Bandit assignment for educational experiments: Benefits to students versus statistical power

Anna N. Rafferty[1], Huiji Ying[1], and Joseph Jay Williams[2]

[1] Computer Science Department, Carleton College, Northfield, MN 55057 USA
[2] School of Computing, Department of Information Systems & Analytics, National University of Singapore, Singapore

Abstract. Randomized experiments can lead to improvements in educational technologies, but often require many students to experience conditions associated with inferior learning outcomes. Multi-armed bandit (MAB) algorithms can address this issue by modifying experiment designs to direct more students to more helpful conditions. Using simulations as well as modeling data from previous educational experiments, we explore the statistical impact of using MAB for experiment design, focusing on the tradeoff between acquiring statistically reliable information and the benefits to students. Results suggest that MAB experiments can lead to much higher average benefits to students than traditional experimental designs, but at least twice as many participants are needed to attain power of 0.8 and the false positive rate is doubled. Optimistic prior distributions in the MAB algorithm can mitigate the loss in power to some extent, without meaningful reductions in benefits or further increasing false positive rates.

Randomized controlled experiments are commonly used in educational technologies to address issues of interest to both curriculum designers and researchers. These experiments typically assign half of students to one version of technology components and half to another, investigating questions like whether video or text explanations will be better. This experimental design fulfills the goal of collecting as much information as possible, but is indifferent to benefits for learners: even if one condition is clearly ineffective, half of students experience it.

Multi-armed bandit (MAB) algorithms offer a potential solution for presenting more effective conditions more frequently. that considers the utility of different versions of content. MABs optimize expected *reward*, such as choosing a condition based on students' learning gains. Traditionally, MABs have been used for applications like selecting online ads [10], but they have also been used in education to discover what version of a system to give to learners [11, 12].

However, using MABs instead of traditional experiment designs raises a tension between maximizing benefits to students and the information gained about differences between conditions [6, 8]. Because MABs unevenly divide students across conditions, statistical power to detect effects can be decreased, limiting the inferences that can be drawn from experiments. Work like [6] 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, such as the effects on power.

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, and demonstrate how optimistic priors can improve power while maintaining student benefits. We then consider a case that may arise when MABs are used in activities like homework, when more (or less) able students tend to engage with activities earlier than others. We show that such biases can have profound impacts on both Type I and II error rates in experiments, with less-able students engaging first to increasing Type I error 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.

1 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 making condition assignment dependent on previous students' results. This raises the question of how MAB assignment affects statistical conclusions drawn from the data. We investigate via simulations using Thompson sampling, a MAB algorithm with logarithmic bounds on regret growth [1] that performs well in practice [4]. Thompson sampling corresponds to weighted randomization based on the expected value of the target outcome measure (reward). It maintains a distribution over rewards for each condition (based on data collected so far), and chooses a condition by sampling from each distribution and selecting the condition with highest sampled value. We expect trends in these simulations to hold for other regret-minimizing MAB algorithms; we focus on Thompson sampling as its policy is readily interpretable by behavioral researchers.

In the simulations, we varied the type of reward (experimental outcome measure) and true effect size, and examined (a) rewards per student, (b) statistical power to detect effects, and (c) how likely the direction of an effect is to be erroneously reversed ("Type S errors" [7]), and (d) the rate of incorrectly rejecting the null hypothesis (Type I error rates). We also investigated the impact of changes to the prior distributions in Thompson sampling, providing guidance for researchers when using these methods in their work.

1.1 Simulation methods

Across simulations, we varied: method of condition assignment (MAB versus uniformly at random), reward type, true effect size, and number of participants (sample size). Each set of parameters was used for 500 simulations.

Reward Models: Thompson sampling models the reward distribution for each condition. We considered both binary rewards (e.g., whether a student completes an activity) and real-valued rewards (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.

For binary rewards, the likelihood of success on each trial is Bernoulli-distributed, and a (conjugate) Beta prior is placed on each condition. For real-valued rewards, the likelihood for the reward on each trial is $\mathcal{N}(\mu, \sigma)$, where μ and σ are unknown. A Normal-Gamma prior is used for each condition, which is conjugate to the normal distribution parameterized via the mean μ and precision $\lambda = \sigma^{-1}$. In both cases, separate posterior distributions are maintained for each condition.

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 [5], and normally-distributed reward thresholds were 0.2, 0.5, and 0.8 for Cohen's d [5]. 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 and then variance set to achieve the desired effect size; means were dependent on the prior, as described below. When the conditions did not differ, the effect size was zero: binary reward simulations had condition means equal to 0.5, and normally-distributed rewards had equal condition means, with the same variances as for the non-zero effect sizes.

Sample sizes: For each effect size, we computed m, the sample size achieve 0.8 power with equally balanced conditions given Type I error rate (false positives) of 0.05. 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 $\text{Beta}(1, 1)$ prior. For normally-distributed rewards, condition means were set to -0.5 and 0.5 for non-zero effect sizes, and the prior was $\text{NG}(0, 1, 1, 1)$ (i.e., prior on the mean has mean zero). *Prior above* placed the prior mean above both conditions (binary rewards: $\text{Beta}(1.5, 0.5)$ prior; normally-distributed rewards: means -1.0 and -0.5 with $\text{NG}(0, 1, 1, 1)$ prior). *Prior below* placed the prior mean below both conditions (binary rewards: $\text{Beta}(0.5, 1.5)$ prior; normally-distributed rewards: means 0.5 and 1.0 with $\text{NG}(0, 1, 1, 1)$ prior). When effect size was zero, the same priors were used for binary rewards, and for normally-distributed rewards, the means were 0 (*prior between*), -0.5 (*prior above*), and 0.5 (*prior below*). 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.

1.2 Results

Conditions differ: When conditions have different benefits for students, the goal is to detect that the difference is reliable and assign more students to the better

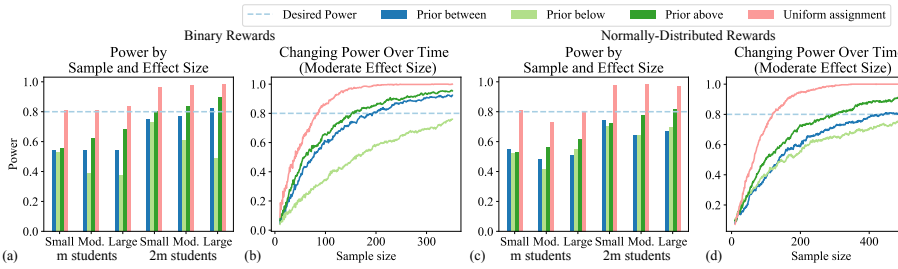

Fig. 1. Power by assignment and outcome measure. Decrease in power is similar across effect sizes. Power with m participants (expected power = 0.8), and $2m$ participants ((a) binary rewards and (c) normally-distributed rewards). Power over time, with a total of $4m$ participants ((b) binary rewards and (d) normally-distributed rewards).

Table 1. Logistic regression for predicting power (effect size $\neq 0$).

Predictor	Coefficient	t	p
Intercept	-1.04	19.1	$< .0001$
Uniform sampling	1.24	36.9	$< .0001$
Normal reward	-0.311	9.45	$< .0001$
Effect size	0.0210	1.05	0.296
Sample size	1.05	53.1	$< .0001$

Table 2. Logistic regression for predicting Type I errors (effect size = 0).

Predictor	Coefficient	t	p
Intercept	-2.47	88.8	$< .0001$
Uniform sampling	-0.700	15.5	$< .0001$
Normal reward	0.218	6.63	$< .0001$
Sample size	0.000192	11.7	$< .0001$

condition. MAB assignment without an optimistic or pessimistic prior (*prior between*) decreased power from an expected 0.80 to 0.54 for binary rewards and 0.51 for normally-distributed rewards (Figure 1). Doubling the sample size raised power closer to the desired 0.80 (0.78 and 0.69 respectively). As Figures 1(b) and (d) show, there are diminishing returns of more students: evidence for the superiority of one condition leads to assigning few students to the alternative.

For this and later analyses, we use logistic regression to test which simulation parameters impact the dependent variable (here, power); the 500 runs of each simulation provide multiple samples to test which factors have reliable impacts. Factors were assignment type, reward type, true effect size, and sample size relative to that expected to obtain 0.8 power. This analysis confirms that MAB assignment led to lower power than uniform assignment (see Table 1). Normally-distributed rewards also lowered power, driven by the MAB simulations; these rewards have a larger range than binary and thus one reward can cause a greater change in condition probabilities.

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 2). Thus, while decreased power means more participants will be necessary to detect an effect, higher rewards (i.e., more benefits to students) occur while the experiment is running.

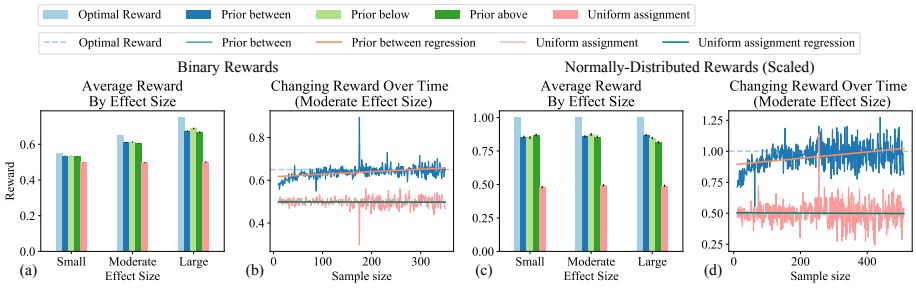

Fig. 2. Benefits to students based on sampling type, compared to assigning all students to the best condition (light blue bar): MAB assignment obtains higher rewards, with increasing reward over time. (a) and (c) show differences in average reward per step with small, moderate, and large effect sizes. Error bars represent one standard error. (b) and (d) show trend of reward growth over time with *moderate* effect size.

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.

The choice of prior in the MAB simulations impacted power (coefficient *prior below* $= -0.69$, $t(35994) = 23.5$, $p < .0001$; coefficient *prior between* $= -0.32$, $t(35994) = 10.9$, $p < .0001$). 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 69% of simulations, compared to 62% for *prior between*, and 55% for *prior below*. This is partially driven by very unbalanced simulations: 7% 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, providing better evidence to estimate the means. This same behavior increases Type S errors slightly for the less optimistic *prior between* ($t(35994) = 2.66$, $p < .01$) and *prior below* ($t(35994) = 3.48$, $p < .001$). However, Type S errors are still extremely rare (0.21% for the prior with the highest rate). Average reward is only modestly decreased with more optimistic priors (Figure 2).

Conditions do not differ: We first confirmed via linear regression that using assignment type did not impact rewards, which was expected given equally effective conditions. We then examined the false significance rate in simulations with no underlying differences between conditions (Type I errors). We aggregrated across prior types for MAB assignment because these did not have a statistically significant impact.[3] As shown in Figure 3, MAB assignment increased the Type

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

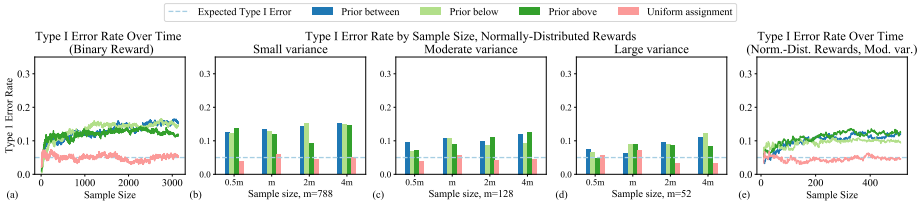

Fig. 3. MAB assignment increases Type I error rate over uniform assignment. (a)/(e) Type I error rates by sample size for binary (a) and normally-distributed rewards (e). (b)-(d) Type I error rates for normally-distributed rewards by variance of conditions.

I error rate: 5.2% of simulations using uniform assignment found a significant difference between conditions, while 9.7% of simulations using MAB assignment found a difference. 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 (see Table 2). There was a slightly higher Type I error rate for normally-distributed rewards than for binary rewards, 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.

In conclusion, simulations when conditions do and do not differ reveal that optimistic priors can help mitigate lowered power from MAB assignment without increasing Type I error rates. Benefits to students approach that of the better condition, which may justify doubling sample sizes to attain sufficient power.

2 MAB-assignment in educational experiments

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 [9].

2.1 Methods

Each of the 22 experiments included three outcomes [9]: 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). 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

[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 ($p < .1$).

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.

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.

2.2 Results

Parameter simulations: As shown in Figure 4(a), MAB assignment 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 $d = 0.70$). Figure 4(b) confirms the simulation findings as MAB assignment decreases power for the *completed* measure. Counterintuitively, MAB assignment increases power for the *problem count* measure. This is due to the high variability for *problem count*, since 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.
Outcome simulations: As shown in Figure 4(c), 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 rewards 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. Because MAB assignment adapts to the reward signal, it improved reward even in some cases where 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

[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.

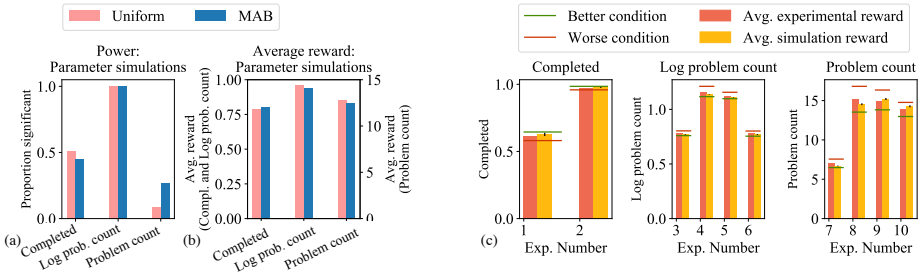

Fig. 4. Results based on educational experiments. (a-b) Power (a) and reward or cost per step (b) averaged across the *parameter* simulations. In the reward graph, higher bars are better for *completed*; lower bars (i.e., lower cost) are better for the other measures. (c) Observed rewards in ten previously conducted experiments, and average reward using MAB assignment. "Better" and "worse" conditions are the average observed values for each condition in the experiment. Error bars indicate one standard error.

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.

3 Discussion

Randomized experiments can identify more effective educational strategies, but there are practical and ethical concerns about giving students 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 a common MAB algorithm (Thompson Sampling) versus traditional uniform random assignment. Our simulations demonstrate that MAB assignment can benefit students. However, this comes at a cost of reliability: while reversing the sign of an effect is rare, higher rates of Type I and Type II errors occurred, showing that the uneven distribution of students across conditions lowers power and that the measurement errors from MAB assignment (documented in, e.g., [6]) can lead

to incorrect rejections of the null hypothesis. While an optimistic prior lessened the impact on power, sample sizes would still need to be doubled to achieve 0.8 power. 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 motivation for educational experimentation with MAB assignment was to encourage teachers to allow experiments in their classes as their students would directly benefit. Our results suggest that MAB experiments could serve 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 methodology and goals of a program of research, but has the potential to address the needs of both teachers and researchers.

There are several limitations to this work. First, we focused only on experiments with two conditions. 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.

Second, we focused on a regret-minimizing algorithm, rather than MAB algorithms for identifying the best condition [2] or trading off rewards and measurement accuracy [6]. 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.

Finally, when considering biased patterns of students engagement, we did not examine detecting or correcting for such bias. Explicitly modeling temporal bias or using MAB algorithms that do not assume stationary rewards [3] could ameliorate some of the negative effects on power and Type I errors, and exploring these approaches in educational settings is an interesting area for future work.

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 in predictable ways that researchers should take into account when employing these methods.

References

1. Agrawal, S., Goyal, N.: Analysis of thompson sampling for the multi-armed bandit problem. In: Mannor, S., Srebro, N., Williamson, R.C. (eds.) Proceedings of the 25th Annual Conference on Learning Theory. vol. 23, pp. 39.1–39.26. PMLR, Edinburgh, Scotland (2012)
2. Audibert, J.Y., Bubeck, S.: Best arm identification in multi-armed bandits. In: COLT-23th Conference on Learning Theory-2010 (2010)
3. Besbes, O., Gur, Y., Zeevi, A.: Stochastic multi-armed-bandit problem with non-stationary rewards. In: Advances in neural information processing systems. pp. 199–207 (2014)
4. Chapelle, O., Li, L.: An empirical evaluation of thompson sampling. In: Advances in neural information processing systems. pp. 2249–2257 (2011)
5. Cohen, J.: Statistical power analysis for the behavioral sciences. Routledge, 2 edn. (1988)
6. Erraqabi, A., Lazaric, A., Valko, M., Brunskill, E., Liu, Y.E.: Trading off rewards and errors in multi-armed bandits. In: International Conference on Artificial Intelligence and Statistics (2017)
7. Gelman, A., Carlin, J.: Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. Perspectives on Psychological Science 9(6), 641–651 (2014)
8. Liu, Y.E., Mandel, T., Brunskill, E., Popovic, Z.: Trading off scientific knowledge and user learning with multi-armed bandits. In: Educational Data Mining 2014 (2014)
9. Selent, D., Patikorn, T., Heffernan, N.: Assistments dataset from multiple randomized controlled experiments. In: Proceedings of the Third (2016) ACM Conference on Learning@ Scale. pp. 181–184. ACM (2016)
10. Tang, L., Rosales, R., Singh, A., Agarwal, D.: Automatic ad format selection via contextual bandits. In: Proceedings of the 22nd ACM international conference on Conference on information & knowledge management. pp. 1587–1594. ACM (2013)
11. Whitehill, J., Seltzer, M.: A crowdsourcing approach to collecting tutorial videos–Toward personalized learning-at-scale. In: Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale. pp. 157–160. ACM (2017)
12. Williams, J.J., Kim, J., Rafferty, A., Maldonado, S., Gajos, K.Z., Lasecki, W.S., Heffernan, N.: Axis: Generating explanations at scale with learnersourcing and machine learning. In: Proceedings of the Third (2016) ACM Conference on Learning@ Scale. pp. 379–388. ACM (2016)