To draw or not to draw? Examining the necessity of problem diagrams using massive open online course experiments

Zhongzhou Chen,[1,3,*] Neset Demirci,[3,†] Youn-Jeng Choi,[2] and David E. Pritchard[3]

[1]*Department of Physics, University of Central Florida, Orlando, Florida 32816, USA*
[2]*Department of Educational Studies in Psychology, Research Methodology & Counseling, University of Alabama, Tuscaloosa, Alabama 35487, USA*
[3]*Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*

(Received 5 April 2016; published 17 February 2017)

Previous research on problem diagrams suggested that including a supportive diagram, one that does not provide necessary problem solving information, may bring little, or even negative, benefit to students' problem solving success. We tested the usefulness of problem diagrams on 12 different physics problems (6A/B experiments) in our massive open online course. By analyzing over 8000 student responses in total, we found that including a problem diagram that contains no significant additional information only slightly improves the first attempt correct rate for the few most spatially complex problems, and has little impact on either the final correct percentage or the time spent on solving the problem. On the other hand, in half of the cases, removing the diagram significantly increased the fraction of students' drawing their own diagrams during problem solving. The increase in drawing behavior is largely independent of students' physics abilities. In summary, our results suggest that for many physics problems, the benefit of a diagram is exceedingly small and may not justify the effort of creating one.

DOI: 10.1103/PhysRevPhysEducRes.13.010110

I. INTRODUCTION

As physics instructors, we often feel obliged to accompany the problems we write with a figure or a diagram. It is common wisdom that a figure or a diagram can help students understand and solve the problem, especially when it involves complex spatial relations or unfamiliar objects. In a typical commercial physics textbook, a significant fraction of problems are accompanied by a diagram. While some diagrams provide necessary problem solving information that is not conveyed in the problem body, such as circuit diagrams, free body diagrams, or voltage phasor diagrams, many are *supportive diagrams*—a term we use when all of the information necessary for solution is already unambiguously included in the problem body, and the diagram adds no new information relevant to problem solving.

Often, creating a "good looking" diagram or figure often consumes several times the time and effort of creating the problem text itself, and requires additional proofreading to determine whether figure and text are fully in accord, especially when the problem is to be published in either a printed textbook or an online platform. Therefore, it is important to measure how much a supportive diagram helps students solve physics problems in order to determine whether the benefit justifies the cost of creating one. Since all the diagrams involved in this study are supportive, we will refer to them simply as "diagrams" in the remainder of this paper.

A number of cognitive learning theories suggest, although somewhat indirectly, that supportive diagrams can still be potentially beneficial to problem solving. Paivio's dual coding hypothesis [1] and the ensuing multimedia learning theories [2,3] imply that, when parallel verbal and visual channels are utilized to convey information, significantly fewer cognitive resources are required, leading to more accurate processing and freeing up more cognitive capacity that can be delegated to problem solving. However, both of those theories relate to teaching rather than problem solving. From a problem-solving perspective, the extra mental effort required to generate a visual representation from text can potentially deepen understanding of the problem, suggesting that drawing a diagram may have benefits over simply being presented with one.

More importantly, the cognitive theories do not take into consideration the impact of a diagram on students' problem solving *behavior*, which can be a more dominant factor in a real instructional setting. Evidence supporting a stronger behavioral impact comes from a series of recent experiments by Lin, Maris, and Singh [4–6] which found that for the problems involved in their study the accompanying diagrams provide no detectable benefit for problem solving, and in some cases significantly hurt performance.

[*] Zhongzhou.Chen@ucf.edu
[†] ndemirci@gmail.com

Based on this observation, the authors suggested that in the presence of a problem diagram, students are less likely to draw one on their own, which in some cases leads to shallow processing of the problem. In other words, their study suggests that in certain cases, the presence or absence of a diagram can indeed significantly impact students' problem solving behavior, which in turn influences their problem solving outcome.

One of the major limitations of those studies is that only a very small number of problems (2 in each study) and students (<60 each group) were studied. This limits the ability of the study to provide a general suggestion to instructors on whether or when to include a diagram. In addition, some problem diagrams may only be beneficial for students at a particular level of proficiency. Heckler [7] reported that prompting weaker students to construct free body diagrams can actually hurt their problem solving performance.

Recent developments in online educational technology, especially massive open online courses (MOOCs), provide an opportunity to address the limitations of previous studies [8–10]. The large number of participants in our MOOC ensures adequate statistics for detecting small differences or looking at a subsection of students, even with moderate response rate. Adding or removing a diagram from a problem is much easier in an online platform, enabling us to test a large number of problems covering different topics, involving objects of various familiarity, and with different levels of visual and spatial complexity. Finally, MOOC students present a much wider distribution of background knowledge compared to students in a typical brick and mortar classroom. This allows us to study the impact of diagrams on students with different abilities. In addition, students' problem solving behavior can be studied by appending survey questions following the problem.

By utilizing the "split test" feature of the edX platform [11], which allows the instructor to randomly present different materials to students in a MOOC, we seek to disentangle the complicated relationships among a problem diagram, students' problem solving performance, students' drawing behavior, and students' background ability. More specifically, this study addresses the following research questions in the context of a calculus based introductory mechanics course:

(1) Do diagrams in general have an impact on students' problem solving performance (either percentage of correct answers or time spent on problem solving)? If so, to what extent?
(2) Do diagrams given with problems change students' problem solving behavior, or more specifically, their decision to draw their own diagram?
(3) How does spontaneously drawing a diagram influence problem solving outcome? (as compared to being prompted to draw a diagram in Ref. [7]?)
(4) Do students with different physics ability react differently to the presence or absence of a diagram?
(5) What types of problems (if any) are more likely to require a diagram?

II. MATERIALS AND METHODS

The experiments described in this paper are conducted in 8.MReVx Mechanics Review, an edX MOOC ran by the RELATE group at MIT [12,13] from May 29th to September 14th 2014. This MOOC covers most of the topics in a typical college level introductory mechanics course, and is designed for students with some existing knowledge of Newtonian mechanics. In summer 2014, the course received ∼11 000 registrations, with over 500 students' receiving certificates.

A. Controlled AB experiment on the edX platform

The edX platform allows the course creator to create controlled AB experiments by splitting the student population into two or more groups (called "partitions"), and presenting each group with a different version of content, such as a problem or a series of problems or html pages.

Every student who tries to access the experimental course content for the first time is randomly assigned to one of the groups at the time of first access. This assignment strategy insures that each group will have approximately the same number of students despite the fact that MOOC students may randomly choose not to access the experimental content.

The instructor can choose to keep the same group assignment (partition) for multiple contents in different locations of the course. For example, we can make sure that all students who received a diagram in problem 1 will receive no diagram in problem 2. The instructor can also use a different partition for different experiments in the same course, which reduces systematic bias between different groups, and prevents interference between earlier and later experiments in the same course. In the current study, we kept the group partition for the two problems within the same experiment, and used a different partition for each experiment, as described in detail in Sec. II C.

B. Structure of 8.MReVx Mechanics Review

The 2014 iteration of 8.MReVx consists of 12 required units and 2 optional units, with each unit designed to be about a week long. A typical unit contains three sections: instructional e-text (with embedded checkpoint problems), homework, and quiz. All components of a single unit are released and are due at the same time. To accommodate for the varying schedule of MOOC students, each unit is released at least 4 weeks ahead of the due date.

All checkpoint problems, homework, and quiz problems are graded. Students earn a certificate for the course if they obtain a minimum of 60% of total course credit. Most graded problems in the course allow multiple attempts. All quiz problems allow 3 attempts, whereas numeric and symbolic response problems on the homework allow for up to 10 attempts. The number of attempts on any multiple choice problem equals half of the available choices to prevent exhaustive guessing. There is no time limit for completing any of the problems, as long as they are completed before the due date of the unit. Quiz and homework problems are weighted almost the same towards the final course grade. The only difference between a quiz problem and a homework problem is that the correct answer to the quiz problem is only released after the unit due date, whereas the correct answer to the homework problem is available to students after they finish (get the correct answer or depleted all the attempts) the problem. Data analysis showed little (if any) differences in the completion rate of homework vs quiz problems [14].

C. Experiment design

A total of six AB experiments with identical design were implemented throughout the first eight units of the course. We chose to implement the experiments in the first eight units mainly because of two reasons: First, based on previous experience there is a significant drop of activity towards the end of the course, since only 60% of the course credit is necessary for a certificate. Second, two different experiments were conducted in the homework and quiz section of units 9–12 [13]; therefore, the diagram experiment was restricted to the first eight units to avoid potential interference between experiments.

Each experiment involved two problems chosen from either the homework or the quiz section of a given unit, so the entire study involves twelve different problems in total. The problems were chosen from the first eight units of the course, covering kinematics, Newton's laws, circular motion, conservation of momentum, and conservation of energy.

The two problems had all the necessary information required for solution expressed unambiguously in the problem text. For each problem, we created two versions with identical problem text: one with a diagram (DG) and one with no diagram (NDG).

In each two-problem experiment, the student population was randomly partitioned into two groups, A and B (Fig. 1). Group A saw the first problem in DG format and the second problem in NDG format. Group B saw the two problems in the same order, but the DG and NDG conditions were reversed. The group assignment for each experiment is independent, reducing systematic bias in the population.

Immediately after each problem, we presented students with the following survey question:

When solving this problem, (check all that apply)
(i) I drew one or more diagrams
(ii) I wrote down some equations
(iii) I did the problem entirely in my head
(iv) I used some other means to solve the problem

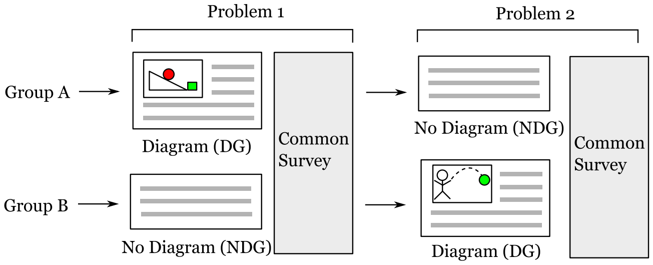


FIG. 1. Experiment design. Each experiment consisted of a pair of problems. The same design was used for all 6 experiments conducted.

D. Obtaining students' physics skill through item response theory

A dichotomous item response theory (IRT) model was used to obtain the students' skills (ability) and item difficulty and item discrimination parameter estimates. Although multiple attempts were allowed to students in our test, only "first attempted correct" was considered as a correct response. To get the IRT ability values to represent students' skills, 604 students who tried more than 50% of all items were selected and 1197 items (checkpoint, homework, quiz, and final test) were used to estimate IRT. The students' skills were estimated by the BILOG-MG IRT computer program [15]. A two- parameter logistic item response theory (2PL IRT) model was applied using a marginalized maximum likelihood estimation (MMLE) method [16]. The person centering method was applied as an identification problem constraint and a concurrent calibration equating method was used to put all IRT values on the same scale for comparisons of skills and item parameters [17].

Using IRT, the physics skills (θ) of the student population obeys a Gaussian distribution with mean of 0 and width of 1. In this study, we define weak students as those with $\theta \leq -0.5$, median students as those with $-0.5 < \theta \leq 0.5$, and strong students as those with $\theta > 0.5$. To ensure the accuracy of skill estimation, IRT analysis is only performed on those who attempted $> 50\%$ of all the problems in the course.

E. Student populations used in the analysis

In a MOOC environment, students have more freedom to choose the assignments that they wish to complete. As a result, the number of students completing each problem (including its survey question) varies and drops as students either drop out of or complete the course, as shown in Table II.

As previously explained, each experiment contained two problems, and the partitions were different for each

TABLE I. Number of students included in the analysis for each problem. We only included students who answered both the problem and the survey question.

Problem	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	Total
N(DG)	480	432	327	303	303	230	296	257	284	224	241	187	3564
N(NDG)	473	428	354	282	272	239	283	283	269	228	257	185	3553

TABLE II. Number of IRT eligible students with different physics skills in the NDG condition.

NDG	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	Total
N(Low)	50	56	58	49	27	38	58	53	44	37	35	36	541
N(Medium)	98	100	94	100	87	91	91	105	89	92	72	85	1104
N(High)	61	60	62	55	60	66	58	71	55	65	70	47	730

experiment. For example, the 480 students in DG condition of P1 are largely the same students as the 428 students in P2 NDG condition (except for a few that completed only one of the two problems). The 327 students in the P3 DG condition came randomly from both the P1 DG and P1 NDG conditions, also including those who did not attempt P1. Since our experimental design minimized systematic bias in the population, it is reasonable to treat each observation as independent. The last column contains the total number of observations for each condition.

Not every one of those students completed $> 50\%$ of all the course problems. Therefore, IRT skill can only be calculated for a fraction of those students. The number of IRT eligible students for each problem is listed in Table II.

Notice that the total number of IRT available observations in the DG and NDG conditions is the same for the adjacent pairs of problems. This is because we switched the conditions between the two problems in each experiment, and every student who completed $> 50\%$ of the course problems completed both problems in each experiment.

F. Obtaining time on task information for problem solving

The “time on task” information for each problem is obtained by analyzing the click stream data from the edX platform. For a single student, the time spent on a single problem is defined as the time between the first access

TABLE III. Rating rubric for problem-diagram pairs.

Code	Question	Detailed explanation
Rep_Info	Does this diagram represent a significant fraction of the key information contained in the problem body?	Consider information that is either key to solving the problem or might affect the solution, such as additional variables not necessary for the solution.
Irr_Visual	Rate the amount of irrelevant visual elements, or visual elements clearly inconsistent with the problem description	Such as showing a real car or human, or showing a small ball when the problem says “massive ball.”
Diagram_Helpful	Overall, do you think this diagram will be helpful to students solving this problem?	
Unfamiliar_Objects	What kinds of objects are involved in this problem?	
Non-Visual_Info	What fraction of key information in this problem CANNOT be easily represented visually?	Such as “elastic collision” or “frictionless surface.”
Need_Draw	In general, do you think students need to draw a diagram in order to solve this problem?	
Spatial_Complexity	Spatial complexity of problem	Evaluate the spatial complexity of the problem, by estimating how many geometric relations are needed to solve the problem, such as finding the correct angle, or finding a relevant distance.
Temporal_Complexity	Temporal complexity of problem	Evaluate the temporal complexity of the problem, by estimating how many sequential steps were described in the problem, such as follows: First the person throws a ball, then the ball hits a wall. This would be considered as two steps.

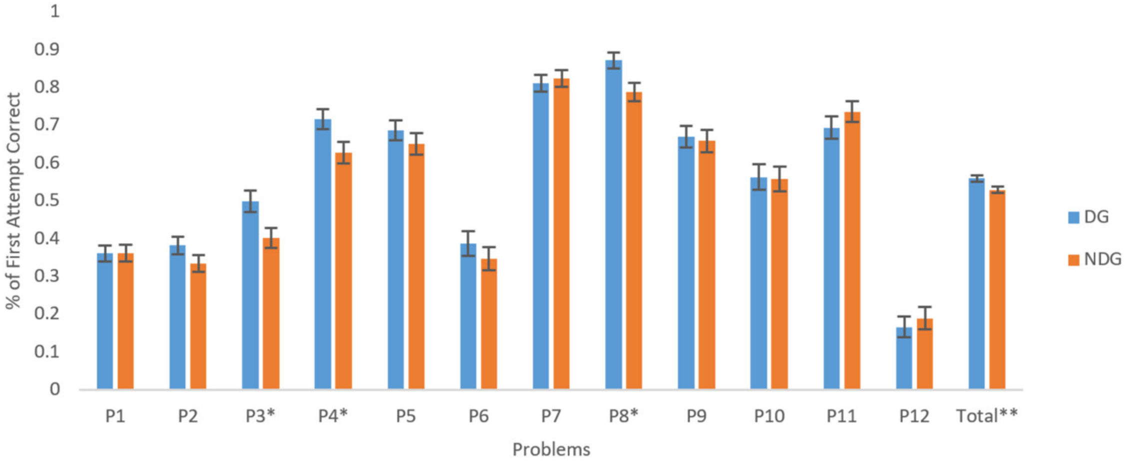


FIG. 2. Percentage of first attempt correct for each problem. The rightmost column is the percentage correct aggregated over all 12 problems. *Difference is significant at the 0.05 level. ** Difference is significant at the 0.01 level. (chi-squared test).

(opening the webpage that contains the problem) and the last attempt on the problem. If the student opened a different edX page during problem solving, the time between opening that page and the next submission on the current problem is excluded from the total time on task. In addition, if no new events were recorded 30 min after the problem was open, we assume that the student was not actively working on the problem, and in that case the problem access event is discarded from the analysis. Similarly, if a student navigated away from the problem in less than 10 seconds, we do not count that time as time spent on solving the problem.

G. Expert evaluation of diagram and problem

We asked 5 physics experts to rate both the problems and the diagrams on the following aspects on a 1–3 scale (Table III), with 1 indicating disagree, and 3 indicating agree, and 2 indicating neutral.

The only exception is the fourth question "Unfamiliar Objects: What kinds of objects are involved in this problem." On this question, the experts were given four choices:

(1) Involves ideal objects such as blocks and pulleys
(2) Involves relatively familiar real world objects such as baseball and bus
(3) Involves relatively unfamiliar real world objects or situations
(4) Involves real world objects or situations that most people have not heard of

In the final analysis, we combined the last two choices, and coded the three choices 1–3, with 1 being ideal physics objects and 3 being unfamiliar real world objects, or objects that most people have not heard of. The reason to code ideal physics objects as one is because we think that physics objects such as blocks and pulleys are very familiar to most of the students in this course, and carry fewer visual features than real objects.

The physics experts were selected from physics faculties, postdocs, and graduate TAs who have experience teaching introductory mechanics.

III. RESULTS

A. Adding a problem diagram slightly improves performance on problem solving

We first look at the impact of including a diagram on the difficulty of physics problems. Problem difficulty is measured by the percentage of correct answers on students' first attempts. As shown in Fig. 2, in most cases the presence or absence of a diagram had little impact on the difficulty of the problem itself. Only 3 out of 12 problems (P3, P4, and P8) showed a significant difference in difficulty between the two conditions ($p < 0.05$, $\chi^2 > 5$).

Since we carefully balanced systematic bias in the population in our experiment design, it is meaningful to add up the data from all 12 problems and compare the overall success rate between the DG vs NDG conditions.[1] As shown on the rightmost column of Fig. 2, the overall correct rate under the DG condition is higher than that in the NDG condition by $3 \pm 0.8\%$. The difference, although small, is still statistically significant due to the large cumulative sample size (~3500 observations per condition, $p < 0.01$, $\chi^2 = 6.9$).

Essentially no differences between the two conditions were observed for either the final attempt correct rate (~87% on average) or the average number of attempts used on each problem (ranging from 1.4 to 3 attempts).

In Figure 3 we compare the median time on task in seconds for solving each problem under the two conditions. Since the distribution of MOOC time data is highly non-normal, with a long, one-sided tail and often more than one

[1] For further discussion on the legitimacy of adding the experiment data, see the Sec. IV.

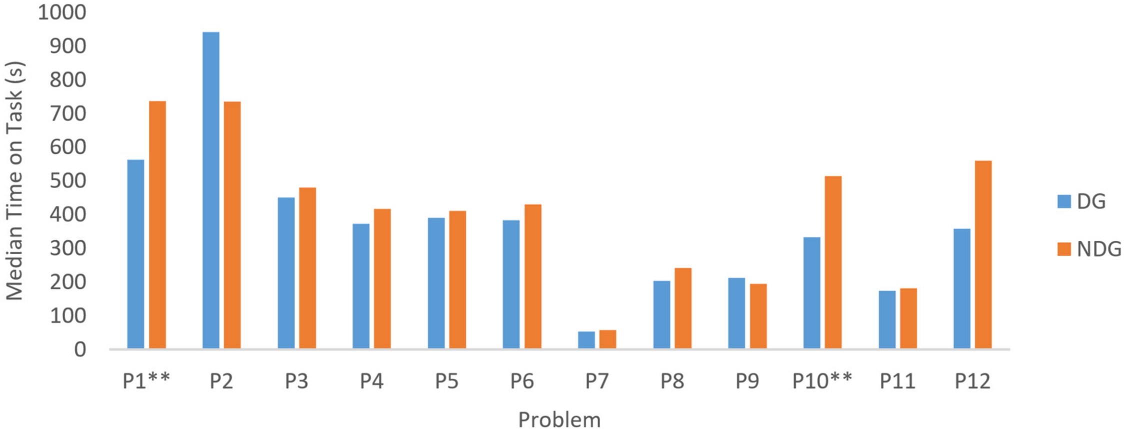


FIG. 3. Comparing median time on task (seconds) for each problem. **Significant at $p = 0.01$ level, Mann-Whitney U test.

peak, we use the Mann-Whitney U test to examine whether the distributions of time data are statistically different from each other. Two of the problems (P1 and P10) showed a significant difference ($p < 0.01$) between the conditions. In both cases, the median time for solving the problem ranges from 400 to 600 s, and the NDG group took ∼200 s longer than the DG group. Note that due to the unique distribution of MOOC time data, it is difficult to capture the difference by one variable such as the mean or the median [18]. For example, although on P2 the median time of the DG condition is longer than that of the NDG condition, the difference in the distributions is insignificant ($p = 0.22$ on a Mann-Whitney U test). More careful treatment of the time data requires sophisticated statistical models, which we deem to be unnecessary for our purpose in this study.

B. Adding a problem diagram reduces students' tendency to draw their own

We also investigated how the presence or absence of a problem diagram impacts students' tendency to draw their own diagram. The fraction of students drawing their own diagram is measured by students' answers on the survey question following each problem. As shown in Fig. 4(a), on 7 out of 12 problems, a significantly lower fraction of students ($p\langle 0.01, \chi^2\rangle 7$, chi-square test) in the DG condition reported drawing their own diagram during problem solving than in the NDG condition. Combining the data across all 12 problems, students in the DG condition are 10% less likely to draw their own diagram than in the NDG condition ($p < 0.001$, $\chi^2 = 65$).

A noteworthy observation is that while for some problems the drawing behavior is highly sensitive to the DG and NDG condition, other problems are far less sensitive. We plot the difference in the percentage of student drawing (DG-NDG) for all problems in Fig. 4(b), and we see that the data points form two distinct groups: one centered around zero (no difference) and the other around −15% (significant difference). As a rough estimate, the standard deviation of the 12 data points is 0.08, while the average standard error of each data point (combined standard error from the two conditions) is 0.04. Therefore, it is unlikely that the differences between the problems arise from random noise.

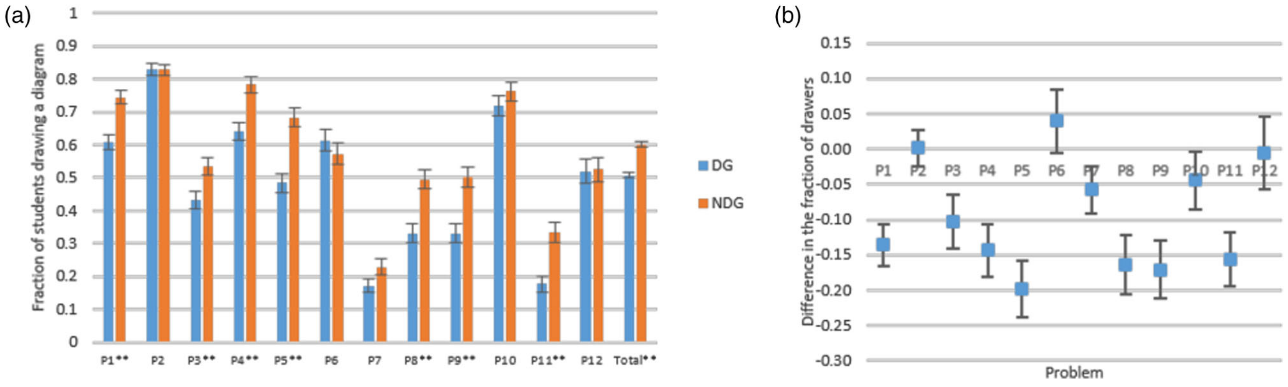


FIG. 4. Comparing the drawing behavior between the two conditions. (a) Percentage of students who chose to draw a diagram during problem solving. The rightmost columns are the drawing percentage aggregated over all 12 problems. *Difference is significant at the 0.05 level. ** Difference is significant at the 0.01 level. (chi-squared test) (b) The difference in percentage between the two conditions on each problem.

C. Adding a diagram might only improve problem solving performance of “nondrawers.”

To further understand how the decision of drawing a diagram interacts with the DG and NDG condition, we divide the subject population into “drawers” and “nondrawers” for each problem, and look at their performance separately.

As shown in Fig. 5(a), the DG and NDG condition has no overall impact on “drawers” first time percentage correct. Indeed, on problems P11 and P12, the DG condition even resulted in significantly worse performance than the NDG condition. On the other hand, “nondrawers” are much more sensitive to the DG and NDG condition. As shown in Fig. 5(b), nondrawers perform significantly better overall in the DG condition compared with the NDG condition, and the DG condition outperformed the NDG condition on P3, P4, and P8, identical to that observed for the entire population.

A possible explanation of this observation is that the drawing behavior helped students overcome the disadvantage of the NDG condition. However, it must be pointed out that the observed difference could also be partly explained by an artifact of the current experimental design. Namely, a student can choose to draw a diagram after their failed first attempt, but cannot become a nondrawer by “undrawing” a diagram at any point. Therefore, the observed first attempt correct rate is an underestimation for drawers and an overestimation for nondrawers. In the Appendix [19], we will discuss this artifact in detail, and argue that while it may partly give rise to the differences shown in Fig. 4, it is probably not the only cause of the observed difference.

For the same reason, comparing the first attempt correct rate between drawers and nondrawers under the same DG and NDG condition is misleading, since the correct rate for drawers will always be less than that of the nondrawers; see the Appendix [19].

D. The impact of diagrams on students with different physics abilities

The impact of problem diagrams on problem solving is likely dependent on student’s physics ability. On one hand, weaker students may have more difficulty generating a visual representation from verbal description, while on the other hand, they are also more likely to benefit from being forced to carefully read and understand the problem body. The very wide distribution of backgrounds and physics abilities in MOOCs naturally lends itself for the investigation of such dependencies. As mentioned in the methods section, we probed these questions by dividing our

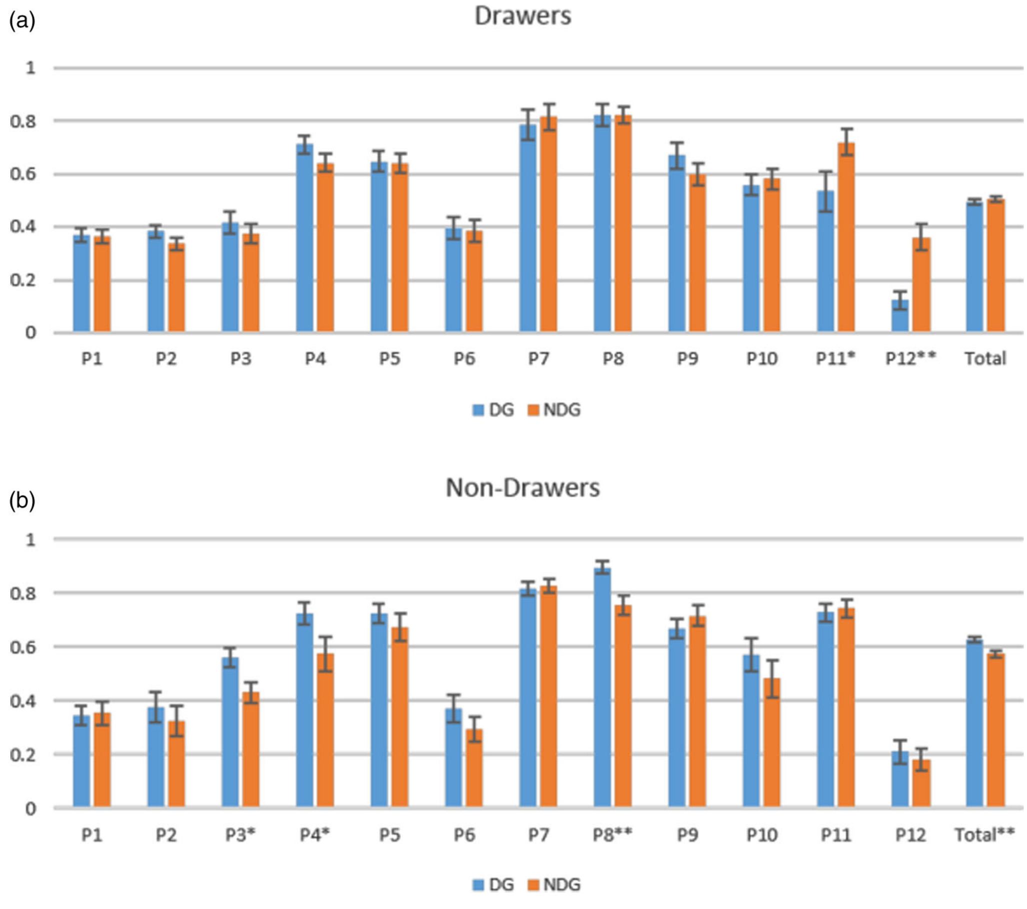


FIG. 5. Comparing the percentage of first attempt correct rate between the two conditions for “drawers” (a) and “nondrawers” (b). *Difference is significant at the 0.05 level. ** Difference is significant at the 0.01 level. (chi-squared test).

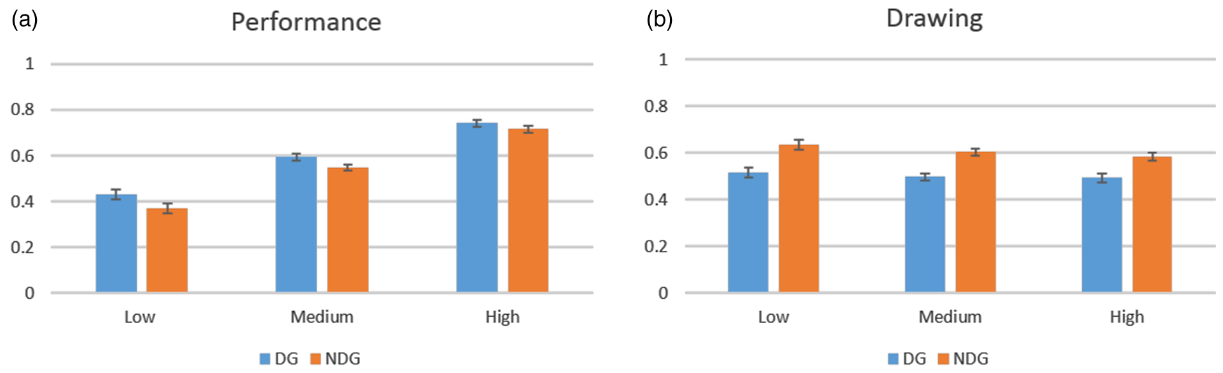


FIG. 6. The impact of DG and NDG condition on students with different skill backgrounds averaged over all 6 experiments. (a) The impact on performance as measured by first attempt correct rate. (b) The impact on drawing behavior.

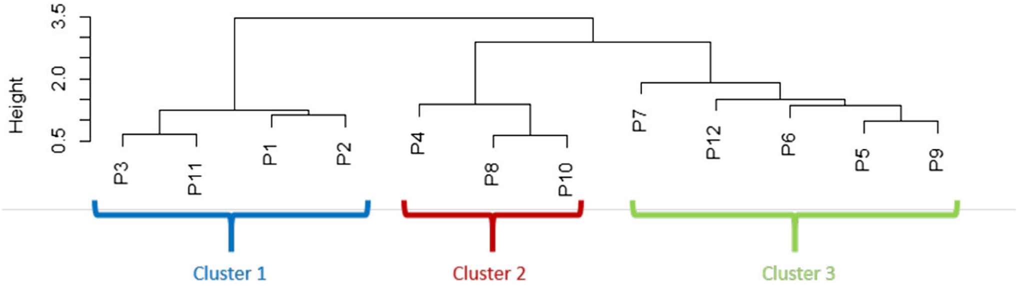


FIG. 7. Dendrogram of cluster analysis on expert's ratings of problems. The height is a metric for measuring the relative distance between items and clusters. The three major clusters are largely independent of either the clustering algorithm or the choice of distance metric.

subjects into three cohorts: high skilled, medium skilled, and low skilled, based on their IRT skill parameter.

As shown in Fig. 6(a), for low and medium skilled students, the overall first attempt correct rate for problems in the DG condition is about 5% higher than that of the NDG condition, and the differences are statistically significant [$p < 0.05$, $\chi^2 = 3.94$ (low skill), $p < 0.03$, $\chi^2 = 5.0$ (medium skill)]. However, for high skilled students, the overall performance difference is smaller ($< 3\%$), and statistically insignificant ($p = 0.26$, $\chi^2 = 1.24$), while still in the same direction.

In contrast, a student's decision to draw their own diagram during problem solving is much more uniform across all skill groups [Fig. 6(b)]. For all three skill levels, an equal fraction of subjects decided to draw a diagram during problem solving (DG: $p = 0.71$, $\chi^2 = 0.7$, NDG: $p = 0.16$, $\chi^2 = 3.56$), while students in the DG condition are ∼10% less likely to do so than students in the NDG condition regardless of their skill.

For each individual problem, the sample size drops to ∼30–50 people in each skill level, making it difficult to identify any significant differences.

E. What kinds of problems are more sensitive to the DG and NDG condition?

To understand if problems that are more sensitive to the DG and NDG condition have common features, we asked physics experts to rate 8 different features of the problems on a 1–3 scale. We then performed cluster analysis (ward.D2 method using Euclidian distance) on the average rating of the 12 problems involved, the result of which is shown in the dendogram in Fig. 7.

From this graph it is reasonable to conclude that the 12 problems form three clusters at the relative distance of 2. Plotting the average rating of the three clusters on a radar graph (Fig. 8), we see that problems in cluster 1 are unique

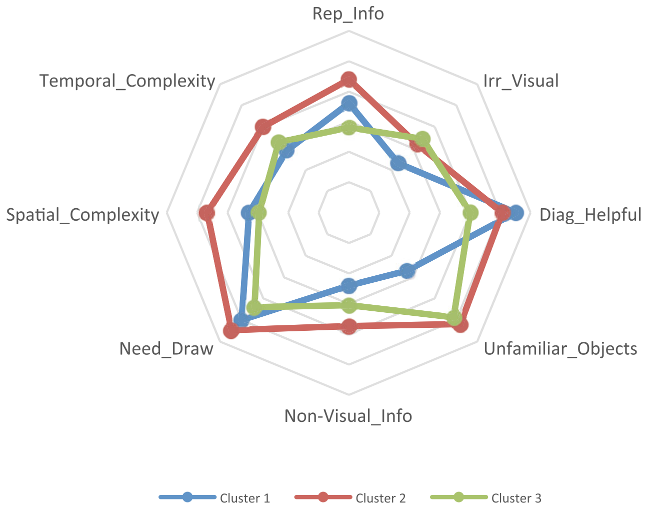


FIG. 8. Radar graph for the average rating of each problem cluster.

TABLE IV. Difference in sensitivity to the DG and NDG condition for each problem. Problems that showed a statistically significant difference to the DG and NDG condition in either difficulty, time, or drawing are labeled 1 in the corresponding column.

Problem	Cluster	Difficulty sensitive	Time sensitive	Draw sensitive
P1	1	0	1	1
P2	1	0	0	0
P3	1	1	0	1
P11	1	0	0	1
P4	2	1	0	1
P8	2	1	0	1
P10	2	0	1	0
P5	3	0	0	1
P6	3	0	0	0
P7	3	0	0	0
P9	3	0	0	1
P12	3	0	0	0

in that they involve ideal physics objects such as blocks on pulleys (“Unfamiliar_Objects”). Cluster 2 stands out as problems that have complicated spatial or temporal information (“Spatial_Complexity” and “Temporal_Complexity”) and that the diagrams represent more key information from the problem. Cluster 3 is a loosely formed cluster that is the least challenging on visual or spatial information, rated low on the helpfulness of diagram, and involves unfamiliar or real-world objects. (See examples in Sec. IV.)

In Table IV, we list the different types of sensitivity of each problem to the DN and NDG condition, grouped by clusters. Most problems in clusters 1 (ideal physics objects) and 2 (spatially demanding) are sensitive to the DG and NDG condition in one way or another, although with only 12 problems the difference is not statistically significant (Fisher’s exact test $p = 0.22$). More specifically, problems that are sensitive in terms of either difficulty or time on task, all belong to the first two clusters.

IV. DISCUSSION

A. Impact of problem diagram on problem solving

Perhaps the most surprising observation of this study is how little students benefit from a problem diagram. Even with the large sample size provided by MOOC, significant differences between the two conditions are only observed for 3 out of 12 problems, with the largest difference at 10% and the overall difference at merely 3%.

We did observe significant differences for the few problems that involve exceptionally complicated visual, spatial, or temporal information. For example, P4 (Fig. 9) involves a spatially complicated object that is very unfamiliar to most students.

P8 (Fig. 10) deals with a somewhat unnatural sequence of events.

However, for most problems at the level of spatial or temporal complexity commonly seen for physics problems, such as the three examples shown here (Fig. 11), no difference in performance was observed between the two conditions.

P12 is particularly surprising, not only because it involves an object (trampoline) that can be unfamiliar to students, especially international students, but also because for “drawers,” the DG group performed significantly worse than the NDG group. This means that the given problem

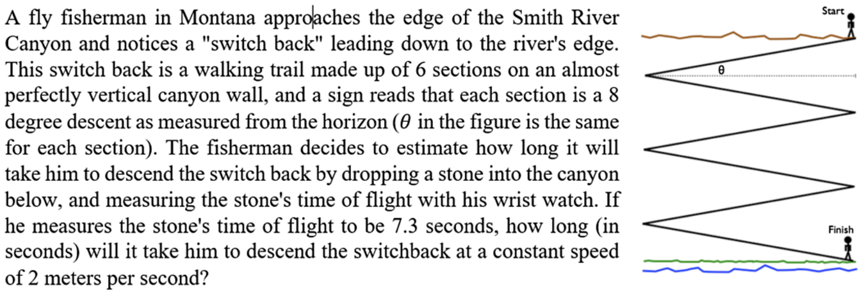

A fly fisherman in Montana approaches the edge of the Smith River Canyon and notices a "switch back" leading down to the river's edge. This switch back is a walking trail made up of 6 sections on an almost perfectly vertical canyon wall, and a sign reads that each section is a 8 degree descent as measured from the horizon (θ in the figure is the same for each section). The fisherman decides to estimate how long it will take him to descend the switch back by dropping a stone into the canyon below, and measuring the stone's time of flight with his wrist watch. If he measures the stone's time of flight to be 7.3 seconds, how long (in seconds) will it take him to descend the switchback at a constant speed of 2 meters per second?

FIG. 9. Diagram and problem text of problem P4.

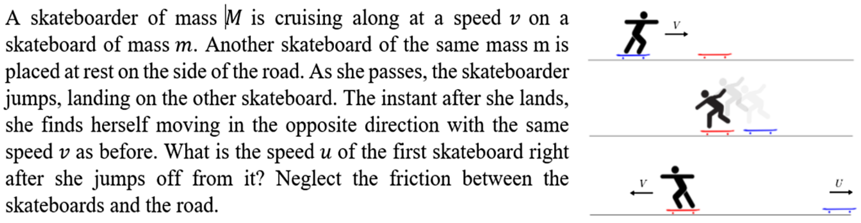
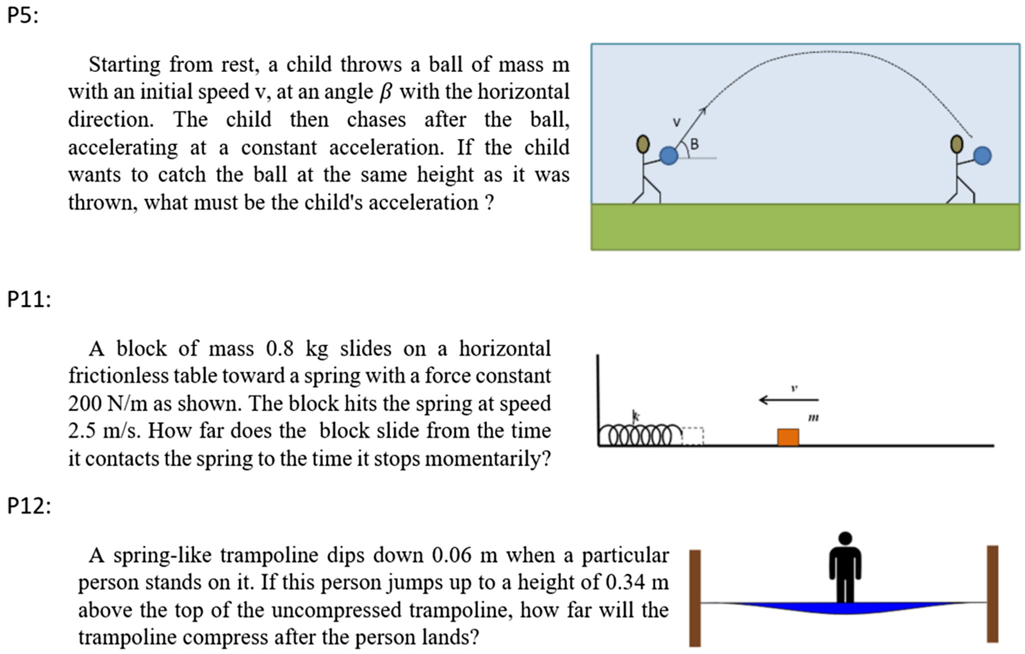

A skateboarder of mass M is cruising along at a speed v on a skateboard of mass m. Another skateboard of the same mass m is placed at rest on the side of the road. As she passes, the skateboarder jumps, landing on the other skateboard. The instant after she lands, she finds herself moving in the opposite direction with the same speed v as before. What is the speed u of the first skateboard right after she jumps off from it? Neglect the friction between the skateboards and the road.

FIG. 10. Diagram and problem text for problem P8.



P5:

Starting from rest, a child throws a ball of mass m with an initial speed v, at an angle β with the horizontal direction. The child then chases after the ball, accelerating at a constant acceleration. If the child wants to catch the ball at the same height as it was thrown, what must be the child's acceleration ?

P11:

A block of mass 0.8 kg slides on a horizontal frictionless table toward a spring with a force constant 200 N/m as shown. The block hits the spring at speed 2.5 m/s. How far does the block slide from the time it contacts the spring to the time it stops momentarily?

P12:

A spring-like trampoline dips down 0.06 m when a particular person stands on it. If this person jumps up to a height of 0.34 m above the top of the uncompressed trampoline, how far will the trampoline compress after the person lands?

FIG. 11. Diagram and problem text for problems P5, P11, and P12.

diagram likely interfered with students' ability to draw a productive diagram on their own. In this case, it may be that the depiction of the trampoline interfered with students' ability to treat it as a normal spring.

Together, those results show that even though the benefits predicted by conventional wisdom and the dual-coding hypothesis may still exist, the effect size is small in an *in vivo* situation and only significant in the most extreme cases. For the majority of "normal" physics problems, our findings are consistent with previous studies [5,6,20] indicating that the benefit of providing a diagram is small, and in certain cases may even hurt problem solving.

B. Impact of problem diagram on student's decision to draw a diagram

For the 12 problems in this study, the fraction of students who drew a diagram during problem solving ranged from less than 20% to over 80%, indicating that "drawing a diagram" is more of a deliberate decision rather than a fixed habit for most students. To further verify this point, we analyzed the behavior pattern for 307 students who completed at least 5 out of 6 experiments as well as the accompanying surveys. Among those students, only 26% consistently (in at least 4 out of 6 experiment) drew both diagrams, and 16% consistently draw no diagrams.

In contrast, a student's decision to draw is very sensitive to the DG and NDG condition on 7 out of 12 problems: when the problem diagram is removed, students are 10% more likely to draw their own. Furthermore, neither the decision to draw a diagram nor the difference between the DG and NDG conditions depend on students' physics ability. In other words, weaker students and stronger students reach much the same decision when deciding whether they need to draw a diagram or not during problem solving.

Interestingly, for most problems about one-half of the students still choose to draw their own diagram when a problem diagram was provided. However, this is not so surprising considering the fact that most diagrams provided are simply a depiction of the problem situation with little extra information. In addition, the fact that the diagrams are presented on the computer screen rather than on a piece of paper where a student can directly sketch might also have contributed to the observation. Finally, notice that since we only included students who completed the optional survey, there might also be a self-selection effect (e.g., only the more motivated students are likely to report drawing a diagram).

C. On the validity of adding data from all 12 problems

In this study we reported the "average total" difference between the two conditions by averaging over all 12 problems. This requires us to address a couple of caveats.

1. Students drop out steadily throughout the study (see Table I), with the first three problems averaging nearly 400 and the last one just under 200. This means that the students who persisted are weighted more heavily. However, it is almost impossible to tell whether the student population is becoming stronger or weaker, as both "good" and "bad" students have equally good reasons to drop an online course. Furthermore, we did not observe any qualitative difference between the results of earlier experiments and those from the later ones, suggesting that those who dropped out may not have performed very differently in the experiments from those who persisted.

2. Students' decision to draw or not draw a diagram may be influenced by their problem solving habit in addition to the experiment conditions. Indeed, when analyzing data from ~300 students who completed at least 5 out of 6 experiments, we found that 26% of them consistently draw both diagrams for at least 4 experiments, while 16% consistently choose to not draw a diagram. However, we point out that since the DG and NDG condition is flipped for each student in each experiment, and that we are only interested in the difference between the conditions, students who draw and do not draw diagrams for both problems in a single experiment do not contribute to the difference, therefore they do not affect the result.

Because of the above mentioned reasons, we believe that the average differences reported in the results section provide meaningful and valid information despite the caveats.

V. SUMMARY AND RECOMMENDATION FOR INSTRUCTORS

For instructors, the study shows that for common physics problems like the ones involved in the current study, the benefit of adding a supportive diagram is quite small except for a few problems where the physical situation is relatively unfamiliar to many students.

On the other hand, omitting a problem diagram results in about 10% more students drawing their own diagram during problem solving. Our observations seem to suggest that students who drew their own diagram can better compensate for the loss of a given problem diagram. However, we note that the evidence is not conclusive under the current experimental design, as discussed in detail in the Appendix [19].

The observation that the drawing decision is independent of students' physics skills suggests that weaker students are equally motivated to draw a diagram compared to high skilled students, although the quality of the diagram may not be as good.

A. Recommendations

In light of our research, is it advisable (or even useful?) for instructors to not include nonessential supportive diagrams with their homework and exam questions. Given that diagrams give only a ~3% improvement in student success, *and* the fact that they reduce the chances that a student will draw a diagram on their own by about 10%, it appears that any advantages of including a diagram may well not justify the resources and effort required to create it.

B. Advantages and shortcomings of this study and suggested directions for future research

The use of MOOC AB experiments enabled us to make some unique contributions to the diagram and no diagram literature. Online technology allowed testing a much larger number of problems and students than previous studies. The large sample sizes of MOOCs allow for the detection of small effects in real learning environments. On top of detecting the existence of the effect—in this case the benefit of adding a diagram—we are also able to measure the size of the effect and decide whether the benefits justify the cost. In addition, the relatively large number of problems enabled us to explore how the characteristics of individual problems impact the effect of adding or removing a diagram.

Another advantage particular to MOOCs is the wide distribution of skills present in the class (approximately twice the standard deviation of our on-campus classes). This allowed us to compare the difference in performance and behavior among students with significantly different physics abilities.

The most prominent shortcoming of our study is that we requested no information on the type of diagrams drawn by students. We would like to know whether a student drew a sketch of the physics situation containing little additional information, or a free body diagram that represents a key step toward solving the problem. Hopefully with the rapid advancement in online education technology, we will be able to easily collect and analyze students' diagrams in future studies.

A second shortcoming is that while we measured the impact of diagramming on immediate problem solving, we did not measure its impact on knowledge transfer from one problem to the other. A problem diagram may enhance transfer by facilitating the visualization of the physical situation, or impede transfer by increasing the level of specificity of students' understanding [21]. Unfortunately, edX.org does not allow for easy control over the order in which students complete problems, which would be a nice addition.

Third, in the current study we did not measure the impact of diagrams on student engagement. It is possible that while including a diagram has little impact on problem solving success, it may improve student's engagement and reduce the rate of attrition. However, since we rotated the DG and NDG conditions between student populations in the current experiments, our experimental design provides little information on student engagement. This question could be answered in future research with a different experimental design.

Finally, a common concern among researchers is that while MOOCs provide large sample size, the demographics of MOOC students can be quite different from the average college or high-school students, in terms of both age and educational background [9]. As a result, it is not clear to what extent results obtained from MOOC experiments (or experiments conducted in any online learning environment) provide valuable information for improving traditional classroom teaching. This is indeed a legitimate concern, and we advise our readers to keep this fact in mind when interpreting the results of the current study. On the other hand, we argue that the disparity in demographics is no reason to dismiss the validity of online experiments altogether. In fact, it could be argued that similar differences exist between students enrolled in top research universities and those enrolled in community colleges. Instead, we hope that more research can be conducted to carefully evaluate the extent to which background differences affect the outcomes of experiments, and find more reliable experiment designs and data analysis methods to fully utilize the advantages of online AB experiments.

ACKNOWLEDGMENTS

We thank Dr. Giora Alexandron and Dr. Qian Zhou for their help with the data mining process, as well as Dr. Theodore Ducas for insightful discussion and comments on the manuscript. We gratefully acknowledge that the work was supported by MIT.

[1] A. Paivio, *Mental Representations: A Dual Coding Approach* (Oxford University Press, Oxford, England, 1986).

[2] R. E. Mayer, *Multimedia Learning* (Cambridge University Press, Cambridge, England, 2001).

[3] W. Schnotz, Towards an integrated view of learning from text and visual displays, Educ. Psychol. **14**, 101 (2002).

[4] A. Maries and C. Singh, Should students be provided diagrams or asked to draw them while solving introductory physics problems?, AIP Conf. Proc. **1413**, 263 (2012).

[5] S.-Y. Lin, A. Maries, and C. Singh, Student difficulties in translating between mathematical and graphical representations in introductory physics, AIP Conf. Proc. **1513**, 250 (2013).

[6] A. Maries and C. Singh, A good diagram is valuable despite the choice of a mathematical approach to problem solving, *AIP Conf. Proc.* 1513, 31 (2013)..

[7] A. F. Heckler, Some consequences of prompting novice physics students to construct force diagrams, Int. J. Sci. Educ. **32**, 1829 (2010).

[8] F. Han, K. Veeramachaneni, and U.-M. O'Reilly, Analyzing millions of submissions to help MOOC instructors understand problem solving, in *NIPS Workshop on Data Driven Education* (Lake Tahoe, Nevada, 2013).

[9] K. F. Colvin, J. Champaign, A. Liu, Q. Zhou, C. Fredericks, and D. E. Pritchard, Learning in an introductory physics MOOC: All cohorts learn equally, including an on-campus class, Int. Rev. Res. Open Distrib. Learn. **15**, 1 (2014).

[10] J. Reich, Rebooting MOOC research, Rebooting MOOC research, Science **347**, 34 (2015).

[11] edX edX Documentation: Creating Content Experiments, http://edx.readthedocs.org/projects/edx-partner-course-staff/en/latest/#.

[12] G. Alexandron, J. Antonio Ruiperez Valiente, Z. Chen, and D. E. Pritchard, Using multiple accounts for harvesting solutions in MOOCs, in *L@S2016* (Edinburg, Scotland, 2015).

[13] Z. Chen, C. Chudzicki, D. Palumbo, G. Alexandron, Y.-J. Choi, Q. Zhou, and D. E. Pritchard, Researching for better instructional methods using AB experiments in MOOCs: Results and Challenges, Res. Pract. Technol. Enhanc. Learn. **11**, 9 (2016).

[14] C. A. Chudzicki, Learning experiments in a MOOC (Massive Open Online Course), Massachusetts Institute of Technology, 2015.

[15] M. F. Zimowski, E. Muraki, R. J. Mislevy, and R. D. Bock, BILOG-MG 3.

[16] F. B. Baker and S.-H. Kim, *Item Response Theory: Parameter Estimation Techniques*, 2nd ed. (Marcel Dekker, New York, 2004).

[17] R. J. de Ayala, *The Theory and Practice of Item Response Theory* (Guilford Press, New York, 2009).

[18] A. Lamb, J. Smilack, A. Ho, and J. Reich, Addressing common analytic challenges to randomized experiments in MOOCs, in *Proceedings of the Second ACM Conf. Learn. @ Scale—L@S '15* (Vancouver, BC, Canada, 2015), pp. 21–30.

[19] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.13.010110 for a detailed discussion of the artifact in the current experimental design mentioned in the Results section.

[20] A. Maries and C. Singh, To use or not to use diagrams: The effect of drawing a diagram in solving introductory physics problems, AIP Conf. Proc. **1513**, 282 (2013).

[21] D. T. Brookes, B. H. Ross, and J. P. Mestre, Specificity, transfer, and the development of expertise, Phys. Rev. ST Phys. Educ. Res. **7**, 010105 (2011).