies the assumptions of a standard t-test, we use the non-parametric Brunner-Munzel test for testing for a difference between conditions in the collected data. This test assumes neither normality nor equivariance of the distributions from which the samples are drawn. We consider the test to detect an effect if and only if $p < .05$.

To determine if a statistic differed reliably based on condition allocation method, we use generalized linear regression, with factors for algorithm (with uniform as reference group) and ASSISTments dataset. We report two decimal places except for small or similar values.

4.2 Results

4.2.1 Datasets with effects: ASSISTments-GES

Mean average reward: All MAB allocations were associated with significantly higher benefits to students, as measured by mean average reward, than uniform allocation (coefficient for TS: 0.80; Exp3: 0.19; Exp3.1: 0.10; Exp3++: 0.15; all $p < .001$). TS collected data with highest mean average reward (16.34), followed by Exp3 (15.74; 24% of the reward gain of TS over uniform), Exp3++ (15.69; 19% of TS over uniform), Exp3.1 (15.64; 12% of TS over uniform), and Uniform (15.54). For a breakdown by dataset, see Figure 1a.

Power: TS and Exp3 allocations had significantly lower power than uniform (coefficient for TS: -1.42; Exp3: -0.18; both $p < .001$), while allocation with the other two algorithms did not (Exp3.1: -0.05, $p = .22$; Exp3++: -0.07, $p = .11$). On average, TS collected data with lowest power (0.40), followed by Exp3 (0.59), Exp3++ (0.61), Exp3.1 (0.61) and Uniform (0.62). Figure 1b shows a breakdown by dataset.

Measured arm means: Our work replicates prior work showing that TS underestimates the worse arm in binary trials [22] and further finds a small underestimation of the better arm (arm 1 [better arm] coefficient: -0.05; arm 2 [worse arm] coefficient: -0.76; all $p < .001$). The Exp algorithms resulted in more accurate measurement of arm means: while Exp3 and Exp3++ also underestimate both arms, the extent to which they underestimate the worse arm is much less than TS (for Exp3: arm 1 coefficient: -0.02, $p < .05$; arm 2 coefficient: -0.04, $p < .05$; for Exp3++: arm 1 coefficient: -0.02, $p < .01$; arm 2 coefficient: -0.04, $p < .05$). Arm mean estimates using data from Exp3.1 did not differ significantly from those derived from uniform allocation (arm 1 coefficient: 0.004, $p = .71$; arm 2 coefficient: -0.007, $p = .69$). See Figure 2a for a qualitative comparison by dataset and Figure 2b for a worst-case analysis showing results for the trial with the most inaccurate $\hat{\mu}$.

4.2.2 Datasets without effects: ASSISTments-RC

FPR: TS allocation was associated with significantly higher FPR than uniform allocation (coefficient for TS: 0.56, $p < .001$), while allocation with the other three algorithms was not (coefficient for Exp3: 0.0065, $p = .88$; for Exp3.1: -0.0019, $p = .97$; for Exp3++: -0.0047, $p = .91$). On average, TS collected data with the highest FPR (0.0867) and is followed by Exp3 (0.0517), Uniform (0.0514), Exp3.1 (0.0513) and Exp3++ (0.0512). As shown in Figure 3a, TS inflates FPR for almost all datasets, while the other algorithms have FPR $< .06$ for the vast majority of datasets.

Measured arm means: Since the two arms are identical and within a trial, their average values are not independent, we arbitrarily examine one of the two arms for each trial. Allocation using TS, Exp3, and Exp3++ was associated with significantly lower estimates of arm means compared to uniform allocation (coefficient for TS: -0.42; Exp3: -0.0320; Exp3++: -0.0274; all $p < .001$); note that the bias for TS is much larger than for Exp3 and Exp3++. Similar results

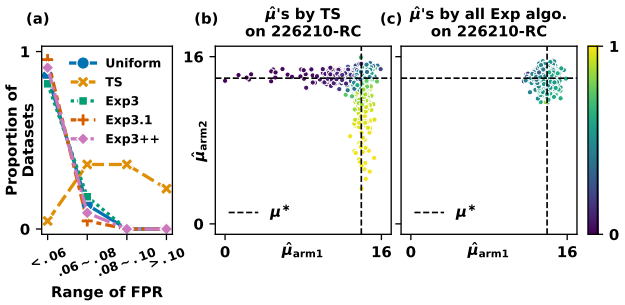

Figure 3: (a) TS allocation results in more datasets with inflated FPR than uniform and adversarial allocations. Remaining figures show the distribution of $(\hat{\mu}_{\text{arm1}}, \hat{\mu}_{\text{arm2}})$ pairs of TS (b) and of all Exp algorithms (c) on dataset 226210-RC, with hue showing $N_{\text{arm1}}/N_{\text{total}}$. Clearly, TS has trials where one arm is very badly estimated while Exp algorithms don't. Hue offers an explanation: for TS (b), $\hat{\mu}_{\text{arm1}}$ is roughly unbiased when arm 1 is pulled more often (a proxy of appearing to be better) and negatively biased when it is pulled less often (a proxy of appearing to be worse early-on); Exp algorithms (c) allocate more evenly, indicating that it explored both arms beyond initial under/over-estimation.

hold if we arbitrarily analyze the other arm instead. Exp3.1 did not have a significant negative bias with one choice of arm for analysis (coefficient: -0.0092, $p = .23$), but did for the other (coefficient: -0.0173, $p < .05$), suggesting a weak negative bias in estimated mean. This replicates results for TS from prior work [22] and shows the same underestimation can occur for adversarial bandits, albeit with much smaller magnitude of underestimation: in all cases, the MAB algorithm underestimates the mean of the arm that appears, due to random sampling, to be worse, and then samples the arm that appears to be better more often, leading the estimate of the other arm to be below its true mean. See Figure 3b and c for a comparison between TS and Exp algorithms.

Overall, these results show that experimental design using the Exp family of algorithms can collect data that meets researchers' needs better than a purely stochastic bandit algorithm, with higher power, lower FPRs, and more accurate condition measurement. These algorithms do not benefit students as much as the purely stochastic TS, but all Exp algorithms do improve on uniform allocation. Exp3++ provides a balance between Exp3.1 and Exp3, achieving 19% of the reward gain of TS, requiring no hyper-parameters, and demonstrating only a slight underestimation of arm means.

5. POWER AND UNEQUAL VARIANCE

The ASSISTments simulations demonstrate that across a range of educational experiments, the Exp algorithms performed better when measured in terms of researchers' concerns, like power and arm mean estimates, while TS attained greater benefits for students. More generally, both TS and Exp aim to optimize reward, but they do so with different assumptions about the environment, meaning that the specific characteristics of an experiment will influence how well each performs on both student- and researcher-centric measures. However, because the ASSISTments experiments do not vary systematically from one another, they are not an ideal platform for exploring the impact of *specific* experimental characteristics on the MAB algorithms' performance compared to one another and compared to uniform allocation. We thus turn to simulations with constructed datasets to explore the impact of one experimental characteristic: the relative variability of the two conditions.

Ideally, pedagogical interventions increase equity and narrow achievement gaps between students, indicated by lower variability among students who experience the intervention, but in the non-ideal case they also could widen these gaps. Interestingly, differences in condition variability affect what allocation of students is best for power: uneven allocation that places more students in the more variable condition will result in higher power than uniform allocation, at the potential cost of reward. In the simulations that follow, we examine how allocation using TS and Exp algorithms juggles power with reward differently, as well as whether Exp algorithms' adversarial assumptions and sensitivity to condition variability make them improve upon uniform allocation for measures like power that are researchers' primary concern.

5.1 Methods

5.1.1 *Two-arm scenarios*

To systemically investigate the impact of having conditions that differ in variability, we construct artificial scenarios with two arms that have normally distributed rewards. We fix the expected reward of arm 1 as 1 and that of arm 2 as 0, and vary whether the better or worse arm has a higher variance and the magnitude of the differences in variance; specifically we consider the following 19 scenarios:

$$\{(1,1)\} \cup \underbrace{(\{1\} \times \{2:10\})}_{\text{Worse arm has higher SD}} \cup \underbrace{(\{2:10\} \times \{1\})}_{\text{Better arm has higher SD}}$$

where each tuple gives the standard deviation (SD) of arm 1 followed by arm 2.

As in Section 4, rewards are interpolated into $[0, 1]$ for the MAB algorithms. Rewards are first clipped to $[-30, 31]$, where -30 is the mean of the worse arm (0) minus three times the maximum possible SD (10) and 31 is the mean of the better one (1) plus three times the maximum possible SD. We run 10000 trials for each scenario-algorithm combination, and each trial includes 250 simulated students.

5.1.2 *Data collection and analysis*

Data collection is described in Section 4.1.3. For analysis, we use Welch's test (t-test that does not assume equal variances) instead of the Brunnzer-Munzel test, as raw rewards are normally distributed and the clipping range is wide.

To determine if a statistic for data collected using the MAB algorithms differed reliably from that for data collected using uniform allocation, we use generalized linear regression, with factors for algorithm (with uniform as reference group), standard deviation of the arm with variable variance, and the interaction between the two.

5.2 Results

As shown in Figure 4, the impact of allocation method on power differed systematically based on whether the better

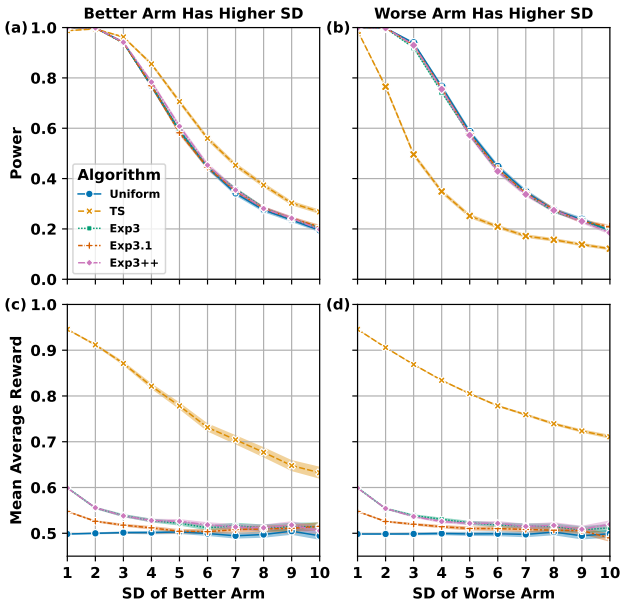

Figure 4: TS has higher power than uniform when the better arm is more variable (a), but lower power when the worse arm is more variable (b); Exp algorithms and uniform perform similarly. In all cases, MABs outperform uniform, with TS attaining the highest reward (c-d). Error bands: ± 1.96 SE.

or worse arm had higher variance. TS allocation was associated with significantly higher power than uniform allocation when the better arm was more variable (coefficient: 0.20; $p < .001$; Figure 4a), but significantly lower power when the worse arm was more variable (coefficient: -1.94, $p < .001$; Figure 4b). In contrast, the three Exp algorithms performed qualitatively similarly to uniform allocation in both scenarios, with a small but reliable decrease in power for Exp3 and Exp3.1 when the worse arm was more variable (coefficient for Exp3: -0.13, $p < .001$; Exp3.1: -0.09, $p < .01$). No other power differences were detected. While the potential of TS to increase power in some situations seems promising, researchers will not know in advance whether to expect TS to increase or decrease power, and the decreases in power from TS when the worse arm has higher variance are larger than the increases in the opposite case. Further, designing an intervention where the better arm is more variable is generally undesirable: it corresponds to a scenario in which an intervention that helps students on average also widens gaps among individual students.

As expected and shown in Figure 4c-d, TS had larger reward compared than the Exp algorithms, but all MAB algorithms resulted in reliably higher reward – i.e., benefits to students – than uniform allocation (better arm more variable: coefficient for TS: 0.47; Exp3: 0.07; Exp3.1: 0.03; Exp3++: 0.07; worse arm more variable: TS: 0.45; Exp3: 0.08; Exp3.1: 0.04; Exp3++: 0.07; all $p < .001$).

6. DISCUSSION

Pedagogical experiments are a useful tool for improving education, but allocating students to conditions uniformly at random can pose challenges, with larger subject pools leading to more students experiencing an inferior educational condition but smaller subject pools potentially decreasing the ability of researchers to differentiate conditions with certainty. Our results suggest that adversarial bandit algorithms offer a way to increase the proportion of students assigned to a better condition with limited compromises to the conclusions that can be drawn from the experimental results. Hyper-parameter free adversarial bandit algorithms like Exp3++ thus offer a researcher-friendly option for experimental design that performs well even with non-standard outcome distributions, as we saw in the ASSISTments simulations. While TS improved power when the better condition had higher variance, it also decreased power when the worse arm had higher variance. Researchers are unlikely to know which of these scenarios applies before collecting results, and this unpredictability of the impact of TS, coupled with the higher FPR and lower power on average in the real educational datasets, may make stochastic bandits less attractive to researchers.

One limitation of this work is its use of simulations rather than new experiments. Importantly, simulations are needed to measure power and FPR, but field testing with adversarial algorithms is also needed to assess their real-time feasibility as well as the existence and impact of temporal trends in outcomes. All studied algorithms run in real-time on a single CPU, but students may complete homework contemporaneously, with some assigned a condition prior to another student finishing and thus without incorporating the outcome of that student. This could be handled by batch updating [17,21], but there has been limited exploration of the impact of batching on analyzing the collected data. Temporal trends may occur in experiments if, say, higher prior knowledge students complete homework first, or multiple schools participate in an experiment at different times. Adversarial bandit algorithms should outperform stochastic ones in these situations, but empirical study is necessary.

A second limitation is our assumption that one arm is better for all students, without regard for personalization. While personalization is an exciting area for future work, incorporating personalization in bandit algorithms for education poses its own ethical conundrums, including potentially lesser outcomes for less well-represented groups (see [14]).

Future work should also examine other scenarios where bandit algorithms might increase power. Here, we focused only on conditions with differing variance, but other interesting scenarios include experiments with more conditions, with predicted outcome distributions that exhibit particular non-standard characteristics (e.g., bimodality), or with different analysis goals than detecting if two conditions differ in mean. One or more types of bandit algorithms (e.g., stochastic, adversarial, or best-arm identification [23]) may be best for each scenario; based on their weak environmental assumptions, we believe adversarial bandits may be reasonable in all of these scenarios. Work in optimal experiment design (OED; see, e.g., [24]) shows the potential of non-uniform allocation to increase the information gained from an experiment. Bandit algorithms often require less setup and a priori knowledge than OED, and thus identifying information gain benefits of these algorithms in particular settings could benefit both researchers and student participants.

7. REFERENCES

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

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

[3] A. C. Atkinson. Selecting a biased-coin design. *Statistical Science*, 29(1):144–163, 2014.

[4] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

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

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

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

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

[9] S. Finkelstein, E. Yarzebinski, C. Vaughn, A. Ogan, and J. Cassell. The effects of culturally congruent educational technologies on student achievement. In *International Conference on Artificial Intelligence in Education*, pages 493–502. Springer, 2013.

[10] N. T. Heffernan. Assistments data: gender. https://sites.google.com/site/assistmentsdata/an-explanation-on-how-to-interpret-our-data-sets/gender, 2014. [Online; accessed 21-February-2022].

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

[12] C. Kaibel and T. Biemann. Rethinking the gold standard with multi-armed bandits: Machine learning allocation algorithms for experiments. *Organizational Research Methods*, 24(1):78–103, 2021.

[13] A. S. Lan and R. G. Baraniuk. A contextual bandits framework for personalized learning action selection. In T. Barnes, M. Chi, and M. Feng, editors, *Proceedings of the Ninth International Conference on Educational Data Mining*, pages 424–429, 2016.

[14] Z. Li, L. Yee, N. Sauerberg, I. Sakson, J. J. Williams, and A. N. Rafferty. Getting too personal(ized): The importance of feature choice in online adaptive algorithms. In *Proceedings of the 13th International Conference on Educational Data Mining, EDM 2020, Fully virtual conference, July 10-13, 2020*, pages 159–170. International Educational Data Mining Society, 2020.

[15] Y.-E. Liu, T. Mandel, E. Brunskill, and Z. Popovic. Trading off scientific knowledge and user learning with multi-armed bandits. In J. Stamper, Z. Pardos, M. Mavrikis, and B. McLaren, editors, *Proceedings of the 7th International Conference on Educational Data Mining*, pages 161–168, 2014.

[16] J. D. Lomas, J. Forlizzi, N. Poonwala, N. Patel, S. Shodhan, K. Patel, K. Koedinger, and E. Brunskill. Interface design optimization as a multi-armed bandit problem. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 4142–4153, 2016.

[17] T. Mandel, Y.-E. Liu, E. Brunskill, and Z. Popović. The queue method: Handling delay, heuristics, prior data, and evaluation in bandits. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 2849–2856, 2015.

[18] P. McGuire, S. Tu, M. E. Logue, C. A. Mason, and K. Ostrow. Counterintuitive effects of online feedback in middle school math: results from a randomized controlled trial in ASSISTments. *Educational Media International*, 54(3):231–244, 2017.

[19] B. M. McLaren, T. van Gog, C. Ganoe, M. Karabinos, and D. Yaron. The efficiency of worked examples compared to erroneous examples, tutored problem solving, and problem solving in computer-based learning environments. *Computers in Human Behavior*, 55:87–99, 2016.

[20] X. Nie, X. Tian, J. Taylor, and J. Zou. Why adaptively collected data have negative bias and how to correct for it. In *International Conference on Artificial Intelligence and Statistics*, pages 1261–1269, 2018.

[21] D. Provodin, P. Gajane, M. Pechenizkiy, and M. Kaptein. The impact of batch learning in stochastic bandits. *NeurIPS 2021 Workshop on Ecological Theory of Reinforcement Learning*, 2021.

[22] A. N. Rafferty, H. Ying, and J. J. Williams. Statistical consequences of using multi-armed bandits to conduct adaptive educational experiments. *Journal of Educational Data Mining*, 11(1):47–79, 2019.

[23] D. Russo. Simple bayesian algorithms for best-arm identification. *Operations Research*, 68(6):1625–1647, 2020.

[24] E. G. Ryan, C. C. Drovandi, J. M. McGree, and A. N. Pettitt. A review of modern computational algorithms for bayesian optimal design. *International Statistical Review*, 84(1):128–154, 2016.

[25] E. M. Schwartz, E. T. Bradlow, and P. S. Fader. Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science*, 36(4):500–522, 2017.

[26] S. L. Scott. Multi-armed bandit experiments in the online service economy. *Applied Stochastic Models in Business and Industry*, 31(1):37–45, 2015.

[27] Y. Seldin and G. Lugosi. An improved parametrization and analysis of the exp3++ algorithm

for stochastic and adversarial bandits. In *Conference on Learning Theory*, pages 1743–1759. PMLR, 2017.

[28] Y. Seldin and A. Slivkins. One practical algorithm for both stochastic and adversarial bandits. In *International Conference on Machine Learning*, pages 1287–1295. PMLR, 2014.

[29] D. Selent, T. Patikorn, and N. Heffernan. ASSISTments dataset from multiple randomized controlled experiments. In *Proceedings of the Third ACM Conference on Learning at Scale*, pages 181–184. ACM, 2016.

[30] W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

[31] S. S. Villar, J. Bowden, and J. Wason. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: A review journal of the Institute of Mathematical Statistics*, 30(2):199–215, 2015.

[32] C. Walkington, V. Clinton, and A. Sparks. The effect of language modification of mathematics story problems on problem-solving in online homework. *Instructional Science*, 47(5):499–529, 2019.

[33] J. J. Williams, J. Kim, A. Rafferty, S. Maldonado, K. Z. Gajos, W. S. Lasecki, and N. Heffernan. Axis: Generating explanations at scale with learnersourcing and machine learning. In *Proceedings of the Third ACM Conference on Learning at Scale*, pages 379–388. ACM, 2016.

[34] J. J. Williams, A. N. Rafferty, A. Ang, D. Tingley, W. S. Lasecki, and J. Kim. Connecting instructors and learning scientists via collaborative dynamic experimentation. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 3012–3018. ACM, 2017.

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

[36] S. F. Williamson, P. Jacko, S. S. Villar, and T. Jaki. A bayesian adaptive design for clinical trials in rare diseases. *Computational statistics & data analysis*, 113:136–153, 2017.

[37] M. Xu, T. Qin, and T.-Y. Liu. Estimation bias in multi-armed bandit algorithms for search advertising. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26, pages 2400–2408. Curran Associates, Inc., 2013.

[38] J. Yao, E. Brunskill, W. Pan, S. Murphy, and F. Doshi-Velez. Power-constrained bandits. *arXiv preprint arXiv:2004.06230*, 2020.