Multiple Choice vs. Fill-In Problems: The Trade-off Between Scalability and Learning

Ashish Gurung
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
agurung@andrew.cmu.edu

Kirk Vanacore
Worcester Polytechnic Institute
Worcester, Massachusetts, USA
kpvanacore@wpi.edu

Andrew A. McReynolds
Worcester Polytechnic Institute
Worcester, Massachusetts, USA
aamcreynolds@wpi.edu

Korinn S. Ostrow
Worcester Polytechnic Institute
Worcester, Massachusetts, USA
ksostrow@wpi.edu

Eamon S. Worden
Worcester Polytechnic Institute
Worcester, Massachusetts, USA
elworden@wpi.edu

Adam C. Sales
Worcester Polytechnic Institute
Worcester, Massachusetts, USA
asales@wpi.edu

Neil T. Heffernan
Worcester Polytechnic Institute
Worcester, Massachusetts, USA
nth@wpi.edu

ABSTRACT

Learning experience designers consistently balance the trade-off between open and close-ended activities. The growth and scalability of Computer Based Learning Platforms (CBLPs) have only magnified the importance of these design trade-offs. CBLPs often utilize close-ended activities (i.e. Multiple-Choice Questions [MCQs]) due to feasibility constraints associated with the use of open-ended activities. MCQs offer certain affordances, such as immediate grading and the use of distractors, setting them apart from open-ended activities. Our current study examines the effectiveness of Fill-In problems as an alternative to MCQs for middle school mathematics. We report on a randomized study conducted from 2017 to 2022, with a total of 6,768 students from middle schools across the US. We observe that, on average, Fill-In problems lead to better post-test performance than MCQs; albeit deeper explorations indicate differences between the two design paradigms to be more nuanced. We find evidence that students with higher math knowledge benefit more from Fill-In problems than those with lower math knowledge.

CCS CONCEPTS

• **Applied computing** → **Computer-assisted instruction**; *Interactive learning environments.*

KEYWORDS

Causal Inference, Learning Experience Design, Multiple Choice Questions, Fill-In Problems, Learning Outcomes

LAK '24, March 18–22, 2024, Kyoto, Japan

ACM ISBN 979-8-4007-1618-8/24/03.
https://doi.org/10.1145/3636555.3636908

ACM Reference Format:
Ashish Gurung, Kirk Vanacore, Andrew A. McReynolds, Korinn S. Ostrow, Eamon S. Worden, Adam C. Sales, and Neil T. Heffernan. 2024. Multiple Choice vs. Fill-In Problems: The Trade-off Between Scalability and Learning. In *The 14th Learning Analytics and Knowledge Conference (LAK '24), March 18–22, 2024, Kyoto, Japan.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3636555.3636908

1 INTRODUCTION

The rapid growth in technology and the ability to produce educational material accessible by innumerable learners has led to the development and adoption of Computer Based Learning Platforms (CBLPs) across educational sectors. With access to the internet, learners across the world use CBLPs in the form of Massive Open Online Courses (MOOCs), Learning Management Systems (LMSs), and standalone online learning platforms. The past two decades have seen a drastic rise in the implementation and utilization of online educational materials [17, 23]. With this growth, Learning Experience (LX) designers are tasked with the critical responsibility of ensuring that learning materials are not only effective but also scalable. A key difficulty faced by LX designers is finding a balance between the educational merits of different problem types and the feasibility of employing these problem types effectively at scale. This balance is crucial for maximizing the impact and reach of these CBLPs.

Broadly, the instructional strategies leveraged by LX designers in terms of the problem types can be classified into two categories: open-ended and closed-ended problems. Closed-ended problems, such as Multiple Choice Questions (MCQs), 'Check all that Apply', and 'Arrange in the Correct Order', lend themselves more easily to the integration of automated grading and instant feedback. These features, along with on-demand help, can significantly enhance the learning experience for users. This approach allows for a more scalable and efficient way of delivering instruction, particularly in educational settings, with automation playing a key role in reducing the demand on instructors' time and resources [4]. Alternatively,

incorporating automated grading and instant feedback in open-ended problems such as short answer questions, essays, and fill-in problems is more challenging. While closed-ended problems are more suitable for automation compared to open-ended problems, researchers have raised concerns regarding their use [12, 32, 49]. They point out that closed-ended problems can be susceptible to recognition and guessing, which can lead to shallow learning. In contrast, open-ended problems are often viewed as more rigorous and allow the instructor to infer the learners' understanding and comprehension of the topic from their answers [12, 49]. While rigor and thoroughness are highly desirable in educational settings, it is also important to acknowledge that the higher level of difficulty and the demand for rigorous engagement can be cognitively taxing on the learners. This strain often stems from the need to exercise recall over recognition, and generation over selection, which inherently requires higher cognitive effort.

Research into the comparative effectiveness of close-ended and open-ended activities in facilitating learning is somewhat mixed. Some studies have found a preference for traditional open response problems (ORP) over MCQs [1, 12, 32], while others underscore the merits of MCQs [16, 48, 50]. Furthermore, others suggest there's little to no difference between the two formats [30, 31, 44, 49]. Amidst this backdrop of varied and sometimes conflicting findings when comparing traditional ORPs to MCQs, Fill-In problems emerge as an intriguing point of discussion. Fill-In problems share several characteristics with MCQs, including the benefits of automated grading, immediate feedback, and availability of on-demand help. Furthermore, a distinct and linear relationship exists between MCQs and Fill-In problems, enabling the straightforward conversion of one format to the other, thus offering versatility in the design of learning activities.

In this paper, we investigate the application of Fill-In problems as an alternative to MCQs in mastery-based activities. To this end, we conducted an *in-vivo* randomized study aimed at exploring the relative difficulty of utilizing Fill-In problems as an alternative to MCQs. We then explore the influence of these problem types on learners' performance in mastery-based activities and a post-test with a more complex transfer task upon acquiring mastery. Finally, we explore the potential heterogeneity in the effectiveness of the problem types across learners with varying mathematical prior performances. Specifically, we explore the following research questions:

(1) Is there a difference in the difficulty between equivalent MCQ and Fill-In problems?
(2) Does the use of different problem types (MCQ vs. Fill-In) on mastery-based activities impact students' learning?
(3) How does the effectiveness of problem type vary among learners with differing levels of mathematical proficiency?

2 PRIOR WORKS

2.1 MCQs & Fill-In Problems

Over the years, various prior research has explored the efficacy of utilizing MCQs over ORPs and found mixed results, with some finding MCQs to be more beneficial [16, 48, 50], others finding ORPs to be more beneficial [1, 12, 32], and others finding little to no difference between the two problem types [30, 31, 44, 49]. However, prior exploration regarding the feasibility of the two problem types has shown ORPs to be more costly towards instructor resources than MCQs [4]. Beyond their usage, it is also crucial to acknowledge that other contextual factors can influence the use of one problem type over the other, as each has its unique advantages and disadvantages. For example, MCQs can be particularly beneficial in assessing large cohorts of students *en masse*(*i.e.* SAT, TOEFL) [26, 35, 38]. On the other hand, ORPs have been shown to provide a more accurate assessment of learners' understanding and comprehension in various STEM-related subjects compared to MCQs [16, 48]. In fact, prior studies have reported on inflation of grades when utilizing MCQs over ORPs in STEM-related subjects [16, 48].

While MCQs are widely used in CBLPs, researchers have expressed concerns regarding their usage. MCQs can be susceptible to synthesis, guessing, and recognition due to the use of distractors[1] resulting in shallow learning [10, 19, 35]. Consequently, there have been concerns regarding the reliability and validity of the use of MCQs when inferring learners' knowledge and ability [14]. In particular, the presence of distractors in MCQs can inadvertently trigger learners' recall of topics they are attempting to work on, thereby diminishing the effectiveness of MCQs in fostering more rigorous learning and developing a deeper understanding of the topic [42]. Furthermore, MCQs, due to their design, are relatively less conducive to fostering creative thinking and idea generation [9], which are crucial skills for comprehensive learning. In contrast, Open Response Problems (ORPs) inherently require students to demonstrate higher-level thinking and reasoning for each problem, thereby eliminating the guessing element commonly associated with MCQs [35].

2.2 Mastery-Based Learning Activities

In recent decades mastery-base learning activities have become a common pedagogical technique, especially in CBLPs. Rather than assuming learning upon completion of certain activities associated with the material, mastery-based learning requires learners to demonstrate knowledge and understanding of the concepts before progressing to the next topic [8]. Mastery-based learning approaches have shown to reduce variance in student aptitude [2, 29], increase long-term retention of knowledge [29], change student attitude towards content [2, 29], and increase self-belief [2, 18].

One of the primary features of mastery-based learning is to provide students with the ability to practice the skills that allows the teachers to assess their students' abilities while facilitating learning opportunities. CBLPs, by design, have an advantage when implementing mastery-based assignments, as the activity can adapt to the student's performance. Various CBLPs have explored the implementation of mastery-based assignments using various approaches. While some platforms, such as Khan Academy [28, 34], and ASSISTments [20], have explored using an arbitrary threshold of N-Consecutive Correct Responses (N-CCR), others have relied on more precise measures of mastery using Knowledge Tracing (KT) models [13]. KT models predict student performance in future

[1]Distractors, also referred to as "Lures" in some academic settings, are incorrect answers in a multiple-choice question designed to mislead students away from the correct answer by providing false information.

problems by leveraging their past performance on similar or related skills. Both N-CCR and KT approaches have their merits and flaws; N-CCR is more explainable and interpretable by teachers, whereas KT models are harder to understand for the teachers but are more accurate at estimating learner mastery. While a heuristic of N-CCR could be considered rather simplistic, Kelly et al. (2015) [25] reported that a N-CCR design, with N = 3, has comparable performance in estimating mastery to more sophisticated KT models. Additionally, Prihar et al. (2022) [36] have reported on the benefits of using N = 2, 4, and 5 as a threshold and found N = 3 to be an optimal threshold. While a simple N-CCR design is easy to implement [22, 25] and interpret, some have expressed concerns regarding the risks of inequitable outcomes due to the use of N-CCR design's assumption across students with different learning rates [15]. Although concerns about the use of N-CCR and its potential impact on creating inequitable outcomes are significant, the study by Koeginder et al. (2023) [27], which reports a surprising consistency in students' learning rates under ideal conditions suggests a possible avenue to both mitigate the concerns related to inequity stemming from varying learning rates and an opportunity to revisit the risk of inequity in outcomes on mastery-based activities due to the methodology utilized in estimating mastery.

3 CURRENT STUDY

3.1 Experimental Design

The current study comparing MCQs and Fill-In problems was conducted using ASSISTments [20], a CBLP popular among middle school math teachers in the United States. In this experiment, we developed two mastery-based activities focused on the mathematical concepts of 'Greatest Common Factor' (GCF) and 'Evaluating Expressions' (EE). These activities were designed in accordance with the Common Core State Standards [33], with the GCF activity developed using the grade 6 curriculum and the EE activity developed using the grade 7 curriculum.

As illustrated in Figure 1, each problem set in our study included a mastery learning component followed by a post-test. The students are randomized to one of two problem types in the mastery learning components: MCQs or Fill-In problems. A N-CCR design with N=3 for the mastery-based activity is utilized in both conditions to estimate mastery, i.e., students need to correctly answer 3 problems in a row to demonstrate their mastery of the content. If a student is incorrect on their first attempt or asks for hints, the consecutive correctness counter is reset to zero. During the assignment, students have the option to request up to three hints, with the bottom-out hint giving away the answer to the problem. Additionally, the system also imposes a daily limit of 10 problems per condition. However, if a student correctly answers the 9th or 10th problem, they are allowed to attempt up to 11 or 12 problems, respectively, to demonstrate mastery. Students unable to demonstrate mastery within the first 10 problems are required to wait until the following day to continue with the activity.

Upon demonstrating mastery, students are asked to take a two-problem post-test. These problems are transfer items on the same topic as the experiment. These items required that students to apply the mastered knowledge component to a relatively more complex problem on the post-test. Examples of the problems in the experiment (MCQs vs. Fill-In problems) and the post-test are illustrated in Figure 2. As demonstrated in Figure 2, post-test problems are relatively more complex in comparison to the problems in the mastery learning component. The problem complexity was increased on the post-test by increasing the dimensionality of the problem from 2 to 3 and using a more complex sentence structure on the problem. The objective here is to assess the performance of the students who demonstrated mastery on a more complex transfer task as a representation of their learning [47]. We chose to utilize Fill-In questions as they align more closely with the problem type that would be utilized in a traditional experimental setup where the post-test would likely be conducted using a traditional paper and pencil approach to assess the students' learning on the transfer item.

3.2 Description of Dataset

The data was collected across five school years in the United States (2017-18, 2018-19, 2019-20, 2020-21, 2021-22). During this time, the assignments were made available to middle school teachers who use ASSISTments as an instructional tool by assigning mastery-based activities to their students as part of their lessons. During our study, 192 teachers assigned the two problem sets to 383 classes. A total of 6774 students participated in the experiment. A small number of students, 20, worked on both problem sets. In such instances, we only included the student data from the first participation and dropped the other records.

In addition to data on student performance, hint usage, and time to first attempt per problem in the mastery-based components of the experiment, the student performance on the post-test items and the average prior percent correctness across all problems the students worked on the CBLP prior to participating in the experiment (prior performance) were also calculated.

3.3 Descriptive Statistics

The descriptive statistics on student behavior in the mastery learning component are presented in Table 1. While there was no significant difference in the average number of problems taken to reach mastery between the MCQ and Fill-In conditions, other behavioral differences were notable. Specifically, students in the Fill-In condition, on average, accessed more hints and took more time before submitting their first responses compared to their counterparts in the MCQ condition. Intriguingly, despite every incorrect attempt and hint request resulting in a loss of 33% partial credit, students in the Fill-In group achieved a higher average score on the mastery components. This suggests that while students in the MCQ condition likely made more attempts than those in the Fill-In group, the students in the Fill-In condition were able to recognize they needed help, request it, and effectively utilize it to the problem.

3.4 Analysis Plan

To address our research questions, we conducted three analyses of the experimental data. The first (Analysis 1: Section 4) addresses the differences in problem difficulty caused by problem type within the mastery learning activity and explores students' performance patterns within each activity. Next (Analysis 2: Section 5), estimates

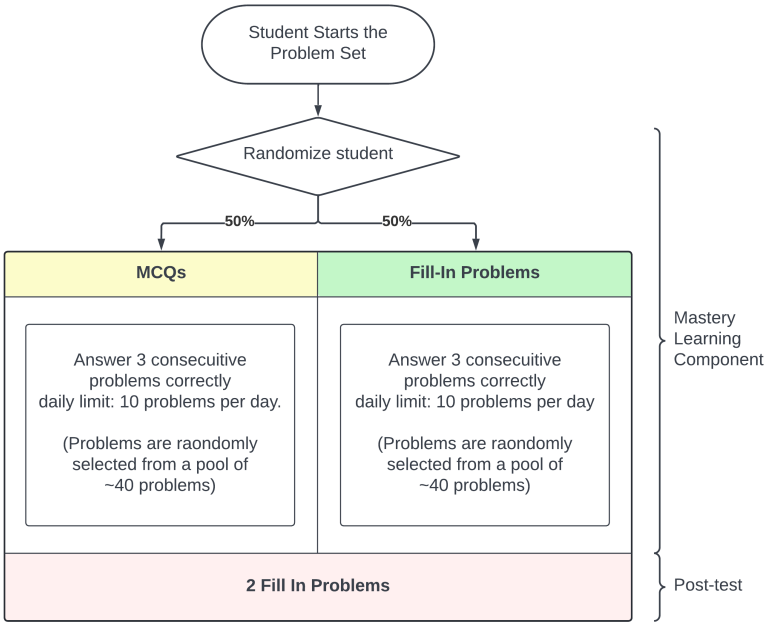

Figure 1: The experiment has two conditions where students are assigned a mastery-based assignment with MCQs or Fill-In problems. Upon acquiring mastery of the content, students are asked to answer two problems in the post-test to examine the student's performance on more complex transfer tasks.

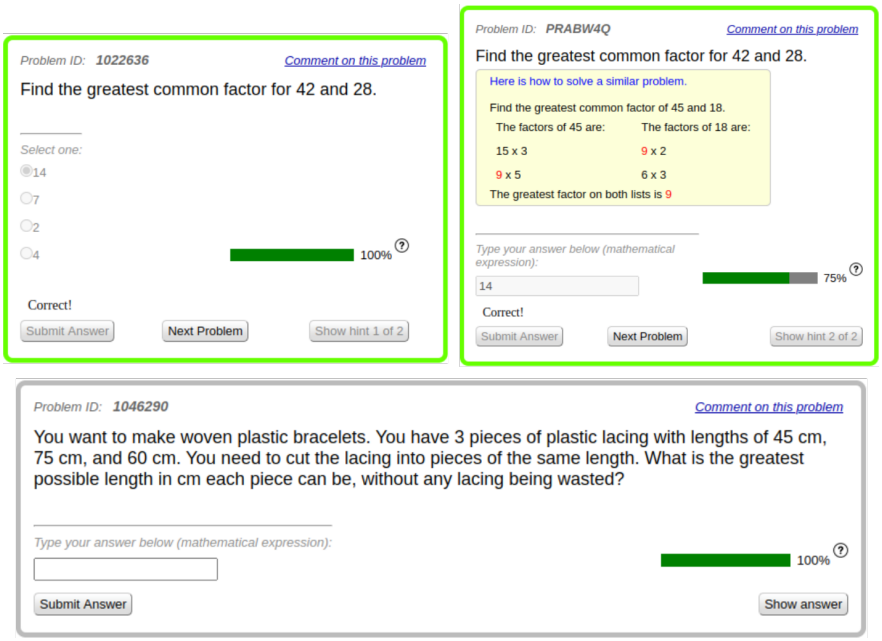

Figure 2: The equivalent MCQ and Fill-In Problems from "Problem Set 1: Greatest Common Factor" (Marked in Green). The relatively more complex transfer task in the post-test (Marked in gray).

the effects of the problem type on students learning, as represented by their performance on the post-test. Finally (Analysis 3: Section 6), evaluates whether the effect of problem type varies based on

Table 1: Descriptive Statistics exploring the total problem to mastery, total hint usage, time to the first response by the students, and the average score on the problems in the mastery learning component across conditions.

	All		Fill-In		MCQ		T-Test	
	Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*	*t-Statistic*	*p*
Average problems to mastery	4.94	2.99	4.92	2.81	4.97	3.16	0.52	0.603
Average hints access	0.58	1.88	1.04	2.55	0.15	0.56	-16.63	< 0.001
Average time to first response (sec)	61.32	457.61	74.61	417.29	48.76	492.42	-1.99	0.046
Average score on mastery component	86.21	16.98	88.80	13.81	83.77	19.15	-10.60	< 0.001

students' prior mathematical proficiency. The following three sections contain detailed information regarding the methodology and the results of these analyses.

4 ANALYSIS 1: ASSESSING THE DIFFERENCES IN DIFFICULTY BETWEEN MCQS AND FILL-IN PROBLEMS

The students were randomized into Fill-In and MCQ conditions with equivalent problems across conditions in the mastery component, i.e., the MCQ and Fill-In problems had the same problem body and answers. The difficulty of the problem type can be evaluated by comparing student performance across conditions. This analysis allows us to understand whether different problem types can influence student performance during mastery-based activities. Furthermore, estimating differences in difficulty can help contextualize potential differences in performance and learning outcomes across conditions.

4.1 Methods

The relative difficulty between Fill-In and MCQs was estimated using linear regression models with robust standard errors using the *estimator* [39] package in R [37]. First, we ran a model only utilizing student data from the first problem they attempted in the mastery learning component. This method allows us to isolate the difference in performance caused by problem types from any potential learning or attrition that could occur as the students work through the mastery component.

Equation (1) represents the difficulty estimation model. Let Y_{ij} indicate whether student i was correct on the first attempt of problem j. Let $Fill\text{-}In_i$ indicate whether student i was randomized to receive Fill-In problems, and P be an indicator for each problem the students attempted. Since assignment to problems and problem type were randomized and the problems j were equivalent across conditions, the effect of the Fill-In condition (β_1) is an unbiased causal effect. Note that if a student saw a problem but did not submit a response, we considered their responses to be incorrect. This step ensured that differences in dropout rates did not bias the results.

$$Y_{ij} = \beta_0 + \beta_1 Fill\text{-}In_i + \sum \beta_j P_j + e_{ij} \quad (1)$$

To examine how students performed on subsequent problems in the mastery learning component, we reran this analysis for the first 10 problems the students can attempt before reaching the daily limit without exhibiting mastery. Although these are no longer unbiased estimates of causal effects of problem types–due to the potential spillover effects from learning on previous problems and the differences in samples due to acquisition of mastery and attrition rates across conditions–the differences are still informative of students' learning experiences.

4.2 Results

Students performance on the first problem in the mastery learning component of the experiment differed significantly based on their treatment assignment such that students in the MCQ condition outperformed those in the Fill-In condition by an estimated five percentage points (β_1 = -0.05, SE = 0.012, $p < 0.0001$). This coefficient is an unbiased estimate of the difference in difficulty caused by problem type.

Figure 3 displays the average performance (lines) and samples (shading) by condition across the first ten problems in the mastery learning component of the experiment. Notably, for the first problem, students in the MCQ condition perform better than their peers in the Fill-In condition. However, this difference dissipates for the subsequent two problems before reversing, with the students in the Fill-In condition outperforming their peers in the MCQ condition. This finding suggests that the difficulty experienced early on by the students in the Fill-In condition potentially benefited the students later in the activity. Nevertheless, it is important to acknowledge a potential confound to this explanation as the samples may differ across conditions on problem sequences greater than one. This possibility of potential confound is illustrated in Figure 3, using the difference in the shaded area (*e.g.,* the percent of students within each condition who started the problem in that problem sequence) on the second and third problem before the students could master the knowledge component. Further exploration of this potential confound is reported in Section 5.2.

5 ANALYSIS 2: IMPACT OF FILL-IN PROBLEMS AND MCQS ON STUDENT LEARNING

Analysis 1 (Section 4) shows that Fill-In problems are more difficult than MCQs. This finding suggests that students tackling Fill-In problems exerted more effort to achieve mastery. This heightened challenge might obscure their recognition of substantial learning progress and potentially lead to negative emotions. Such feelings could result in unproductive behaviors, including gaming [3], wheel-spinning [6], or even dropping out of the activity entirely. Thus,

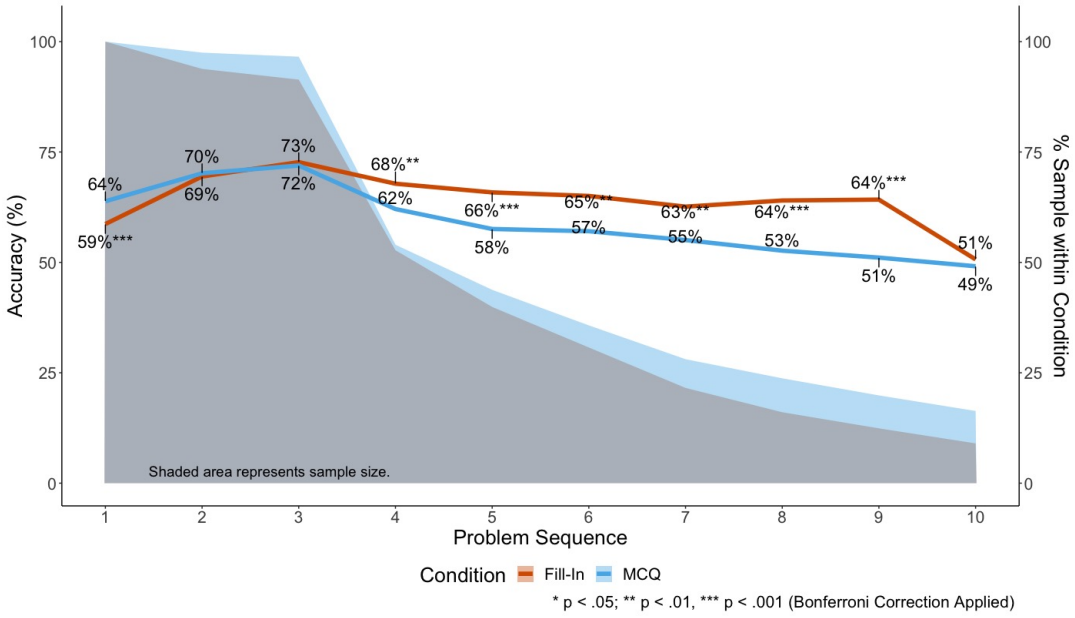

Figure 3: Accuracy During Mastery Component by Problem Type and Problem Sequence

The lines represent the average accuracy of students within the condition. Students in the Fill-In condition have lower accuracy when starting the problem set, but this trend switches after the third problem. The shading represents the percent of original samples within each condition. Not that slightly more students master or attrit in the Fill-In than the MCQ condition early on. This difference is explored further in Section 5.2.

there is the potential that the increased difficulty caused by Fill-In responses may have an adverse effect on student learning.

However, Analysis 1 also showed that students in the Fill-In condition outperformed those in the MCQ condition later in the masterly learning activity. Conversely, MCQs may induce students to engage in shallow learning, such as employing educated guesses and deducing answers, effectively recognizing and synthesizing the information. They could potentially gain only perfunctory mastery of the concept by deliberating over the provided choices in the MCQs instead of taking full advantage of the learning opportunities. Thus, it is unclear which problem type is more conducive to learning based on differences in problem difficulty alone.

In Analysis 2, we evaluate the impact of problem type (MCQ vs Fill-In) within mastery-based activities on student learning. To measure the effectiveness of each approach, we analyze students' performance on the post-test problems. As explained in Section 3.1, these post-tests are designed to assess how well students can apply their recently mastered knowledge to more complex problems within the same topic (*i.e.* transfer problems).

5.1 Methods

Evaluating the impact of problem type on students' learning requires two steps. First, since only the students who mastered the knowledge component took the post-test, and some students who mastered the knowledge did not start or complete the post-test, we must ensure that this attrition does not bias our outcomes. This assessment is critical given that Fill-In problems, being more difficult than their MCQ counterparts, might prompt students with lower knowledge levels to attrit. Once we establish whether attrition is balanced across conditions, we can estimate the effects of the problem types on student learning as measured by their performance on the post-test.

To evaluate whether attrition was balanced across conditions, we employed two tests. First, we compared the difference in attrition rates between conditions to thresholds established by the U.S. Institute of Education Sciences (IES) [21]. Next, we conducted a Chi-Square test to estimate whether the attrition differed significantly between conditions.

To evaluate whether students are more likely to learn while working on Fill-In problems than MCQs, we estimated a mixed-effect logistic regression model using the `glmer` package in R [5]. We regress indicators of whether the students got each individual post-test problem correct on the first attempt on a binary indicator for the Fill-In condition. We use this method because averaging the post-test problem together would not have created a continuous variable, as there were only two post-test problems for each problem set. Therefore, a linear regression likely has a poor model fit. Using a logistic regression model, we can treat each post-test problem individually – this also allows us to include students even if they did not complete both post-test problems. We include random interprets for post-test problems to account for differences in problem difficulty and students because students completed multiple post-test problems. We also include random intercepts for the students' classes because their classroom context could influence their learning behaviors, and students are often grouped within classes with students of similar abilities.

For any given post-test question j completed by student i, the model for the likelihood of correctness is represented by Equation (2) where γ_0 is the fixed intercept, μ_i is the random intercept for each student, μ_c is the random intercept for each student's class, and μ_j is the random intercepts for each post-test problem. Let $Fill\text{-}In_i$ be a binary indicator for whether a student is in the Fill-In condition. The coefficient, γ_1, for the Fill-In problems, represents the difference in likelihood of the students in the Fill-In condition answering the post-test items correctly in comparison to the students in the MCQs condition. Because assignment to condition was random, γ_1 is the causal effect of Fill-In problems on mastery.

$$logit(\text{Student } i \text{ Gets Post-Test Problem } j \text{ Correct}) = \gamma_0 + \gamma_1 \text{Fill-In}_i + \mu_i + \mu_c + \mu_j \quad (2)$$

5.2 Attrition

Table 2 details the experiment's attrition rates and the balance test statistics. The overall attrition rate was 27.07%. Attrition occurred at three distinct levels: Firstly, 18.02% of students did not demonstrate mastery in the learning component and thus were unable to take the post-test. Secondly, 9.91% of students achieved mastery but did not commence the post-test. Thirdly, a subset of students only completed one problem of the post-test. These students were not excluded from our analysis and are not reflected in the overall attrition figure. Based on the results of both the IES threshold and the Chi-squared test, suggesting that attrition was balanced across conditions at all levels. However the differences in mastery rates between the conditions were only marginally non-significant (p = 0.006), thus we choice to incorporate a robustness check of our effects estimation presented below.

5.3 Results

Table 3 presents the parameters of the model used to estimate the effect of problem type on post-test performance. There was a positive causal effect of Fill-In problems on student performance on transfer tasks. Specifically, students who engaged with Fill-In problem sets were significantly more likely to provide correct responses in the post-test compared to those who worked through MCQs (γ_1 = 0.23, SE = 0.06, $p > 0.001$). Putting this on the probability scale, students in the MCQ condition had a 27% probability of getting either of the transfer problems correct, whereas students in the Fill-In condition had a 31% probability of getting the transfer item correct. Notably, this higher likelihood of correctly answering post-test problems persisted even after adjusting for the variance attributable to individual students, their respective classes, and the problems themselves.

In our analysis, we also examined the variances of μ, τ, to evaluate the variance in post-test performance attributed to students and their classes. Over one-third, 34% of the variance was associated with the random intercepts. A substantial portion of the performance variance was associated with individual students (τ_i = .83; 17% of the variance). However, the class environment also played a substantial role, accounting for a considerable proportion of the variance (τ_c = .60; 12% of the variance). This finding indicates the importance of the student's learning environment and peer group in their performance. As part of our analysis of the potential heterogeneity in outcome across different prior knowledge among students, we also delve deeper into the impact of the learning environment on student outcomes in the upcoming section, Section 6.

Notably, this analysis included only post-test problems that students attempted, thus excluding students who did not master the knowledge component and those who did not start the post-test. Furthermore, some students only completed one problem on the post-test. Although we showed that the differences in attrition across conditions were not significant, it is still possible that they biased our outcomes. Thus, we reran the analysis, coding all of the students who did not master or attrited in any way with a zero for each post-test problem. The results of this robustness check model revealed no difference in significance or magnitude of the causal effect.

6 ANALYSIS 3: HETEROGENEITY IN THE IMPACT OF PROBLEM TYPES DIFFERENT LEVELS OF PRIOR PERFORMANCE

Finally, to address our last research question of whether the effect of problem type problems on learning varies based on students' prior ability, we ran one final analysis. This allows us to assess a potential nuance of how Fill-In problems benefit student learning. As detailed in Section 4, Fill-In problems were more difficult than the equivalent MCQs. As such, students of varying mathematical proficiency may derive different benefits from each problem type. It's plausible that higher-knowledge students potentially benefit more from the critical thinking, retrieval, and recall required to solve the Fill-In problems, which lack the options provided in MCQs. On the other hand, lower-knowledge students might find the availability of the options in MCQs beneficial in developing intuition and learning the concept, as distractors can highlight potential misconceptions and gaps in knowledge. In this section, we explore the potential heterogeneity in the benefits of the different problem types across students with different prior performances.

6.1 Methods

To explore whether the effect of problem type varies by prior performance, we added an interaction between students' prior performance and their experimental condition as shown in Equation 3. The prior performance was calculated as an average of the student's scores from all problems they completed prior to prior to participating in the experiment. In our initial sample, 1,643 students had completed at least ten problems prior to participating in the experiment, and the rest were excluded from the current analysis. This exclusion helps ensure accuracy in inferring the students' mathematical ability, as a limited number of completed problems might not reliably indicate their ability. This exclusion was proportionally distributed across both Fill-In and MCQ conditions–19.96% and 19.72% of students, respectively–ensuring no bias in our estimates due to imbalances in condition-specific sample sizes. In the filtered sample, prior performance scores ranged from 14.29% to 100% with a mean of 70.05% and a deviation of 15.02%. The model standardized the prior performance for better interpretability by z-scoring the prior performance scores.

Table 2: Attrition analysis at four-levels: overall, did not acquire mastery, did not start post-test and did not complete post-test.

	All	*Fill-In*	*MCQ*	$\|Difference\|$	*I*ES Threshold	χ^2	*p*
Overall attrition	27.07%	27.02%	27.13%	0.11%	5.40%	1.24	0.266
Did not reach mastery	18.02%	18.93%	17.18%	1.75%	5.70%	3.37	0.066
Did not start post-test	9.05%	8.09%	9.95%	1.86%	6.00%	1.44	0.231
Did not complete post-test	8.45%	7.92%	8.93%	1.01%	6.30%	1.72	0.190

Table 3: Model estimating the effect of Fill-In Problems on Post-Test Performance

	Model 1		
Predictors	*Log-Odds*	*SE*	*p*
Intercept	-1.02	0.27	**<0.001**
Fill-In	0.23	0.06	**<0.001**
Random Effects			
σ^2	3.29		
τ_{00}	0.83_i		
	0.60_c		
	0.25_j		
ICC	0.34		
N	4940_i		
	363_c		
	4_j		
Observations	9657		

$$logit(\text{Student } i \text{ Gets Post-Test Problem } j \text{ Correct}) = \gamma_0 + \gamma_1 \text{Fill-In}_i + \gamma_2 \text{Fill-In}_i \text{Prior Performance}_i + \mu_i + \mu_c + \mu_j \quad (3)$$

6.2 Results

Table 4 displays the results for Model 5. The main effect (γ_1) is the effect of the Fill-In problem for the students who received the average score because scaled prior performance is centered at the mean. The effect of Fill-In on the likelihood of getting the post-test correct for students with average prior performance effect is non-significant (γ_1 = 0.13, SE = 0.09, p = 0.156). The interaction between the prior performance and the Fill-In problem set is significant and positive (γ_3 = 0.20, SE = 0.10, p = 0.042). To ensure that this effect was not a spurious product of our number of prior problems completed cut point of 10, we ran model 3 varying outputs ranging from one prior problem through ten, which did not change the significance or direction of the effects. Therefore, the effect of Fill-In problems compared to MCQs appears to depend on the student's prior math ability–especially for high-performing students. Fill-In problems led to better post-test performance, while for lower-performing students, the effect was smaller and possibly negative.

Table 4: Exploring potential heterogeneity in the students' performance on the transfer item across problem-types.

	Model 3		
Predictors	*Log-Odds*	*SE*	*p*
Intercept	-3.45	0.24	**0.001**
Fill-In	0.13	0.09	0.156
Prior Performance (Z-score)	0.55	0.08	**<0.001**
Fill-In x Prior Performance (Z-score)	0.20	0.10	**0.042**
Random Effects			
σ^2	3.29		
τ_{00}	0.61_i		
	0.77_c		
	$0._j$		
ICC	0.32		
N	1643_i		
	100_c		
	4_j		
Observations	3253		

We visualize the interaction between Fill-In problems and prior performance in Figure 4 by plotting the predicted probability of a correct response on the post-test for each post-test attempt by Prior Performance for both Fill-In and MCQ conditions. These probabilities were predicted using Model 3. Notably, the visualization shows a negative effect of Fill-In problems for students with lower prior scores. To test whether this effect is significant, we ran a post-hoc[2] model based on Model 5 with prior performance low-end centered so that the main effect will be for the effect of Fill-Ins for students with the lowest prior performance scores. The main effect was not significantly significant γ_{10} = -0.61, SE = 0.38, p = 0.114). In summary, we have strong evidence that the effect of Fill-In problems is greater for students with higher prior performance compared with students with lower prior performance. However, there is insufficient evidence that the effect of MCQs is negative for students with lower performance.

[2]The full model output is available in the supplemental materials.

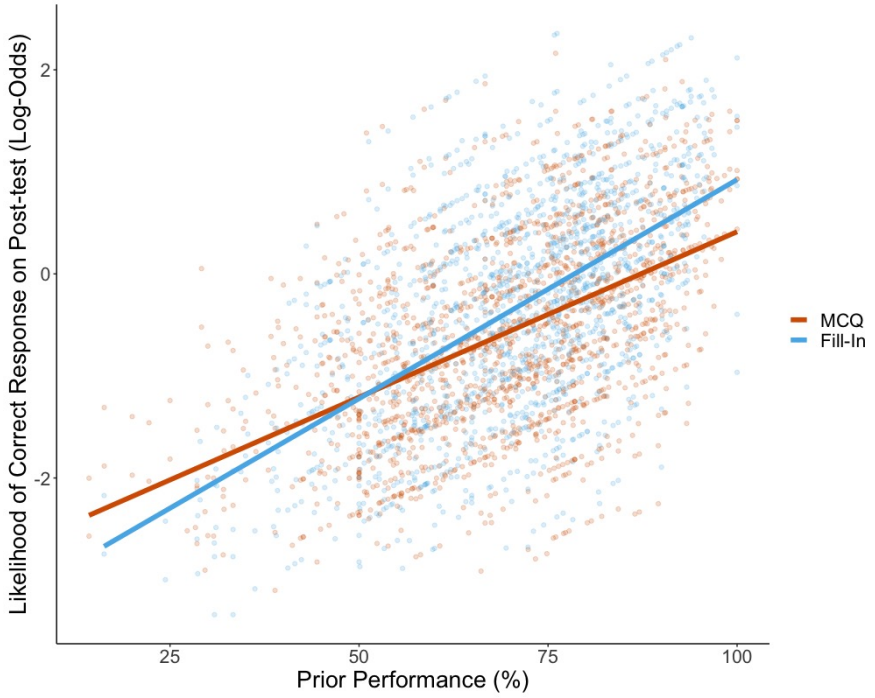

Figure 4: Interaction between prior performance(%) and the likelihood of a correct response on the transfer item on the post-test.

7 DISCUSSION AND FUTURE WORKS

Our analysis found that, on average, problem sets with Fill-In problems were more difficult and led to better learning outcomes than MCQs. However, we observed that the benefits of Fill-In problems have certain contextual constraints. The impact of Fill-In problems on learning was only significant for higher-knowledge students, and there is some evidence that MCQs may be more beneficial for lower-knowledge students. In sum, despite the nuances exposed by our analyses, Fill-In problems have a more positive effect on student learning than MCQs.

One possible explanation for why Fill-In problems were more effective at improving learning than MCQs is that they increase the likelihood that students will participate in productive struggle. Inducing productive struggle has been shown to increase learning [7]. Often this is done by creating *desirable difficulties*—such as varying presentation of content [45], interweaving knowledge components instead of presenting them sequentially [41], spacing content delivery [11] and retrieval practice [24]-–during instruction and practice. ORP problems in non-mathematical problem-solving settings are often associated with retrieval practices. Although mathematical Fill-In problems don't solely rely on retrieval, they do require the student to generate their answers independent of any prompts and, therefore, might have a similar benefit to retrieval activities. Taken together, the findings of higher difficulty and greater post-test performance caused by Fill-In problems also align with findings of desirable difficulties, which are often associated with lower students' performance, even as these design choices can positively affect learning as measured by distal outcomes [40, 43, 46].

Further evidence that the interplay between problem type, difficulty, and learning may be inducing productive struggle can be found in the other differences in student behavior across conditions. In Table 1, we reported that the students in the Fill-In problems invested more time presumably thinking before taking their first action than those in the MCQs because they perceived the problems to be more difficult. They were also more likely to utilize hints. These differences may indicate that Fill-In problems are producing better learning behaviors in students, which may be underlying causal mechanisms producing the differences in learning outcomes. Future research should study these potential processes by which problem types impact student learning. One way of doing this would be to use multiple mediation analysis, to evaluate causal paths that lead from problem type to mastery demonstration to discern and assess how problem types influence student behaviors which, ultimately, cause differences in learning outcomes.

Furthermore, the analysis in Section 6 exploring the heterogeneity of the Fill-In effects indicates that not all students benefit equally from Fill-In problems. We observed that students with higher prior performance benefited more than students with lower prior performance. Although this finding implies that students with lower prior performance might benefit more from MCQs than Fill-In problems; however, we cannot make more substantial claims due to the sparsity of students with low prior performance in the data. Despite this uncertainty, our analysis shows impact differentials for Fill-In based on students' knowledge before beginning the activity. There are some plausible explanations for this phenomenon. High-knowledge students may have the ability to learn the concept addressed in the problem sets but may need the challenge of having to produce

the answers themselves without the MCQ options to truly benefit from the activity. Alternatively, lower-knowledge students may benefit from the options in the MCQs but are less likely to learn the concepts well enough to transfer their knowledge to problems where they must provide the answer independently. Regardless of the underlying mechanisms behind the penalization effect, the finding provides evidence that LX designers and instructors may have to consider adapting problem types to students' needs.

8 LIMITATIONS

There are a few key limitations of our work. First, we conducted experiments on two very specific content areas, where we found evidence that content may influence the effect of the problem type on learning. This research should be replicated using different content areas across different subjects to fully understand the heterogeneity of the impact the problem types can have on student learning. A further limitation of the current study is the lack of student demographic information. The CBLP platform we used in this study does not collect personally identifiable information about the students, per the IRB Protocol; thus, we cannot make any advances in understanding the more fine-grained differences within our sample.

Our work has an experimental design limitation as we only used Fill-In problems for our post-test. Concerns regarding this design limitation are valid, yet, we argue that knowledge, by nature, should be transferable upon mastery and, as such, would be independent of the instrument used during evaluation. While such assumptions regarding transferability can be problematic, the balanced post-test completion rates across conditions indicate that students from both conditions were comfortable with the design of the post-test. However, we feel that using the Fill-In problem is justifiable as Fill-In problems are an accurate measure of student ability. Further exploration using a combination of both MCQs and Fill-In problems would help establish the optimal approach in the design of assignments as the combination of both activities could enhance learning outcomes or, conversely, the switch between problems in the post-test could cause cognitive load leading to higher dropout rates. Similarly, additional work exploring the benefits and drawbacks of other types of close and open-ended activity design would be beneficial to understand their assessment and learning value.

9 CONCLUSION

Overall, findings from the present study present causal evidence that problem types influence how and whether students learn. We observed that, on average, students had better learning outcomes when using mastery-based assignments with Fill-In problems compared to MCQs. We also demonstrated the robustness of our findings by evaluating them across various contextual scenarios, *i.e.*, pre-pandemic, pandemic, and summer sessions. We took a comprehensive approach and evaluated the heterogeneity effects of the two methods, where we observed that high-performing students benefited more from Fill-In problems.

We hope that the findings of this paper can help inform the design of learning experiences on CBLPs, as we provide evidence that problem types have an impact on learning outcomes. While our findings present the potential benefit of using Fill-In problems in designing learning activities, it is important to highlight that different students, as indicated by their prior performance, may require different types of activity design in order to facilitate more effective learning. We believe that LX designers and instructors will benefit from our findings when designing learning and assessment activities where they are continually required to balance the trade-offs between the use of open and close-ended activities to facilitate learning while assessing student knowledge.

ACKNOWLEDGMENTS

We would like to thank NSF (*e.g.*, 2118725, 2118904, 1950683, 1917808, 1931523, 1940236, 1917713, 1903304, 1822830, 1759229, 1724889, 1636782, & 1535428), IES (*e.g.*, R305N210049, R305D210031, R305A170137, R305A170243, R305A180401, & R305A120125), GAANN (*e.g.*, P200A180088 & P200A150306), EIR (U411B190024 & S411B210024), ONR (N00014-18-1-2768), and Schmidt Futures. None of the opinions expressed here are that of the funders.

REFERENCES

[1] Hayri Akay, Danyal Soybaş, and Ziya Argün. 2006. Problem kurma deneyimleri ve matematik öğretiminde açık-uçlu soruların kullanımı. *Gazi Üniversitesi Kastamonu Eğitim Dergisi* 14, 1 (2006), 129–146.

[2] Stephen A Anderson. 1994. Synthesis of Research on Mastery Learning. https://files.eric.ed.gov/fulltext/ED382567.pdf

[3] Ryan Baker, Jason Walonoski, Neil Heffernan, Ido Roll, Albert Corbett, and Kenneth Koedinger. 2008. Why students engage in "gaming the system" behavior in interactive learning environments. *Journal of Interactive Learning Research* 19, 2 (2008), 185–224.

[4] Christine Bastin and Martial Van der Linden. 2003. The contribution of recollection and familiarity to recognition memory: A study of the effects of test format and aging. *Neuropsychology* 17, 1 (2003), 14–24. https://doi.org/10.1037/0894-4105.17.1.14

[5] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67, 1 (2015), 1–48. https://doi.org/10.18637/jss.v067.i01

[6] Joseph E. Beck and Yue Gong. 2013. *Wheel-Spinning: Students Who Fail to Master a Skill.* Lecture Notes in Computer Science, Vol. 7926. Springer Berlin Heidelberg, Berlin, Heidelberg, 431–440. https://doi.org/10.1007/978-3-642-39112-5_44

[7] Robert A Bjork and Elizabeth L Bjork. 2020. Desirable difficulties in theory and practice. *Journal of Applied research in Memory and Cognition* 9, 4 (2020), 475.

[8] Benjamin S. Bloom. 1968. Learning for Mastery. Instruction and Curriculum. Regional Education Laboratory for the Carolinas and Virginia, Topical Papers and Reprints, Number 1. *Evaluation Comment* 1, 2 (May 1968). https://eric.ed.gov/?id=ED053419 issue: 2 container-title: Evaluation Comment volume: 1 ERIC Number: ED053419.

[9] H Douglas Brown and Priyanvada Abeywickrama. 2004. *Language Assessment: Principles and Classroom Practices.* Pearson Education, New York, USA.

[10] Donald R. Cahill and Robert J. Leonard. 1999. Missteps and masquerade in American medical academe: Clinical anatomists call for action. *Clinical Anatomy* 12, 3 (1999), 220–222. https://doi.org/10.1002/(SICI)1098-2353(1999)12:3<220::AID-CA14>3.0.CO;2-K

[11] Nicholas J Cepeda, Harold Pashler, Edward Vul, John T Wixted, and Doug Rohrer. 2006. Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological bulletin* 132, 3 (2006), 354.

[12] TJ Cooney, WB Sanchez, K Leatham, and DS Mewborn. 2004. Open-ended assessment in math: A searchable collection of 450+ questions.

[13] Albert T Corbett and John R Anderson. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction* 4, 4 (1994), 253–278.

[14] Lee J. Cronbach. 1988. *Five perspectives on the validity argument.* Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, US, 3–17. https://doi.org/10.1037/14047-004

[15] Shayan Doroudi and Emma Brunskill. 2019. Fairer but Not Fair Enough On the Equitability of Knowledge Tracing. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge* (Tempe, AZ, USA) *(LAK19)*. Association for Computing Machinery, New York, NY, USA, 335–339. https://doi.org/10.1145/3303772.3303838

[16] Steven C Funk and K Laurie Dickson. 2011. Multiple-choice and short-answer exam performance in a college classroom. *Teaching of Psychology* 38, 4 (2011), 273–277.

[17] Lucinda Gray and Laurie Lewis. 2021. Use of Educational Technology for Instruction in Public Schools: 2019–20. https://nces.ed.gov/pubs2021/2021017Summary.

pdf
[18] Thomas R. Guskey and Therese D. Pigott. 1988. Research on Group-Based Mastery Learning Programs: A Meta-Analysis. *The Journal of Educational Research* 81, 4 (Mar 1988), 197–216. https://doi.org/10.1080/00220671.1988.10885824
[19] Christopher J. Harrison, Karen D. Könings, Lambert W. T. Schuwirth, Valerie Wass, and Cees P. M. van der Vleuten. 2017. Changing the culture of assessment: the dominance of the summative assessment paradigm. *BMC Medical Education* 17, 1 (Dec 2017), 73. https://doi.org/10.1186/s12909-017-0912-5
[20] Neil T. Heffernan and Cristina Lindquist Heffernan. 2014. 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 (Dec 2014), 470–497. https://doi.org/10.1007/s40593-014-0024-x
[21] What Works Clearing House. 2022. Procedures and Standards Handbook (V5). https://ies.ed.gov/ncee/wwc/Docs/referenceresources/WWC-HandbookVer5.0AppIES-508.pdf
[22] David Hu. 2011. How Khan Academy is using Machine Learning to Assess Student Mastery. http://david-hu.com/2011/11/02/how-khan-academy-is-using-machine-learning-to-assess-student-mastery.html
[23] Gallop Inc. 2019. Education Technology Use in Schools. https://www.newschools.org/wp-content/uploads/2020/03/Gallup-Ed-Tech-Use-in-Schools-2.pdf
[24] Jeffrey D Karpicke and Franklin M Zaromb. 2010. Retrieval mode distinguishes the testing effect from the generation effect. *Journal of Memory and Language* 62, 3 (2010), 227–239.
[25] Kim M. Kelly, Yan Wang, Tamisha Thompson, and Neil T. Heffernan. 2015. Defining Mastery: Knowledge Tracing Versus N- Consecutive Correct Responses. In *proceedings of the 8th International Conference on Educational Data Mining*. Association for Computing Machinery, New York, NY, USA, 39–46. http://web.wpi.edu/Pubs/ETD/Available/etd-041416-122623/unrestricted/wang.pdf#page=42
[26] Jindrich Klufa. 2015. Multiple Choice Question Tests – Advantages and Disadvantages. *Mathematics and Computers in Sciences and Industry* (2015), 91–97.
[27] Kenneth R Koedinger, Paulo F Carvalho, Ran Liu, and Elizabeth A McLaughlin. 2023. An astonishing regularity in student learning rate. *Proceedings of the National Academy of Sciences* 120, 13 (2023), e2221311120.
[28] Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M. Henne. 2009. Controlled experiments on the web: survey and practical guide. *Data Mining and Knowledge Discovery* 18, 1 (Feb 2009), 140–181. https://doi.org/10.1007/s10618-008-0114-1
[29] Chen-Lin C. Kulik, James A. Kulik, and Robert L. Bangert-Drowns. 1990. Effectiveness of Mastery Learning Programs: A Meta-Analysis. *Review of Educational Research* 60, 2 (Jun 1990), 265–299. https://doi.org/10.3102/00346543060002265
[30] Jeri L Little and Elizabeth Ligon Bjork. 2015. Optimizing multiple-choice tests as tools for learning. *Memory & Cognition* 43, 1 (2015), 14–26.
[31] Jeri L Little, Elizabeth Ligon Bjork, Robert A Bjork, and Genna Angello. 2012. Multiple-choice tests exonerated, at least of some charges: Fostering test-induced learning and avoiding test-induced forgetting. *Psychological science* 23, 11 (2012), 1337–1344.
[32] Joseph P Magliano, Keith Millis, Yasuhiro Ozuru, and Danielle S McNamara. 2007. A multidimensional framework to evaluate reading assessment tools. In *Reading Comprehension Strategies: Theories, Interventions, and Technologies* (1 ed.), Danielle S. McNamara (Ed.). Vol. 1. Psychology Press, New York, USA, Chapter A Multidimensional Framework to Evaluate Reading Assessment Tools, 107–136.
[33] Council of Chief State School Officers National Governors Association Center for Best Practices. 2010. Common Core State Standards (Mathematics Standards). http://www.corestandards.org/Math/
[34] Institute of Education Sciences; National Center for Education Evaluation and Regional Assistance. 2003. Identifying and Implementing Educational Practices Supported by Rigorous Evidence: A User Friendly Guide. https://doi.org/10.1037/e370412004-001
[35] Murat Polat. 2020. Analysis of Multiple-Choice versus Open-Ended Questions in Language Tests According to Different Cognitive Domain Levels. *Novitas-ROYAL (Research on Youth and Language)* 14, 2 (2020), 76–96.
[36] Ethan Prihar, Manaal Syed, Korinn Ostrow, Stacy Shaw, Adam Sales, and Neil Heffernan. 2022. Exploring Common Trends in Online Educational Experiments. In *Proceedings of the 15th International Conference on Educational Data Mining*. International Educational Data Mining Society, Durham, United Kingdom, 27–38. https://doi.org/10.5281/zenodo.6853041
[37] R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
[38] Dominique P. Rauch and Johannes Hartig. 2010. Multiple-choice versus open-ended response formats of reading test items: A two-dimensional IRT analysis. *Psychological Test and Assessment Modeling* 52 (2010), 354–379.
[39] Alexander Robitzsch, Thomas Kiefer, and Margaret Wu. 2022. *TAM: Test Analysis Modules*. https://CRAN.R-project.org/package=TAM R package version 4.1-4.
[40] Henry L Roediger III and Jeffrey D Karpicke. 2006. Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological science* 17, 3 (2006), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x
[41] Doug Rohrer, Robert F Dedrick, and Kaleena Burgess. 2014. The benefit of interleaved mathematics practice is not limited to superficially similar kinds of problems. *Psychonomic bulletin & review* 21 (2014), 1323–1330.
[42] Kenneth Ruit and Patrick Carr. 2011. Comparison of student performance on "selected-response" versus "constructed-response" question formats in a medical neuroscience laboratory practical examination. *The FASEB Journal* 25, S1 (Apr 2011), 182–186. https://doi.org/10.1096/fasebj.25.1_supplement.182.6
[43] John B Shea and Robyn L Morgan. 1979. Contextual interference effects on the acquisition, retention, and transfer of a motor skill. *Journal of Experimental psychology: Human Learning and memory* 5, 2 (1979), 179.
[44] Megan A Smith and Jeffrey D Karpicke. 2014. Retrieval practice with short-answer, multiple-choice, and hybrid tests. *Memory* 22, 7 (2014), 784–802.
[45] Steven M Smith and Justin D Handy. 2014. Effects of varied and constant environmental contexts on acquisition and retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 40, 6 (2014), 1582.
[46] Steven M Smith and Ernst Z Rothkopf. 1984. Contextual enrichment and distribution of practice in the classroom. *Cognition and instruction* 1, 3 (1984), 341–358.
[47] Nicholas C. Soderstrom and Robert A. Bjork. 2015. Learning Versus Performance: An Integrative Review. *Perspectives on Psychological Science* 10, 2 (March 2015), 176–199. https://doi.org/10.1177/1745691615569000
[48] Brenda Sugrue, Noreen Webb, and Jonah Schlackman. 1998. *The Interchangeability of Assessment Methods in Science. CSE Technical Report 474.* Technical Report. Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.
[49] Xu Wang, Carolyn Rose, and Ken Koedinger. 2021. Seeing Beyond Expert Blind Spots: Online Learning Design for Scale and Quality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–14. https://doi.org/10.1145/3411764.3445045
[50] Xu Wang, Srinivasa Teja Talluri, Carolyn Rose, and Kenneth Koedinger. 2019. UpGrade: Sourcing Student Open-Ended Solutions to Create Scalable Learning Opportunities. In *Proceedings of the Sixth (2019) ACM Conference on Learning @ Scale* (Chicago, IL, USA) *(L@S '19)*. Association for Computing Machinery, New York, NY, USA, Article 17, 10 pages. https://doi.org/10.1145/3330430.3333614