Evolution of response time and accuracy on online mastery practice assignments for introductory physics students

Megan Nieberding* and Andrew F. Heckler

Department of Physics, The Ohio State University, 191 West Woodruff Avenue, Ohio 43210, USA

(Received 29 April 2022; revised 23 March 2023; accepted 15 June 2023; published 16 August 2023)

We have investigated the temporal patterns of algebra ($N = 606$) and calculus ($N = 507$) introductory physics students practicing multiple basic physics topics several times throughout the semester using an online mastery homework application called science, technology, engineering, and mathematics (STEM) fluency aimed at improving basic physics skills. For all skill practice categories, we observed an increase in measures of student accuracy, such as a decrease in the number of questions attempted to reach mastery, and a decrease in response time per question, resulting in an overall decrease in the total time spent on the assignments. The findings in this study show that there are several factors that impact a student's performance and evolution on the mastery assignments throughout the semester. For example, using linear mixed modeling, we report that students with lower math preparation for the physics class start with lower accuracy and slower response times on the mastery assignments than students with higher math preparation. However, by the end of the semester, the less prepared students reach similar performance levels to their more prepared classmates on the mastery assignments. This suggests that STEM fluency is a useful tool for instructors to implement to refresh student's basic math skills. Additionally, gender and procrastination habits impact the effectiveness and progression of the student's response time and accuracy on the STEM fluency assignments throughout the semester. We find that women initially answer more questions in the same amount of time as men before reaching mastery. As the semester progresses and students practice the categories more, this performance gap diminishes between males and females. In addition, we find that students who procrastinate (those who wait until the final few hours to complete the assignments) are spending more time on the assignments despite answering a similar number of questions as compared to students who do not procrastinate. We also find that student mindset (growth vs fixed mindset) was not related to a student's progress on the online mastery assignments. Finally, we find that STEM fluency practice improves performance beyond the effects of other components of instruction, such as lectures, group-work recitations, and homework assignments.

DOI: 10.1103/PhysRevPhysEducRes.19.020111

I. INTRODUCTION

Nearly 55 years ago, Bloom published an article claiming 90% of students can master material an instructor teaches them if the material can be broken down into smaller units and fit to meet the student's needs [1]. He claimed that by tailoring the speed that material is delivered to students, instructors allow students who are comfortable with the material to move on to other topics, while students who are struggling with a topic can spend more time learning the material until the topic is mastered. In the decades following, a number of studies have shown the effectiveness of mastery learning across a number of fields [2], and in the field of university-level physics education, more recent studies have also documented the benefits of mastery learning [3–8].

However, the general notion that mastery learning is an effective instructional strategy oversimplifies the vast complexity of the domain of science, technology, engineering, and mathematics (STEM) learning, the different methods used to demonstrate and investigate mastery, and the target student populations [9]. In other words, there are numerous potentially interacting factors that likely modulate the effectiveness of mastery learning. The natural heterogeneity of topics, methods, and populations in educational settings compels us here to focus on cases that are applicable to important and common STEM educational contexts. Specifically, in this paper, we investigate the extent to which a number of educationally relevant factors, described below, affect mastery learning for students using an online mastery learning application,

*nieberding.17@osu.edu

called STEM fluency, designed for building fluency in basic skills and knowledge necessary for success in algebra-based and calculus-based university-level introductory physics for students in large research universities in the United States.

Rather than focusing on larger grain outcomes like exam scores, final grades, or retention, in this study, we will peer more into the "black box" of the process of learning and investigate the progression of mastery of several topics assigned multiple times throughout the semester. Therefore, we will study two longitudinally measured outcomes aimed at measuring fluency. The first outcome measures the student's response accuracy, which is a commonly studied indicator of mastery of the material. We expect to see improvements in the correctness of students' responses over the course of the semester if students are benefitting from using the STEM fluency application. The second outcome is a measure of speed, such as the time taken for completion of the assignment, and is less commonly studied. We expect to also see improvements in the time that it takes students to answer questions if students are benefitting from the STEM fluency assignments. Task completion time is well known to be a relevant measure of cognitive performance [10,11], especially for tests with time constraints [12]. Completion time has also been used to characterize or predict performance, such as rapid guessing on low-stakes tests [13], copying on homework assignments [14], or course performance for online assignments [15]. Further, response time is a common measure of cognitive load [16,17], which is important for our context since one of the goals of the online learning application, STEM fluency, is to reduce the cognitive load of basic skills, allowing students to solve more complex problems [8].

Task completion time has also been measured as an important factor in modeling learning. This idea has emerged, for example, from work by Newell and Rosenbloom (1980), who demonstrated a very general power-law decrease in task completion times as a function of trial numbers for a wide variety of tasks. The use of response times has especially been used to measure performance or model student mastery of a given "knowledge component" in intelligent tutoring systems [18,19] and in other learning contexts [20], including problem-solving [21]. One must always keep in mind, though, the possible confound of increasing speed due to retesting effects [22].

A. Factors affecting mastery learning

Mikula and Heckler have demonstrated via pretesting and post-testing that STEM fluency was effective in improving student accuracy and speed for a variety of vector math skills, and they also investigated which kinds of feedback were most beneficial in this online mastery learning context [8]. In this study, we will expand to a wider variety of content topics and investigate the extent to which several other important factors affect the progression of mastery learning and performance.

First, we will study a design-level factor. To begin, it is important to note that since we collect data on student performance during practice, each mastery practice assignment for a given topic in this study is also an "*in situ*" assessment of accuracy and speed on that topic. Therefore, in order to first establish whether STEM fluency practice adds educational value, we will compare performance on mastery assignments between "fully trained" students and "partially trained" students for several skill categories. For a given category, fully trained students complete four mastery assignments, with the first assignment starting less than a week after the first lecture on the relevant topic (but before the first homework on the topics is due) and the remaining three assignments starting 1–10 weeks after the first. Partially trained students complete only the third and fourth mastery assignments for that category (as shown in Fig. 1). Therefore investigating whether there are differences in accuracy and speed between conditions in the third mastery assignment will allow us to compare the effect of completing two STEM fluency assignments to no STEM fluency training, keeping in mind that both conditions may have gains due to other course components such as lectures and homework assignments. Further, this design will allow us to compare students practicing twice vs 4 times to determine the extent to which more practice on a given topic is providing significantly more benefit. These results will help us to better calibrate how many times students on average should practice for a given topic.

We will also study the dependency of the evolution of performance on several educationally important student-level factors. The first factor is prior preparation. In this study, we will use ACT (or SAT equivalent) math score as a proxy for prior preparation, at least for math preparation. While we acknowledge some larger potential issues with

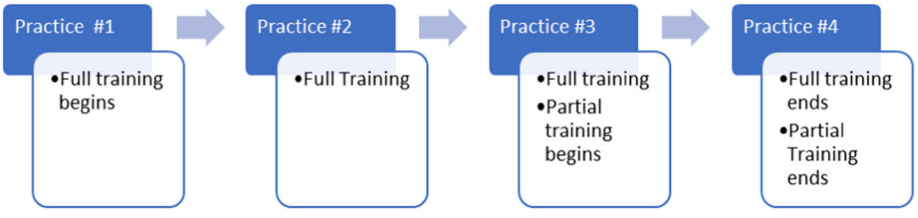

FIG. 1. Study 2 design showing the differences in training between full and partial trained students. Practice 1 begins less than one week after the first lecture on the topic begins. The subsequent practices occur sequentially, each 1–10 weeks after the first.

demographic biases [23], the ACT math score is well documented to be predictive of introductory physics grades [23,24]. Since we are studying the STEM fluency application that is designed to improve the mastery and fluency of basic skills, we are especially interested in the extent to which there is an interaction between preparation and training. Specifically, do students with lower ACT scores have higher gains in performance than students with higher ACT scores? If so, this may support the intention of such practice to especially help underprepared students.

The second student-level factor studied is gender. Studies have documented gender differences in homework scores with women tending to score higher than men [23–25]. More specifically, in a study by one of the authors that is set in the same institution, courses, and assignments as in this study, Simmons and Heckler have documented that women achieve higher scores (i.e., completion rates) on the STEM fluency mastery assignments [23]. Given these gender differences and the critical gender disparities in physics enrollments, it is important to investigate whether there are gender differences in the evolution of performance in these mastery assignments. For example, controlling for ACT score, do men and women start at the same level of accuracy and speed? Is there a difference in gains in accuracy and speed?

The third student-level factor is related to procrastination. Felker and Chen found that rewarding students with extra credit for submitting assignments early encourages low-performing students who typically submit assignments late to complete assignments earlier than they otherwise would have and spend more time studying for the class [26]. This could be an important factor when designing mastery assignments or when developing interventions related to students' procrastination. Additionally, in a recent study, we used submission time as a proxy for measuring student behavior [27]. For example, students who submitted assignments closer to the deadline earned lower grades in the course and completed fewer assignments than students who submitted assignments early. Further, there were differences in procrastination by gender. For this study, we are interested in determining if students who procrastinate have smaller gains in their performance on mastery assignments, as opposed to students who do not procrastinate as much. We might expect if students are procrastinating that they are not completing all the training sessions and that they are spending less time on assignments than students who are submitting assignments early thus impacting the evolution of the progress a student makes on the mastery assignments.

The fourth student-level factor involves the construct of mindset, which can be considered as a theory of "challenge-seeking and resilience" [28]. In this study, we are specifically interested in the claim that students with a growth (as opposed to fixed) mindset persist to overcome challenges and are more resilient to failure [29], though it is important to note that some researchers have found evidence that does not support this claim [30]. Applying this idea to the context of our study, we are interested in determining whether there is a relation between mindset (growth vs fixed) and performance on the assignments, especially for students who initially struggle with the assignments. For example, considering students who do poorly on the initial mastery assignments, do students with a growth mindset perform better on subsequent assignments compared to students with a fixed mindset? We will also consider how overall performance on all mastery assignments is related to the growth mindset, though correlations between mindset and measures of academic achievement, such as exam scores, have been found to be positive but very weak and with wide variation [31,32].

The final student-level factor involves the course grade. While the course grade may be an outcome rather than a predictive variable, for purposes of gaining further insight into how different students evolve in mastery training, we will also present descriptive statistics in the form of graphs for students with different final course grades. For example, do students with lower final grades also have initially lower performance on mastery assignments on basic skills, and do they evolve differently than students with high grades?

In summary, this study investigates the evolution of accuracy and speed in online mastery learning of several basic introductory physics skills on the timescale of weeks. We also investigate several educationally important factors that may affect mastery learning, including design-level factor and several student-level factors. More specifically, our research questions are as follows:

RQ1: How do measures of speed and accuracy evolve during mastery practice over timescales of weeks?

RQ2: To what extent does the initial level and subsequent evolution of measures of speed and accuracy vary with the student-level factors of ACT score, gender, procrastination, mindset, and final grade?

RQ3: To what extent does STEM fluency practice improve accuracy and speed beyond what is learned in other components of the course?

RQ4: To what extent is there benefit in practicing multiple times? For example, do performance gains in both accuracy and speed quickly saturate?

II. METHODS

A. Participants, materials, and design

This investigation is comprised of two studies that were conducted at a large public research university located in the United States Midwest during the first semester of a two-semester calculus-based and algebra-based introductory physics course. The studies included participants from two semesters: Study 1, conducted in the Autumn of 2019, investigated RQ1 and one factor in RQ2, while study 2, conducted in the Spring of 2021, investigated RQ1–RQ4 and employed quantitative statistical modeling. Participants

TABLE I. Data on all enrolled students and study participants.

				Total course enrollment				Study participants			
Study	Semester	Course	Number of sections	Number of students	% Female	Mean grade	Mean ACT	Number of students	% Female	Mean grade	Mean ACT
1	Autumn 2019	Algebra	4	817	61	2.74	27.9	445	68	3.06	28.2
		Calculus	7	1322	25	2.51	30.4	677	29.1	2.82	30.6
2	Spring 2021	Algebra	4	606	64	2.96	27.5	332	68.7	3.19	27.9
		Calculus	3	507	28	2.46	29.1	315	29.8	2.85	29.5

were included in this study if they consented to include their data in this study, which was requested on the first assignment. As shown in Table I, about 52% of all students enrolled were included in study 1 and 58% of all students enrolled were included in study 2. This is significantly below full participation because approximately 70% of students completed the first assignment and about 80% of those students consented to participate. While this participation rate does introduce some potential selection effects in our data, the sample was somewhat representative of the population, as seen in Table I. Specifically, the mean grades were 0.2–0.4 grade points (0.2–0.4 standard deviations) higher, the mean ACT scores were 0.2–0.4 points (0.05–0.1 standard deviations) higher, and the female participation rate was 2%–7% points higher for the study sample compared to all students enrolled in the course.

The course structure included a lecture section with traditional lectures (lecturing most of the time, with occasional lecture demonstrations, and occasional question-and-answer with students), a recitation section comprised of group work and/or quizzes, and a traditional lab section. The Autumn 2019 classes were in person, and, due to the pandemic, the Spring 2021 classes were virtual (including online Zoom lecturers and recitation group work via Zoom rooms), but all other aspects of the course were identical. The graded course components included a set of nonexam components, such as weekly homeworks, participation, and lab grades (30% of total grade) and exam/quiz components (70%). The weekly STEM fluency assignments were included as a nonexam component of the student's grade and accounted for 3% of the student's grade.

During each semester, online STEM fluency units were assigned weekly. The first and last units were pretest and post-tests on topics that are not included in this study, and the remaining units were mastery assignments. Each mastery assignment consisted of 3–5 categories to complete. To complete or "master" a category, students were required to correctly answer three or four questions in a row in that category. In study 1, there was a mix of assignments, some requiring three questions in a row and others requiring four questions. We realized after study 1 that this additional variation in the number of questions in a row required a more complicated data analysis; therefore, in study 2, four questions were required for all assignments.

The students were given feedback on the correctness of their responses immediately after they submitted their answers to each question. If they answered incorrectly on a question, they were given the option to try to answer 2 more times, after which they could choose to view the correct answer. However, if they incorrectly answered a question on the first try for a given category, the counter indicating the number of correct questions in a row for that category would be reset to zero and students would have to answer four more questions in a row to master that category. A student received full credit for each category they mastered and zero credit for categories not mastered, and the grade depended on the proportion of completed categories. An investigation of login and logout time stamps revealed that the vast majority of students completed each assignment in one sitting and typically took 10 to 30 min to complete. The weekly assignment window opened on Tuesdays at noon and closed on Sundays at 11:59 pm.

In this version of STEM fluency, the questions were all in multiple-choice format. The questions and responses were carefully and iteratively designed based on an evidence-based process described by Mikula and Heckler [8] involving feedback from student performance and prior research on student difficulties. While common distractors were often included as answer options, the questions were not designed to be especially difficult or "tricky." Rather, they were designed to be a straightforward and effective practice of specific skills with careful variation in a range of relevant practice dimensions such as representation format, physical context, magnitude, sign and direction of parameters, and which variables are known and unknown. On average, students took between 30 s and 2 min to answer each question.

Each week students practiced four or five different categories, with a total of about 12 categories per semester. We investigated only a portion of the categories covered throughout the semester, namely for each study, we selected *a priori* those that were well developed, assigned multiple times in the semester, and spanned a range of topics. We investigated five categories each in studies 1 and 2, with three of the categories overlapping between the studies. The lack of complete overlap occurred because of uncontrollable differences in assignment schedules between semesters, and we were also interested in increasing the number of

TABLE II. Description of categories students practiced throughout the semester and during which semesters we collected student performance data on that particular category (sample questions for each category are found in Appendix C).

Category	Label	Study	Category description
Free body diagram net force trigonometry	Net Force Trig	1	Given two arrows representing the magnitude and direction of forces on an object, choose the correct expression of the net force in the x or y direction
Rotational kinematics	Rot Kin	1	Using rotational kinematics equations to solve for α, θ, ω, or t
Rotational unit conversion	Rot Unit Conv	1 & 2	Convert between radians and degrees Convert between revolutions and radians Convert from revolution per time to radians per time
Linear vs rotational motion	Lin vs Rot	1 & 2	Convert between $a_{\text{tangential}}$, α, and r; Convert between $v_{\text{tangential}}$, ω, and r
Work done by a constant force	Work	1 & 2	Work done on an object by a constant force is positive, negative, or zero; Work done on an object by a constant force is greater than, less than, or equal to XX Joules
Vector components using trig	Vector Comp	2	Identify the correct trigonometric expression for the x or y component of a vector (and angle) displayed on an x-y plot.
Vector addition	Vec Add	2	Given two vectors on a grid or ijk notation, identify the sum of the resultant vector.

categories studied. Brief descriptions of each practice category are provided in Table II and example items of each category are presented in Appendix B. The practice categories were constructed over a period of several semesters using an evidence-based process described by Mikula and Heckler [8].

The design of study 1 included all students assigned the practice categories indicated in Table II. All students in both courses were assigned the categories either 2, 3, or 4 times spaced throughout the semester. The timing of the assignments is detailed in the figures on the results (Sec. III A).

Study 2 included two conditions, as shown in Table III. All students received full training in some practice categories and partial training in other categories, depending on the condition. As described in Sec. I A, partial training consisted of two practice trials starting at least two weeks after the relevant unit and full training consisted of four practice trials beginning just after instruction starts. The last two practice trials for the full training coincided with the two partial training trials as shown in Fig. 1. We assigned both full and partial training to students to help counterbalance total training time (across categories) and to allow for a within-student analysis of the effect of training. Students were selected in one of the two conditions based on their instructor's lecture section. If an instructor taught two lecture sections, the students in the first lecture section received condition 1 categories while the students in the other lecture section received condition 2 categories. This was done to help control for instructor-level factors that may affect student improvement on the assignments.

B. Performance data

There is some freedom and ambiguity in choosing which performance parameters to use when considering useful outcome measures related to accuracy and speed. Choices include the raw performance data collected during the mastery assignments, consisting of the number of questions attempted (Q_{att}) to achieve mastery, the number of questions answered correctly (Q_{cor}), and the total completion time (T) [or the logarithm of the completion time $\log_{10}(T)$] for each student and each category in each assignment. Other possible choices derived from these measurements include the proportion correct ($Q_{\text{cor}}/Q_{\text{att}}$), the mean response time per question (T/Q_{att}). The choice of variable to investigate certainly depends on the research questions of interest, and it can also depend on the design of the practice assignments. To provide a sense of how these

TABLE III. Study 2 design indicating the full and partial training practice categories for each condition.

Condition	N	Full training	Partial training
1	Algebra: 173 Calculus: 186	Work Vector components Rot. unit conversion	Vector addition Linear vs rot. motion
2	Algebra: 159 Calculus: 129	Vector addition Linear vs Rot. motion	Work Vector components Rot. unit conversion

variables are empirically related, we present correlation tables for a few representative categories in Tables VIII to X in Appendix A. Below, we describe which variables were used in this paper.

For an outcome measure related to accuracy, we chose different measures for study 1 and study 2. As mentioned earlier, study 1 had varying numbers of correct questions in a row required for mastery of different practice categories in different assignments. Therefore, for study 1, we used the proportion correct (Q_{cor}/Q_{att}) as the measure of accuracy. For study 2, the number of questions correct required in a row was the same for all practice categories and all assignments, therefore we chose to use the total number of questions attempted Q_{att} to achieve mastery, which we also view as an informative measure when the study design allows for it. It also complements the results of study 1. There were several additional reasons for this choice. First, considering between Q_{att} and Q_{cor}, we found that for any given category, Q_{att} was essentially empirically interchangeable with Q_{cor}, because their correlations with each other were typically around $r = 0.97$ (see Tables VIII–X in Appendix A). Second, we chose Q_{att} because it is readily interpretable and is relevant to the assumed general student goal of minimizing the number of attempted questions needed to complete the assignment. Further, we chose Q_{att} instead of the proportion correct Q_{cor}/Q_{att}, because the latter can be ambiguous in terms of the number of questions answered in the mastery practice context. To understand this, consider that the goal of mastery practice is to achieve a set number, say 4 questions correct in a row. This could be achieved, for example, by answering 4 out of 6 questions correctly or 8 out of 12 questions correctly, given that only the last four questions were answered correctly in a row in both cases. In the second case, Q_{att} is twice as large, but the proportion correct is the same.

For an outcome measure related to speed, again we chose different and complementary measures for study 1 and study 2. For study 1, we chose the mean response time per question T/Q_{att} for a given category, where T is the total time to complete the practice category since the required number of questions correct in a row for a given category varied by assignment. For study 2, we chose the logarithm of the completion time $\log_{10}(T)$, which we viewed as the preferred measure to use when possible. Specifically, the completion time is readily interpretable, and we suppose that students are more likely to aim to minimize their total time spent on a STEM fluency assignment instead of minimizing how fast they can answer individual questions, which is related but not identical in a mastery assignment. We use $\log_{10}(T)$ for purposes of better data analysis. Specifically, the completion time distributions for each category were skewed right, with skewnesses ranging from 3 to 5, which is outside the range of validity for normally distributed residuals in our model fits. The transformation to $\log_{10}(T)$ results in a more symmetric, normal-like distribution, resulting in better model fits. Note that we will keep in mind the fact that the logarithm is a nonlinear function, because this affects our interpretation of the model results, especially when considering interactions.

A common feature of timing data includes right-skewed distributions with long tails. The very long tails in our context indicate that some students took a very long time to answer the questions (on the order of thousands of seconds per question, i.e., 15 to 20 min per question). One possible explanation for these long response times is that sometimes students leave the assignment open on their computer when they were not actively working on the assignment but are instead engaged with other activities. To account for this tail, we trimmed the top 2.5% of our timing data from each practice category trial, removing that entire entry for that category trial for those students, including the questions attempted, questions correctly answered, their response time, and completion time [33]. For example, a student could have data removed for category 1 and practice trial 2 but still have data included for category 1 and practice trial 3, and for all trials of another category. An examination of the time distributions revealed that this cutoff effectively removed the extreme times. The time distributions also revealed that some students had average response times shorter than 1 would reasonably expect to read the entire question and determine an answer (typically less than 4 to 6 s per question). We believe a portion of these response times were due to students randomly guessing, and since these guesses are not an accurate portrayal of how long it takes a student to complete these assignments, we removed the bottom 2.5% of our timing data. We will assume that the data is "missing at random," and the linear mixed modeling used for our analysis is valid for such missing data [34]. This assumption seems reasonable for the upper time cutoff, but the lower timed cutoff may introduce some bias, and this is a potential threat to a small bias in the results of this study. However, to check, we reran the models in study 2, including the data below the lower cutoff, and found no qualitative differences (e.g., in significance in slopes and interactions) and very minor quantitative differences from the results reported here.

Procrastination was measured using *submission time*, which we define as the time between when the completed STEM fluency assignment was submitted by the student and the assignment deadline time. The smaller the submission time, the more the student procrastinated. For each student, an average submission time was calculated by adding the submission times for each assignment and dividing by the total number of assignments submitted. Note that other aspects of submission times were investigated in more detail by the authors in a previous study [27].

To measure the mindset of our students, we administered a student's personal physics mindset beliefs survey. This mindset survey was only administered to the calculus

students because a different motivational survey study was being conducted at the same time in the algebra course. In order to determine the predictive power of mindset, the survey was administered at the beginning of the first mastery assignment. The personal physics mindset scale contained four items pertaining to the student's beliefs about physics intelligence (see Appendix A), with two items from the Dweck mindset scale [35], and two additional items more aimed at understanding physics and problem-solving in physics. For our dataset, the scale had a level of reliability of Cronbach's $\alpha = 0.8$ indicating that the scale was internally consistent. A high score on the personal physics mindset corresponds to a student with a fixed mindset while a person with a low personal physics mindset score is considered to have a growth mindset.

C. Models

In order to provide more precise quantitative answers to our research questions, in study 2, we employed linear mixed modeling to build and analyze statistical models of the data. At a broad level, these models are somewhat similar to ordinary multiple regression models in that we will estimate regression coefficients to test and quantify relationships, but because of the relatively complex structure of the data in study 2, linear mixed modeling is needed [34]. For example, not only are there within-student repeated measured (practice trials), but students and practice categories are cross-classified clustered data, namely, practice categories are clustered within students and vice versa. Linear mixed modeling also allows for missing data in the cases where some of the students missed some of the assignments, or, as discussed earlier, are trimmed out because of outlier response times on specific categories in an assignment. Students were modeled as a random effect to account for expected variation in student abilities. To account for variation in performance by practice category, we modeled practice categories as fixed factors since there were only five categories, which is too small to reliably model as a random factor. Below, we describe the models used in study 2. Note that the ordering and numbering of the models were chosen to improve the clarity and comparability of the data tables summarizing the results of all of the models.

1. Model 1

Model 1 investigates RQ3 and RQ4. This first model compares the effects of full to partial training on the number of questions attempted during training. To compare full vs partial training, only data from practice trials 3 and 4 were used for this model.

$$(Q_{\text{att}})_{ijk} = \gamma_{00} + \gamma_{\text{cat},i}(\text{Category})_i + \gamma_{\text{trial4}} * (\text{Trial})_k + \gamma_{\text{train}} * (\text{Train})_{ij} + \gamma_{\text{interaction}}(\text{Trial})_k \times (\text{Train})_{ij} + u_{0j} + r_{ijk}. \quad (1)$$

In this model, $(Q_{\text{att}})_{ijk}$ corresponds to the number of questions attempted for category i, student j, and practice trial k. The coefficient γ_{00} represents the overall mean intercept of our model and $\gamma_{\text{cat},i}$ represents the fixed-effect estimate for the average questions attempted for all students in category i. Note that "work" is the reference category. The variable $(\text{Trial})_{\boldsymbol{k}}$ is coded as 0 for practice trial 3 and 1 for practice trial 4 (see Fig. 1). The coefficient γ_{trial4} represents the mean difference in Q_{att} between practice trials 3 and 4. The variable $(\text{Train})_{\boldsymbol{ij}}$ is coded 0 for student j receiving partial training on category i and 1 for full training. Therefore γ_{train} represents the mean effect of full vs partial training on Q_{att} in practice trails 3 and 4. Note that each student receives partial training in some categories and full in others, depending on their training condition (Table III). For model 1, the coefficient $\gamma_{\text{interaction}}$ represents the estimate for the practice trial-by-training interaction and indicates the extent to which full training affects the change in performance between practice trails 3 and 4 compared to partial training. The term u_{0j} represents the random effect of student j, and r_{ijk} is the random error associated with trial k, student j, and category i.

2. Model 2

The second model investigates RQ1 and RQ2. Specifically, this model tested how a student's prior preparation, as measured by ACT math score, is related to the number of questions attempted during practice and how Q_{att} evolves throughout the semester for the full-trained students receiving four practice trials for every category. To investigate evolution during the semester in model 2, we analyzed only the performance for the first and last practice trials for the fully trained students who were assigned four practice trials.

$$(Q_{\text{att}})_{ijk} = \gamma_{00} + \gamma_{\text{cat},i}(\text{Category})_i + \gamma_{\text{init fin}} * (\text{Init Fin})_k + \gamma_{\text{ACT}} * (\text{ACT})_j + \gamma_{\text{interaction}}(\text{Init Fin})_k \times (\text{ACT})_j + u_{0j} + r_{ijk} \quad (2)$$

Several terms are the same as for model 1 described above. The variable $(\text{Init Fin})_k$ indicates the first or last practice trial on the practice category i for student j. This variable is coded as either 0 for initial practice or 1 for final practice. Therefore $\gamma_{\text{init fin}}$ represents the change in Q_{att} from the initial to final practice trials. The variable $(\text{ACT})_j$ the mean-centered ACT math score of student j. Therefore, γ_{ACT} represents an estimate of the extent to which Q_{att} depends on the ACT score. For model 2 the coefficient $\gamma_{\text{interaction}}$ is an estimate of the $(\text{Init Fin})_k \times (\text{ACT})_j$ interaction, namely how the change in performance depends on ACT score.

3. Model 3

To further study RQ2, model 3 investigates the extent to which student performance (i.e., Q_{att}) and the evolution of student performance on mastery assignments is related to gender, which here is considered only as a binary term (male or female) since this is how it was recorded in the university database from which the gender was reported for this study.

$$\begin{aligned}(Q_{\text{att}})_{ijk} = {} & \gamma_{00} + \gamma_{\text{cat},i}(\text{Category})_i + \gamma_{\text{init fin}} * (\text{Init Fin})_k \\ & + \gamma_{\text{ACT}} * (\text{ACT})_j + \gamma_{\text{female}} * (\text{Gender})_j \\ & + \gamma_{\text{interaction}}(\text{Init Fin})_k \times (\text{Gender})_j + u_{0j} + r_{ijk}\end{aligned} \tag{3}$$

Several terms are the same as for model 2 described above. The variable $(\text{Gender})_j$ is coded as 0 for male and 1 for female for student j. For model 3, the coefficient $\gamma_{\text{interaction}}$ is an estimate of the $(\text{Init Fin})_k \times (\text{Gender})_{ij}$ interaction, namely how the change in initial-to-final performance depends on gender.

4. Model 4

Model 4 investigates the extent to which student performance and the evolution of student performance on mastery assignments is related to procrastination, as measured by submission time, as defined in the previous subsection. This model also investigates RQ2.

$$\begin{aligned}(Q_{\text{att}})_{ijk} = {} & \gamma_{00} + \gamma_{\text{cat,i}}(\text{Category})_i + \gamma_{\text{init fin}} * (\text{Init Fin})_k \\ & + \gamma_{\text{ACT}} * (\text{ACT})_j + \gamma_{\text{subT}} * (\text{Sub Time})_{ij} \\ & + \gamma_{\text{interaction}}(\text{Init Fin})_k \times (\text{Sub Time})_{ij} + u_{0j} + r_{ijk}\end{aligned} \tag{4}$$

Several terms are the same as for model 2 described above. The variable $(\text{Sub Time})_{ij}$ is the amount of time, in hours, before the deadline that the assignment with category i was submitted by student j. Therefore, low submission times mean the student procrastinated since the student submitted a small amount of time before the deadline. For model 4, the coefficient $\gamma_{\text{interaction}}$ is an estimate of the $(\text{Init Fin})_{ij} \times (\text{Sub Time})_{ij}$ interaction, namely how the change in performance depends on submission time.

5. Model 5

Model 5 tested how a student's personal physics mindset impacted the student's performance and evolution of performance, which is relevant for RQ2. Because we are specifically interested in determining if the mindset is predictive for students who struggle, for model 5, we limit the population to students who have mean scores above the median proportion correct for the initial practice trials because a student who is struggling on the assignments is less accurate and will answer *more* questions.

$$\begin{aligned}(Q_{\text{att}})_{ijk} = {} & \gamma_{00} + \gamma_{\text{cat},i}(\text{Category})_i + \gamma_{\text{init fin}} * (\text{Init Fin})_k \\ & + \gamma_{\text{ACT}} * (\text{ACT})_j + \gamma_{\text{mind}} * (\text{Mindset})_j \\ & + \gamma_{\text{interaction}}(\text{Init Fin})_k \times (\text{Mindset})_j + u_{0j} + r_{ijk}\end{aligned} \tag{5}$$

Several terms are the same as for model 2 described above. The variable $(\text{Mindset})_j$ is the physics mindset score for student j. For model 5, the coefficient $\gamma_{\text{interaction}}$ is an estimate of the $(\text{Init Fin})_k \times (\text{Mindset})_j$ interaction, namely how the change in performance depends on physics mindset.

6. Models 6–10

For models 6–10, the equations are identical to models 1–5 except that the outcome variable $(Q_{\text{att}})_{ijk}$ is replaced with the outcome variable $\log_{10}(T)_{ijk}$, corresponding to the logarithm base 10 of the total time that student j spends on category i in practice trial k.

III. RESULTS

A. Study 1—Trends in the evolution of performance

We begin by presenting graphical representations of the evolution of mean response time per question and proportion correct for several categories practiced 3–4 times spaced throughout the semester (see Fig. 2). Let us discuss several observations prompted by Fig. 2. The first is that there are notable decreases in time per question and increases in proportion correct for each category. Second, there is a significant variation in response time and accuracy between categories, ranging by an order of magnitude in time and a factor of 2 in accuracy. Third, for the calculus-based course, the accuracies show signs of plateauing for all but the lowest accuracy category. Likely related to this observation is that these same categories also show signs of reaching an asymptote in the decrease of time per question. Essentially, students are maintaining the same accuracy and time per question over multiple practices. However, for the algebra-based course, there is far less of an indication of plateauing in accuracy for any category. Rather there are signs of continued substantial improvement in accuracy, perhaps indicating the benefit of more practice. The same trend appears for the time per question, namely that there are no signs of reaching an asymptote in decrease in time per question. These observations are relevant to RQ4 regarding the benefits of continued practice, which appears to depend on the initial level of performance and the population. Finally, the calculus-based students tend to be a little faster and more accurate than the algebra-based students. This difference at least qualitatively

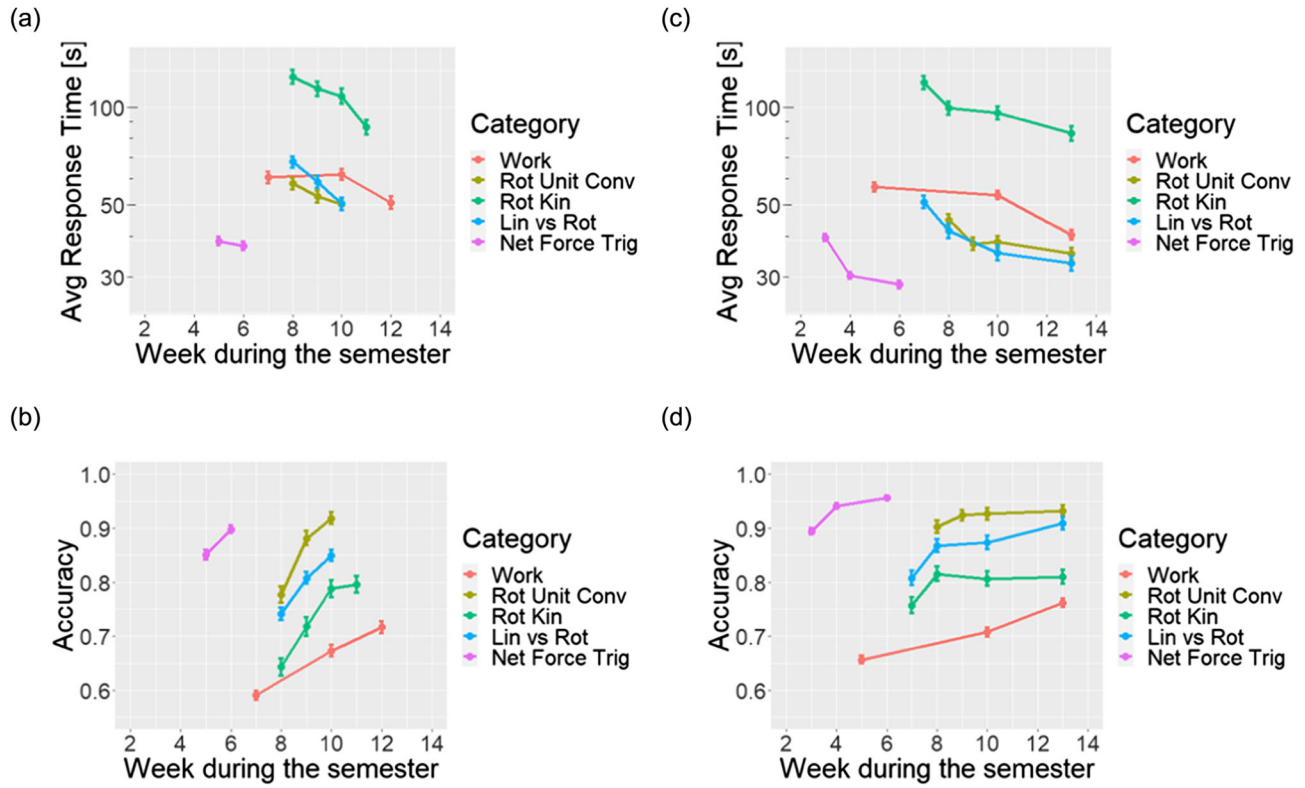

FIG. 2. Study 1 data (a) algebra students' mean response time spent for multiple essential skills categories at multiple practices times. (b) algebra students' mean accuracy for multiple essential skills categories at multiple practice times. (c) Calculus students' mean response time spent for multiple essential skills categories at multiple practice times. (d) Calculus students' mean accuracy for multiple essential skills categories at multiple practice times724 [Note: lines are drawn only to help pair data points from the same category. Some categories included practice sessions in between the initial and final practice].

is consistent with observations of plateauing for the former but not the latter.

To gain more insight into subpopulations of students, Figs. 3 and 4 display the evolution from first to last practice trial grouped by students receiving different final grades. There are several notable features of these graphs. First, all groups are improving on average, and there is some indication for some practice categories that performance gaps are narrowing. However, overall, it appears that students receiving an A grade began and ended as the fastest and most accurate students and those receiving a D grade began and ended as the slowest and least accurate. Note that using the final grade is *post hoc* grouping variable rather than a predictive one. To get a better sense of whether initially less-prepared students evolve differently than better-prepared students, in study 2, we will investigate ACT math score as a predictive covariate using model 2.

B. Study 2—Factors predicting evolution: Trends and quantitative models

To determine the potential predictive power of the various factors discussed in the introduction on the evolution of performance, Tables IV–VII present the results of Models 1–10, and Figures 5–8 provide visual information that provides more insight into the model results. Overall, and consistent with the results of study 1, these results show clear decreases in the number of questions attempted to achieve mastery Q_{att} and completion time T between the first and last practice trials for both courses. Below, we discuss the results for each factor.

1. Partial vs full practice

The results from model 1 indicate that, compared to partial training, the full training condition had a small but statistically significant beneficial impact on the number of questions attempted and the total time to mastery, as measured by γ_{train} in Tables IV–VII (see also Figs. 5 and 6). Recall that γ_{train} represents the estimated mean difference in performance between conditions in trials 3 and 4, where students in the full training had two training practices before trials 3 and 4, and students in the partial training did not have any training practice before trials 3 and 4. Specifically, on average, the students in the full training condition completed trials 3 and 4 with 2.69 fewer questions attempted and about 103 s faster (per trial) in the algebra-based course and 1.37 fewer questions

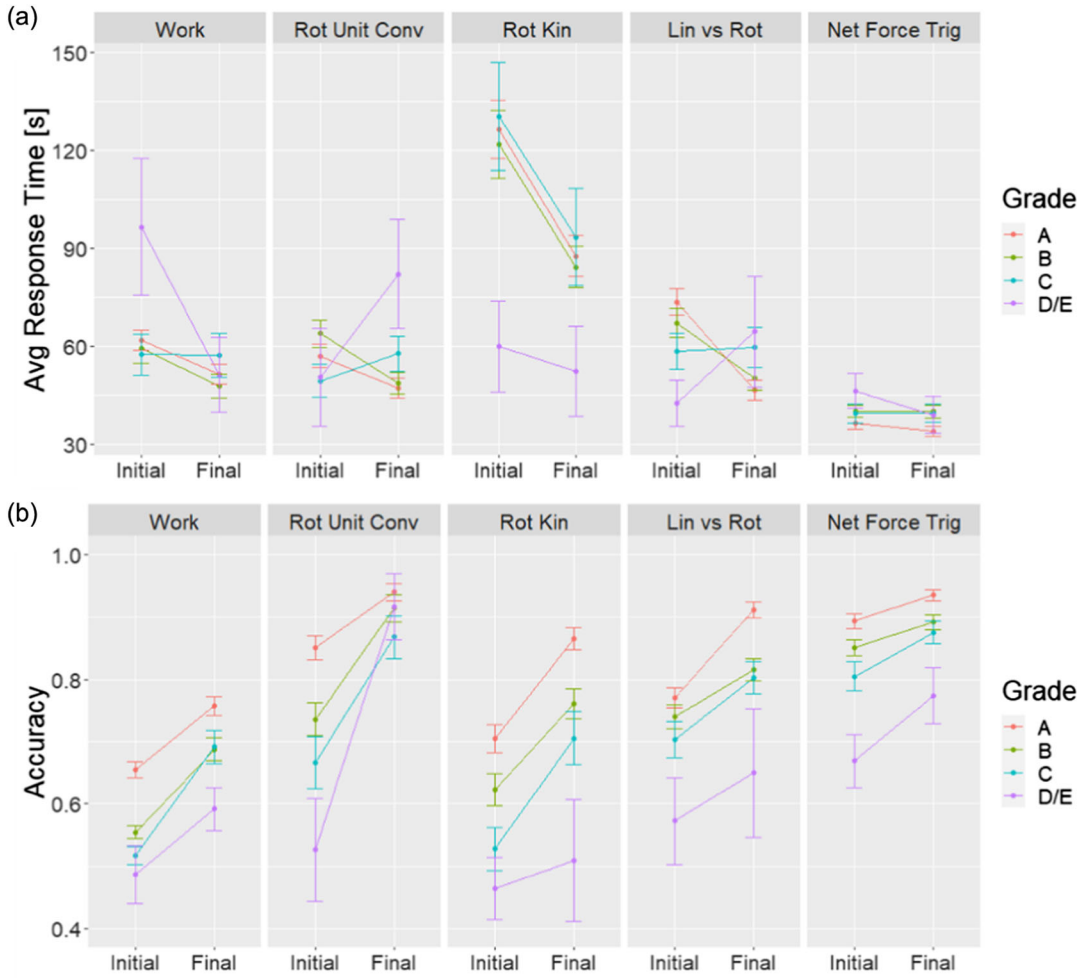

FIG. 3. Study 1 (a) Algebra-based students' average response time spent for each category. The average response times are measured the first- and last-time students saw the categories and subset by the course grade the students earned in the course. (b) Algebra-based students mean accuracy for each category. The mean accuracy is measured the first- and last-time students saw the categories and subset by the course grade the students earned in the course.

attempted and about 140 s faster in the calculus-based course. To get a sense of effect size, consider that the residual standard deviation σ_r for Q_{att} is about 17 and 9 for the algebra-based course and calculus-based courses, respectively. In terms of speed, the full training results in a roughly 5–10 s per question increase in speed compared to the partial training. In the four tables, there was only one significant interaction estimate ($\gamma_{\text{interaction}}$), indicating that the evolution from practice trial 3 to 4 was the same for partial and full training, with the exception of the total time for the calculus-based students where the partially trained students sped up a little faster than the fully trained students.

In summary, students who practiced directly after instruction and the week after became faster and more accurate than the students who only practiced several weeks after instruction. The fact that the fully trained students were faster and more accurate on the third practice trial indicates that STEM fluency practice benefitted students above and beyond benefits from other components of the course, such as the lectures, group-work recitations, and homework assignments. We can see this effect by looking at the third practice trial and comparing students without any STEM fluency training before the third practice trial (i.e., "partially trained students") to the "fully trained" students who completed two STEM fluency assignments before the third practice trial. For many of the practice categories, the "partial practice" students never caught up to the fully trained students. Figures 5 and 6 display the mean total times and Q_{att} by category for each practice trial for these two groups, and graphically confirms the model 1 findings, and can provide deeper insights into the results. For example, for the rotational unit conversion practice category, students in the full training condition were faster and had lower Q_{att} than the partial training in practice trial 3 for both courses. Differences between conditions on practice trial 3 vary by category, though it is not clear why there is such variation.

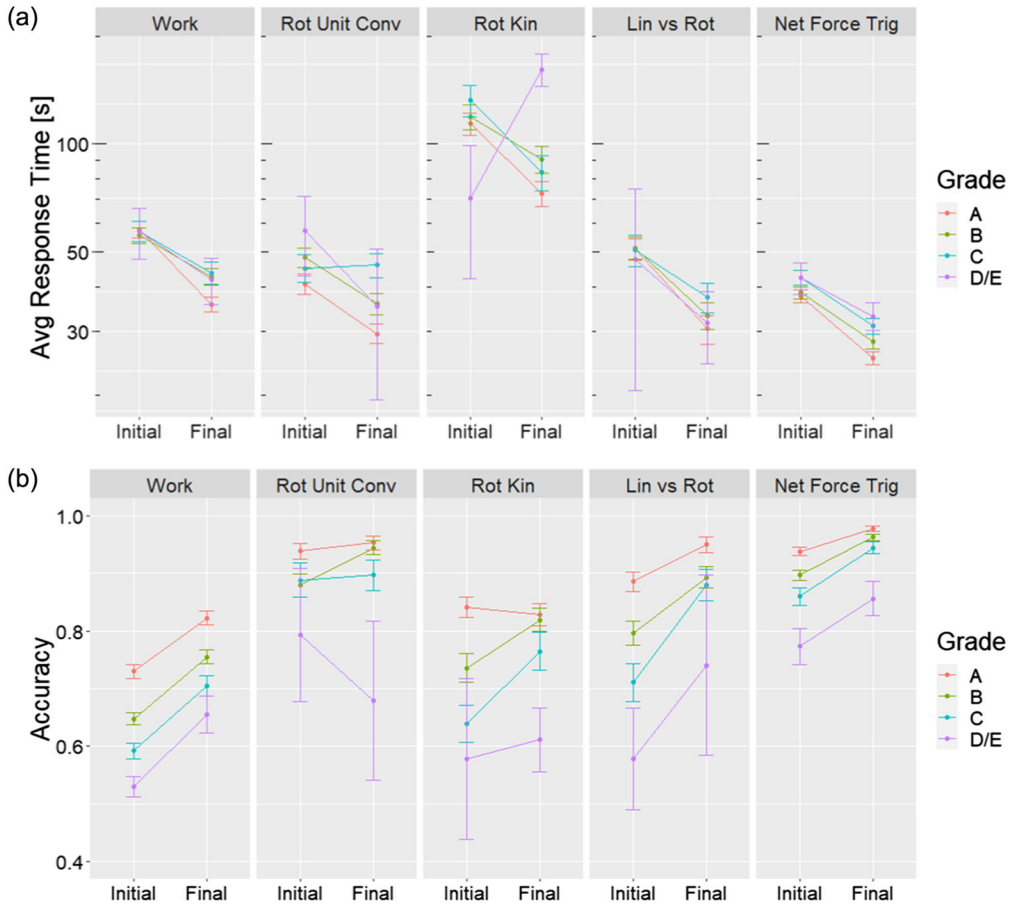

FIG. 4. Study 1 (a) calculus-based students mean response time spent for each category. The mean response times are measured the first- and last-time students saw the categories and subset by the course grade the students earned in the course. (b) Calculus-based students' average accuracy for each category. The mean accuracy is measured the first- and last-time students saw the categories and subset by the course grade the students earned in the course.

TABLE IV. Study 2 model coefficients for algebra physics classes with the number of questions attempted (Q_{att}) for a given practice category as the outcome variable. Values in parentheses are standard errors. Note the ACT scores are mean centered. Bolded numbers are significant at the $p < 0.01$ level (and often significantly lower). An * denotes the cell is significant at the $p < 0.05$ level.

	Algebra students			
	Model 1	Model 2	Model 3	Model 4
Fixed effects	Full and partial training	Full training	Full training	Full training
γ_{00}	**24.12(1.05)**	**23.80(1.14)**	**20.93(1.49)**	**23.70(1.46)**
$\gamma_{\text{cat,Trig}}$	**−9.05(0.99)**	**−5.22(1.34)**	**−5.31(1.34)**	**−5.28(1.34)**
$\gamma_{\text{cat,VecAdd}}$	**−15.28(0.99)**	**−11.39(1.50)**	**−11.52(1.50)**	**−11.32(1.51)**
$\gamma_{\text{cat,RotConv}}$	**−14.34(0.98)**	**−13.14(1.33)**	**−13.15(1.33)**	**−13.17(1.33)**
$\gamma_{\text{cat,LinRot}}$	**−12.34(0.98)**	**−10.91(1.48)**	**−11.08(1.48)**	**−10.90(1.49)**
$\gamma_{\text{init fin}}$	⋯	**−6.21(0.86)**	**−4.40(1.49)**	**−4.98(1.43)**
γ_{train}	**−2.69(0.88)**	⋯	⋯	⋯
γ_{trial4}	−1.12 (0.88)	⋯	⋯	⋯
γ_{ACT}	⋯	**−1.40(0.17)**	**−0.99(0.14)**	**−1.02(0.14)**
γ_{female}	⋯	⋯	**4.83(1.47)**	⋯
γ_{subT}	⋯	⋯	⋯	0.01 (0.02)
$\gamma_{\text{interaction}}$	−0.28 (1.24)	**0.79(0.22)**	−3.54 (1.81)*	−0.04 (0.02)
Random effects				
σ_{u_i}	10.18	5.42	5.29	5.57
σ_r	16.96	15.04	15.09	15.07

TABLE V. Study 2 model coefficients for calculus physics classes with the number of questions attempted (Q_{att}) for a given practice category as the outcome variable. Values in parentheses are standard errors. Note the ACT scores are mean-centered. For model 5, only students scoring below the median on the initial trials are included. Bolded numbers are significant at the $p < 0.01$ (and often significantly lower). An * denotes the cell is significant at the $p < 0.05$ level.

	Calculus students				
	Model 1	Model 2	Model 3	Model 4	Model 5
Fixed effects	Full and partial training	Full training	Full training	Full training	Full training above median
γ_{00}	**16.23(0.53)**	**21.02(0.88)**	**19.96(0.95)**	**20.95(1.08)**	**22.15(1.05)**
$\gamma_{cat,Trig}$	**−5.07(0.53)**	**−3.61(1.07)**	**−3.65(1.07)**	**−3.63(1.07)**	**−4.84(1.30)**
$\gamma_{cat,VecAdd}$	**−10.36(0.54)**	**−12.23(1.23)**	**−12.28(1.24)**	**−12.26(1.24)**	**−14.66(1.67)**
$\gamma_{cat,RotConv}$	**−10.28(0.53)**	**−12.14(1.05)**	**−12.19(1.05)**	**−12.18(1.06)**	**−13.67(1.28)**
$\gamma_{cat,LinRot}$	**−8.07(0.54)**	**−9.82(1.25)**	**−9.92(1.25)**	**−9.85(1.26)**	**−10.96(1.68)**
$\gamma_{init\ fin}$	...	**−4.49(0.73)**	**−3.35(0.86)**	**−3.73(1.08)**	**−4.35(1.00)**
γ_{train}	**−1.37(0.48)**	...	...	...	...
γ_{trial4}	0.38 (0.48)	...	...	...	...
γ_{ACT}	...	**−1.18(0.15)**	**−0.81(0.11)**	**−0.78(0.11)**	**−0.95(0.14)**
γ_{female}	...	...	2.73 (1.19)*	...	...
γ_{subT}	...	...	...	−0.01 (0.02)	...
$\gamma_{mindset}$	...	...	...	...	−0.27 (1.33)
$\gamma_{interaction}$	−1.36 (0.67)	**0.75(0.19)**	−1.74 (1.54)	0.00 (0.02)	0.17 (2.06)
Random effects					
σ_{u_i}	3.29	3.50	3.38	3.46	3.62
σ_r	8.92	12.33	12.41	12.42	13.59

TABLE VI. Study 2 model coefficients for algebra physics classes with total Log base 10 completion time for a given practice category as the outcome variable. Values in parentheses are standard errors. Note the ACT scores are mean-centered. Bolded numbers are significant at the $p < 0.01$ (and often significantly lower). An * denotes the cell is significant at the $p < 0.05$ level.

	Algebra students			
	Model 6	Model 7	Model 8	Model 9
Fixed effects	Full and partial training	Full training	Full training	Full training
γ_{00}	**2.74(0.02)**	**2.86(0.02)**	**2.83(0.03)**	**2.84(0.03)**
$\gamma_{cat,Trig}$	**−0.42(0.02)**	**−0.32(0.03)**	**−0.32(0.03)**	**−0.32(0.03)**
$\gamma_{cat,VecAdd}$	**−0.59(0.02)**	**−0.48(0.03)**	**−0.48(0.03)**	**−0.47(0.03)**
$\gamma_{cat,RotConv}$	**−0.33(0.02)**	**−0.31(0.03)**	**−0.31(0.03)**	**−0.31(0.03)**
$\gamma_{cat,LinRot}$	**−0.40(0.02)**	**−0.36(0.03)**	**−0.36(0.03)**	**−0.36(0.03)**
$\gamma_{init\ fin}$	...	**−0.35(0.02)**	**−0.36(0.03)**	**−0.25(0.03)**
γ_{train}	**−0.09(0.01)**	...	...	...
γ_{trial4}	**−0.05(0.01)**	...	...	...
γ_{ACT}	...	**−0.02(0.0)**	**−0.02(0.00)**	**−0.02(0.0)**
γ_{female}	...	...	0.04 (0.03)	...
γ_{subT}	...	...	...	0.0 (0.0)
$\gamma_{interaction}$	−0.01 (0.02)	−0.0 (0.0)	0.02 (0.04)	**−0.002(0.0004)**
Random effects				
σ_{u_i}	0.17	0.11	0.10	0.11
σ_r	0.28	0.30	0.30	0.30

2. ACT math score

There are three main results from models 2 and 7, which investigate how ACT score might be related to the evolution of accuracy and speed. The first, perhaps as expected, is that the number of attempted questions Q_{att} significantly decreases with increasing ACT score, as estimated by γ_{ACT}. It is important to keep in mind that model 2 uses the mean-centered ACT score, thus a score of zero is at the mean (see Table I), and scores below the mean change the sign of the effect. Therefore, for algebra-based

TABLE VII. Study 2 model coefficients for calculus physics classes with Log base 10 completion time for a given practice category as the outcome variable. Values in parentheses are standard errors. Note the ACT scores are mean centered. For model 10, only students scoring below the median on the initial trials are included. Bolded numbers are significant at the $p < 0.01$ (and often significantly lower). An * denotes the cell is significant at the $p < 0.05$ level.

	Calculus students				
	Model 6	Model 7	Model 8	Model 9	Model 10
Fixed effects	Full and partial training	Full training	Full training	Full training	Full training above median
γ_{00}	**2.68(0.02)**	**2.74(0.02)**	**2.72(0.02)**	**2.77(0.03)**	**2.74(0.02)**
$\gamma_{\text{cat,Trig}}$	**−0.44(0.02)**	**−0.30(0.02)**	**−0.30(0.02)**	**−0.30(0.02)**	**−0.32(0.03)**
$\gamma_{\text{cat,VecAdd}}$	**−0.67(0.02)**	**−0.58(0.03)**	**−0.58(0.03)**	**−0.58(0.03)**	**−0.58(0.04)**
$\gamma_{\text{cat,RotConv}}$	**−0.36(0.02)**	**−0.32(0.02)**	**−0.32(0.02)**	**−0.32(0.02)**	**−0.33(0.03)**
$\gamma_{\text{cat,LinRot}}$	**−0.33(0.02)**	**−0.29(0.03)**	**−0.29(0.03)**	**−0.29(0.03)**	**−0.31(0.04)**
$\gamma_{\text{init fin}}$	…	**−0.23(0.02)**	**−0.23(0.02)**	**−0.22(0.02)**	**−0.23(0.02)**
γ_{train}	**−0.15(0.02)**	…	…	…	…
γ_{trial4}	0.03 (0.01)*	…	…	…	…
γ_{ACT}	…	**−0.02(0.00)**	**−0.02(0.00)**	**−0.02(0.00)**	**−0.02(0.003)**
γ_{female}	…	…	0.04 (0.03)	…	…
γ_{subT}	…	…	…	−0.00087(0.00043)*	…
γ_{mindset}	…	…	…	…	−0.03 (0.03)
$\gamma_{\text{interaction}}$	**0.06(0.02)**	0.0 (0.0)	0.01 (0.04)	0.0(0.0)	−0.03 (0.05)
Random effects					
σ_{u_i}	0.12	0.12	0.11	0.11	0.12
σ_r	0.28	0.28	0.28	0.28	0.29

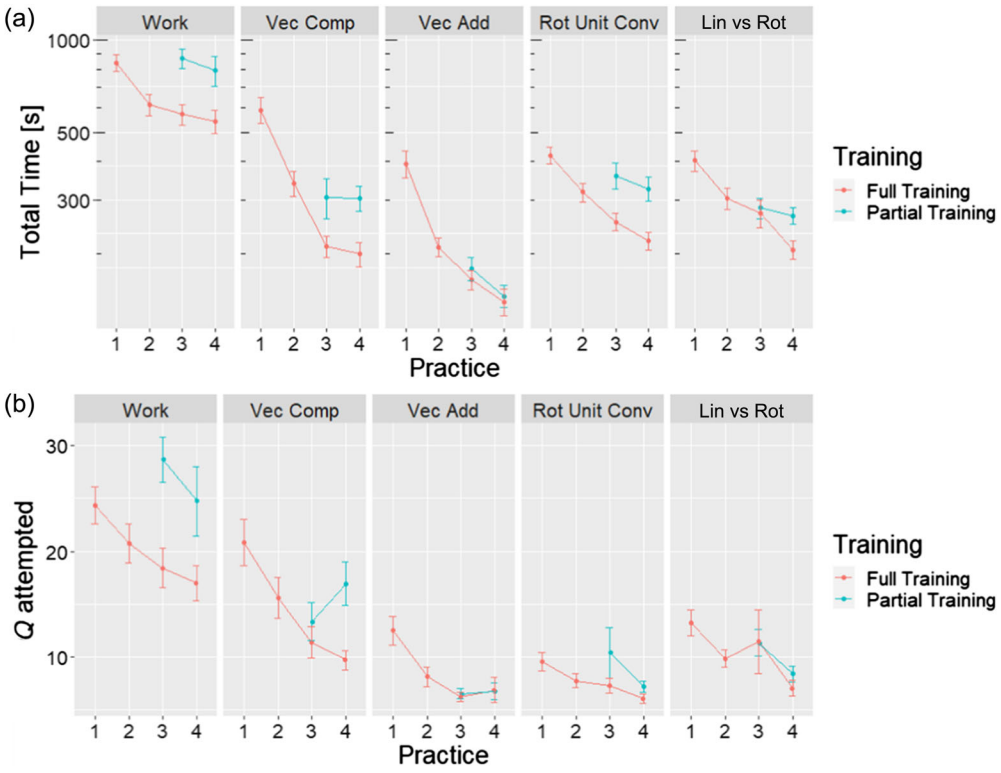

FIG. 5. Study 2 total completion time and questions attempted for each training group evaluated at the same time for the algebra physics class.

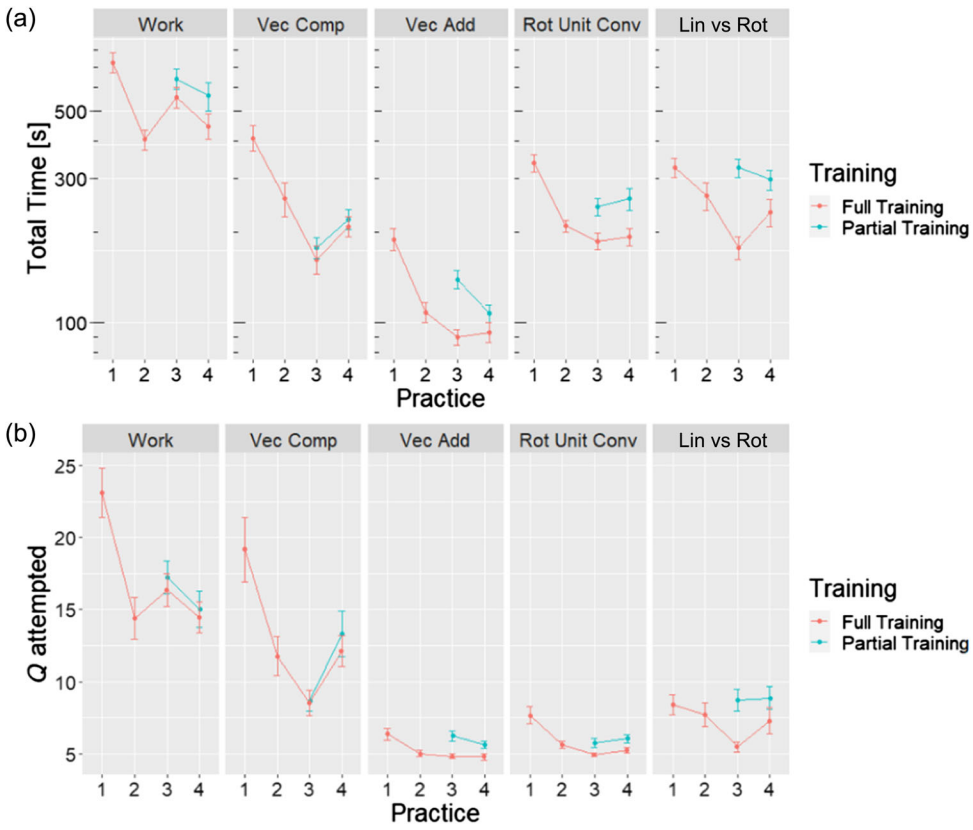

FIG. 6. Study 2 (a) total completion time and (b) questions attempted for each training group evaluated at the same time for the calculus physics class.

students, for example, the estimate is $\gamma_{\mathrm{ACT}} = 1.4$ meaning that for every ACT point above the mean, Q_{att} decreases by 1.4 questions, and for every ACT point below the mean, Q_{att} increases by 1.4 questions. Again, to get a sense of effect size, every ACT point changes $Q_{\mathrm{att}}T$ by about 0.1 residual standard deviation.

The results of model 7 indicate that the completion time also significantly decreases with increasing ACT score. For example, for algebra-based students with a mean ACT score, the time to complete a category is $T = 724$ s on average. But for a student with an ACT score one point above the mean, $T = 692$ s, or 32 s faster. The results for Q_{att} and T could naturally be related. One hint toward this possibility is the fact that while ACT score is moderately to weakly correlated with both ($r \approx 0.1$–0.3) for any practice category, it is not significantly correlated with the time per question ($r < 0.10$), see Tables VIII to X in Appendix A.

Finally, there is a significant ACT-by-practice trial interaction for Q_{att}, as indicated by the estimate of the interaction term in Model 2. For example, following Table IV, consider that Q_{att} decreases by 6.2 questions from the initial to final practice trials for students with the mean ACT score in the Algebra-based course. However, the interaction implies that this decrease is moderated by the ACT math score such that for students with an ACT score one point above the mean the decrease narrows to 5.4 questions and for students with one point below the ACT, the decrease widens to 7.0. In other words, students with lower ACT scores improve more in terms of questions attempted than students with high ACT scores. Roughly, the same effects and magnitude of the effects on Q_{att} and completion time are found for calculus-based students. Figure 7 graphically displays the interaction effect on Q_{att} for both courses.

For both courses, there was no interaction in terms of the logarithm of completion time. However, as mentioned earlier, when interpreting these results, there is an important point to keep in mind due to the non-linearity of the logarithmic function: while there is no interaction in logarithmic time, effectively there could still be an interaction in linear time, so caution must be used in interpreting the result of the model. For example, consider the estimates for model 6 for algebra-based students in Table VI. Students scoring one point above or one point below the mean ACT completed a category on average in about 692 or 758 s, respectively, a difference of 66 s. In the final practice trial, those times become 309 and 338 s, respectively, a difference of 29 s. In other words, logarithmically, there was no interaction (no closing of the gap), but linearly, the gap was cut in half, reduced by 37 s, thus indicating some level of interaction between ACT score and

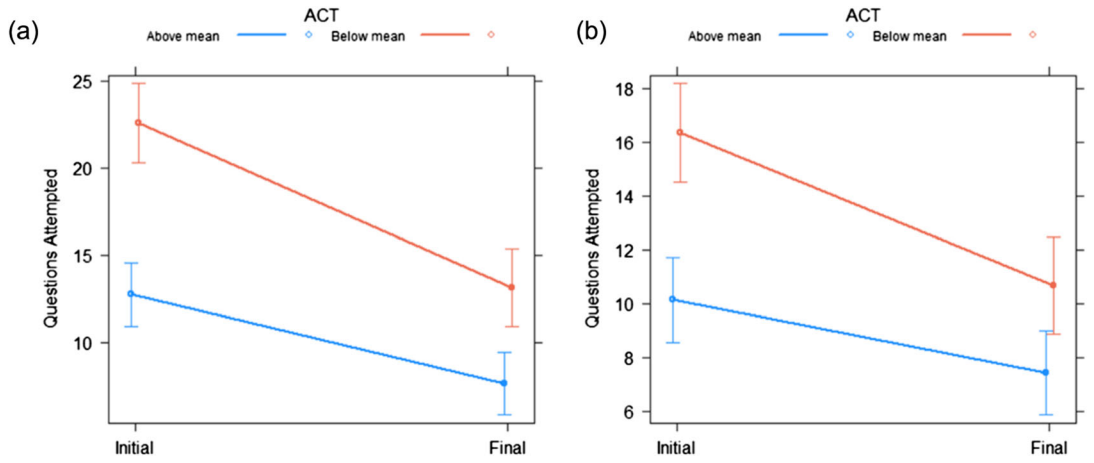

FIG. 7. Study 2 estimated the marginal mean number of questions attempted at the initial and final practice trials across all practice categories, split by students whose ACT score was above or below the mean for (a) algebra-based students and (b) calculus-based students. Error bars are 1 SE. The lines are drawn only to help pair data points from the same category. Some categories included practice sessions in between the initial and final practice.

improvement in completion time, with the time gap closing between high and low ACT students.

3. Gender

The factor of gender, as reported in the university database, was also found to be significant, even accounting for ACT math score, as estimated by γ_{female} in model 3. Specifically, the results of model 3 in Tables IV and V indicate that on average for the first practice trial, the number of attempted questions Q_{att} is 4.8 questions higher for women than for men for the algebra-based course for a practice category, and 2.7 questions higher for women in the calculus-based course. Given that Q_{att} is around 20 questions in the first practice trial, this indicates a difference of about 15%–25% in questions attempted between genders. However, the results of model 8 indicate that there were no such significant differences in completion time. This implies that women tend to answer the questions slightly more rapidly.

There is also a significant gender-by-practice trial interaction for the algebra-based students in model 3. An inspection of Table IV indicates that Q_{att} decreased by 4.4 questions between practice trials 1 and 4 for male students, however, for female students Q_{att} decreased by 7.9 questions. In short, though female students began with a significantly higher Q_{att} than males, this gap essentially reduced to zero by the fourth practice trial. This interaction was not significant for calculus-based students, but the point estimate for the interaction trended in a similar way. There

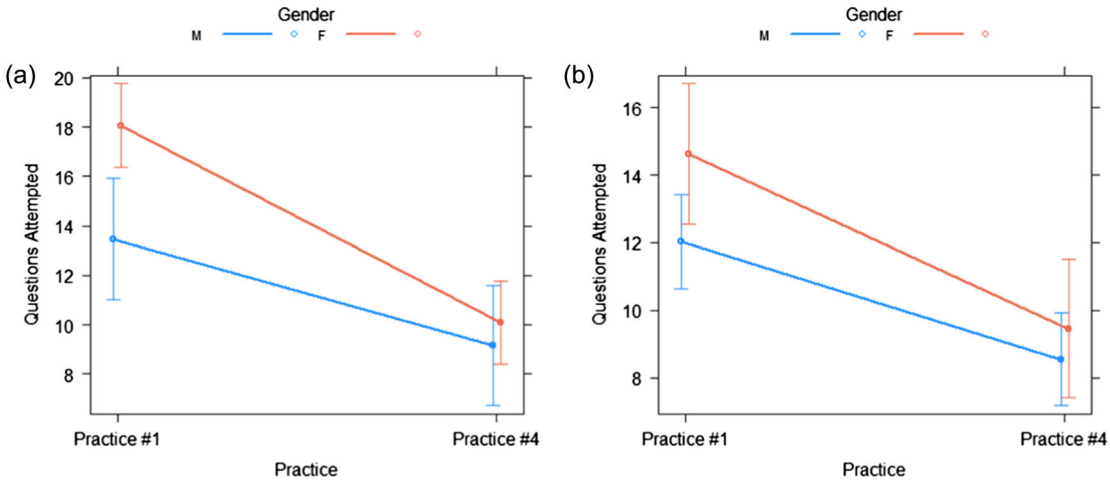

FIG. 8. Study 2 estimated the marginal mean number of questions attempted at the initial and final practice trials across all practice categories, split by gender for (a) algebra-based students and (b) calculus-based students. Error bars are 1 SE. The lines are drawn only to help pair data points from the same category. Note that some categories included practice sessions in between the initial and final practice.

was no interaction effect on completion time. Figure 8 graphically displays the interaction effect for both courses.

4. Submission time

The results of models 4 and 9 indicate that procrastination, as measured by submission time, does not predict any differences in the number of attempted questions Q_{att}. Specifically, in Tables IV and V, for model 4, the estimates for γ_{SubT} and $\gamma_{\text{interaction}}$ are not significantly different from zero. However, procrastination does predict differences in how students evolve during their practice in terms of completion time, even accounting for ACT scores. Recall that submission time is measured in hours and indicates the amount of time before the deadline the assignment was submitted. While Table VI indicates that there was no relation between submission time and completion time (i.e., the time it takes to complete the assignment) for the first practice trial for students in the algebra-based course, there was a submission time-by-trial interaction predicting completion time. Specifically, on average, in their first practice trial, all students completed one practice category in about 692 s regardless of submission time. Students with the mean ACT score who procrastinated and submitted near the deadline decreased their completion time to about 389 s on average per practice category by the last practice trial, but students with the mean ACT score who submitted their assignments 72 h (on average) before it was due decreased their completion time to about 279 s per category. That difference in the decrease of 110 s between procrastinators and nonprocrastinators is substantial considering the original completion time. In short, students in the algebra-based course who procrastinate improved their completion times significantly less than students who do not procrastinate, even controlling for ACT scores.

For the calculus-based students, Table VII indicated that submission time does predict an overall significant difference in the logarithm of completion time. For example, for the first practice trial, students with the mean ACT score submitting near the deadline on average completed a practice category in about 589 s, but students with a mean ACT score who submitted their assignments 72 h (on average) before it was due completed a category in about 510 s. In other words, students in the calculus-based course who procrastinate complete each category about 79 s slower than students who do not procrastinate, even controlling for ACT scores. As discussed earlier with the ACT scores, while the interaction term of the logarithm of time is not significant, there is still a reduction in the completion time gap to 48 s between the last practice trial for procrastinators (355 s) and for nonprocrastinators (307 s).

5. Mindset

The results of models 5 and 10 indicate that the mindset scores do not predict performance or evolution of performance, as estimated by γ_{mindset}. As stated earlier, the analysis of models 5 and 10 only includes those students scoring above the median proportion correct on the initial practice trials.

IV. DISCUSSION AND CONCLUSION

In a series of studies, we have characterized the evolution of accuracy and speed of students responding to questions on online mastery-based assignments repeated throughout the semester covering basic introductory physics skills. To summarize, let us discuss how our results address our research questions, starting with RQ1 and RQ4. Following expected patterns of accuracy and response time learning curves typically found in studies of learning ([18]), both algebra and calculus students on average systematically improved their accuracy and decreased their response time per question on a range of physics topics and categories over multiple repeated spaced practices throughout the semester. While calculus students were slightly faster and more accurate than the algebra students, the STEM fluency assignments were still effective and beneficial to both classroom populations in improving student fluency and performance on the assignments. We noticed the differences in the shapes of the accuracy and response time curves in study 1 reached saturation for some categories (i.e., the student's speed and accuracy plateaued after two trials) while other practice categories, like work, did not reach saturation even after full training. On average, this saturation happened mainly in the calculus-based population, suggesting that for some of the categories studied, one could decrease the number of practice trials without sacrificing gains in performance.

Considering RQ2, we found that several student-level factors were associated with differences in initial performance and evolution. Perhaps most notably, while students with low ACT math scores were initially less accurate and slower than students with high ACT scores, this gap decreased by the final practice trial. This suggests that STEM fluency mastery assignments are a useful tool for instructors to help students refresh important basic skills, and it helps students with lower levels of preparation to catch up.

Regarding differences between genders, women are initially spending the same time as men on assignments but are answering more questions to achieve mastery, even controlling for ACT scores. By the final practice trial, both men and women increased in accuracy, but for algebra-based students, the gap closed: women improved more than men, such that they both ended up with similar accuracies. For the calculus-based students, there was no significant decrease in the gap. For both courses, men and women decreased the time they spend on the assignments by about the same amount.

Combining the results from this study (that the performance gap between men and women is diminished after spaced practices) and previous work (that women complete

more STEM fluency assignments [23] and procrastinate less on the assignments than men [27]) all controlling for ACT score, it leads us to wonder why we are seeing a distinct difference in study habits and evolution of performance between women and men, namely that women have initially poorer performance but appear to be working harder and catching up. This suggests a potentially interesting line of inquiry for future work to present a coherent framework to explain these differences between the two groups.

We were somewhat surprised to find that student mindset is not predictive of the number of questions attempted or completion time for students who struggle initially with the assignments (RQ2). Despite mixed results reported on mindset [28–30], we were expecting that mindset would predict performance on mastery assignments. Specifically, we were expecting to see that students who initially had relatively low accuracy on the mastery assignments but had a growth mindset would improve more than students with a fixed mindset because they would be more resilient to failure, but this was not the case. Ours was a superficial investigation of the factor of mindset, and before we can make any firm conclusions about whether or not mindset is important in this context, further research is needed to perhaps more carefully measure this construct (beyond a four-item scale) and devise a more careful theoretical argument identifying which behaviors it might influence.

In terms of submission time, controlling for ACT scores, students who did not procrastinate reduced their assignment completion time more than students who did procrastinate. This is true even though the number of questions attempted to achieve mastery did not depend on procrastination. In other words, the nonprocrastinators sped up or became more fluent than the procrastinators. We hypothesize that this could occur because the nonprocrastinators are more committed to learning, resulting in their performance improving. Another possibility is that the procrastinators have put themselves in a stressful environment by submitting the assignments late, which results in a lack of improvement in performance. Future work could look further at the individual question level of heavy procrastinators to see how the evolution of the response time per question varies in the final hours before the deadline, seeing if heavy procrastinators are exhibiting rapid guess behavior, meaning they are not rapidly responding to questions before time expires.

Finally, models 1 and 6 in study 2 provided evidence that on average across several practice categories, STEM fluency practice improves both accuracy and speed beyond any gains accrued from traditional lectures (RQ3). Figures 5 and 6 reveal that this added benefit depends on the category and the course, though it is not immediately evident why there is such variation. These results along with the overall STEM fluency learning curves that match general expectations from past learning research help to further validate the STEM fluency materials and design [8] as a useful learning tool, though naturally, the effectiveness is modulated by numerous factors such as those studied here. Example factors of interest for future studies include the timing of spaced and interleaved practice, which has been studied in numerous contexts and could be applied to mastery learning of basic skills in an introductory physics context [35–39].

There are a few limitations to keep in mind when interpreting this work. First, there may be selection effects in our results since about half of our students consented to participate in our study, and our population sample is skewed toward higher mean grades (0.2–0.4 standard deviations) and ACT scores (less than 0.1 standard deviation). Therefore, the sample is slightly underrepresenting low-performing students. Further, study 2 took place during the COVID-19 pandemic, which could have impacted a variety of factors in our study. For example, it could have impacted the motivation of students completing these assignments, though our observations indicate similar overall trends in improvement in accuracy and speed in both studies. Additionally, the studies here were for brief (15–30 min) online STEM fluency assignments that are designed to be low-level difficulty practice sessions. Because these assignments are distinct from traditional, "back-of-the-textbook" homework questions, this impacts our ability to generalize this work to other assignments. Future work could look at how practicing with STEM fluency assignments might impact students' exam performance on problems that cover topics practiced in the STEM fluency assignments. Future work should also investigate if STEM fluency practices help students on homework topics similar to the topics covered in the STEM fluency assignments. This would allow us to discuss the impact of STEM fluency assignments on other important components of the course.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under DUE IUSE Grant No. 1914709.

APPENDIX A: CORRELATION TABLES

Correlation tables for a selection of STEM fluency categories and various variables measured in the study.

TABLE VIII. Correlation Table for the Work done by a constant force category.

	T	$\log_{10}(T)$	T/Q_{att}	Q_{cor}	Q_{att}	Q_{cor}/Q_{att}	Course GPA	Submission time
$\log_{10}(T)$	0.84							
T/Q_{att}	0.46	0.48						
Q_{cor}	0.41	0.46	−0.29					
Q_{att}	0.39	0.43	−0.30	0.98				
Q_{cor}/Q_{att}	−0.29	−0.40	0.34	−0.47	−0.52			
Course GPA	−0.02	−0.04	0.07	−0.09	−0.11	0.19		
Submission time	−0.07	−0.08	−0.03	0.01	0.00	0.06	0.25	
MC ACT score	−0.11	−0.14	0.03	−0.16	−0.19	0.21	0.33	−0.01

TABLE IX. Correlation Table for the Rotational Unit Conversion category.

	T	$\log_{10}(T)$	T/Q_{att}	Q_{cor}	Q_{att}	Q_{cor}/Q_{att}	Course GPA	Submission time
$\log_{10}(T)$	0.85							
T/Q_{att}	0.49	0.62						
Q_{cor}	0.64	0.44	−0.17					
Q_{att}	0.63	0.40	−0.16	0.97				
Q_{cor}/Q_{att}	−0.44	−0.48	0.21	−0.48	−0.47			
Course GPA	−0.08	−0.10	0.05	−0.11	−0.12	0.24		
Submission time	−0.07	−0.13	−0.12	0.02	0.02	0.06	0.24	
MC ACT score	−0.24	−0.28	−0.09	−0.20	−0.20	0.26	0.36	−0.04

TABLE X. Correlation table for the vector components—trig category.

	T	$\log_{10}(T)$	T/Q_{att}	Q_{cor}	Q_{att}	Q_{cor}/Q_{att}	Course GPA	Submission time
$\log_{10}(T)$	0.82							
T/Q_{att}	0.30	0.44						
Q_{cor}	0.66	0.63	−0.20					
Q_{att}	0.68	0.62	−0.20	0.97				
Q_{cor}/Q_{att}	−0.52	−0.65	0.19	−0.56	−0.61			
Course GPA	−0.13	−0.16	0.05	−0.16	−0.16	0.21		
Submission time	−0.03	−0.03	−0.08	0.03	0.04	−0.01	0.23	
MC ACT score	−0.27	−0.29	0.03	−0.27	−0.29	0.35	0.32	−0.06

APPENDIX B: STUDENT'S PERSONAL PHYSICS MINDSET BELIEFS

1	2	3	4	5	6
Strongly disagree	Disagree	Somewhat disagree	Somewhat agree	Agree	Strongly agree

- You have a certain amount of physics intelligence, and you can't really do much to change it.
- Only very few specially qualified people are capable of really understanding physics.
- No matter how much physics intelligence you have, you can always change it quite a bit (R).
- Anyone can become good at solving physics problems through hard work (R).

APPENDIX C: SAMPLE QUESTIONS FOR EACH ES CATEGORY

1. Work done by a constant force sample question

A car is moving at a constant speed along a flat horizontal street as shown in Fig. 9. The wind exerts a constant force $\vec{F}$ on the car while the car moves an amount $\Delta\vec{x}$ as shown above. Determine whether the work done on the car by the force of the wind is positive, negative, or zero.

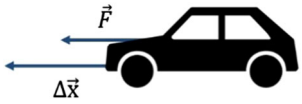

FIG. 9. Image to accompany work done by a constant force sample question.

2. Rotational unit conversion sample question

Shawn White won the gold medal in the 2011 Winter X-Games for completing a "Double McTwist" where he completed two 1260° turns. How many radians did he complete after the first 1260° turn?

(a) $1260° \cdot \frac{\pi \text{ rad}}{180°} = 7\pi$ radians
(b) $1260° \cdot \frac{\pi \text{ rad}}{180°} = 7$ radians
(c) $1260° \cdot \frac{\pi \text{ rad}}{360°} = 3.5\pi$ radians
(d) $1260° \cdot \frac{\pi \text{ rad}}{360°} = 3.5$ radians

3. Vector addition sample question

Note that the practice items varied between ijk arrow representations for both the question and answer formats.

Consider two vectors:

$$\vec{A} = -2\hat{i} - 3\hat{j},$$
$$\vec{B} = -1\hat{j}.$$

Which answer choice (presented in Fig. 10) represents the vector sum $\vec{A} + \vec{B}$?

4. Linear vs rotational motion sample question

A disk has an angular acceleration of $\alpha = 17\ \frac{\text{rad}}{\text{s}^2}$. At what radius is the tangential acceleration equal to $0.22\ \frac{\text{m}}{\text{s}^2}$?

(a) 0.012 m
(b) 3.7 m
(c) 77 m

5. Rotational kinematics sample question

A windmill on a farm rotates at a constant speed and completes one-half of a rotation in 0.5 s. What is its rotation speed?

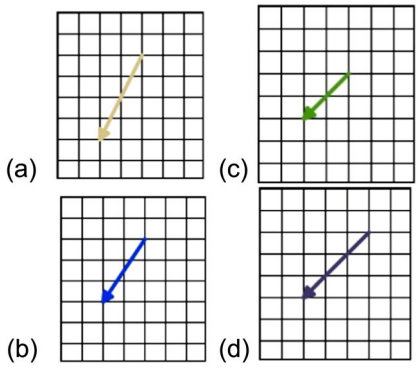

FIG. 10. Vector addition sample question answer choices.

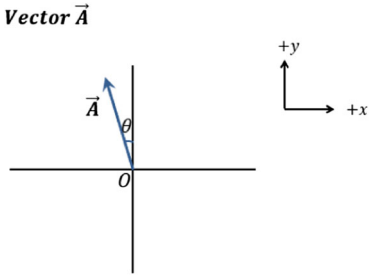

FIG. 11. Image to accompany the vector components sample question.

(a) 6.28 rad/s
(b) 1 rad/s
(c) 0.5 rad/s
(d) 3.14 rad/s

6. Vector components sample question

Note that the angles were reference from either the ±x or ±y axis, and this varied between items.

Vector $\vec{A}$ is shown in Fig. 11. Which of the following options represents A_x, the x component of vector $\vec{A}$?

(a) $-A\sin\theta$
(b) $A\sin\theta$
(c) $-A\cos\theta$
(d) $A\cos\theta$
(e) $A\cos(90° - \theta)$

7. Net force trig sample question

Note that the angles were reference from either the ±x or ±y axis, and this varied between items.

Two forces $\overrightarrow{F_1}$ and $\overrightarrow{F_2}$ are shown in Fig. 12 in free-body diagram form. Which of the options below represents the y component of the net force?

(a) $\Sigma F_y = F_1\sin\theta + F_2\sin\phi$
(b) $\Sigma F_y = F_1\cos\theta - F_2\sin\phi$
(c) $\Sigma F_y = -F_1\sin\theta + F_2\cos\phi$
(d) $\Sigma F_y = F_1\cos\theta - F_2\cos\phi$

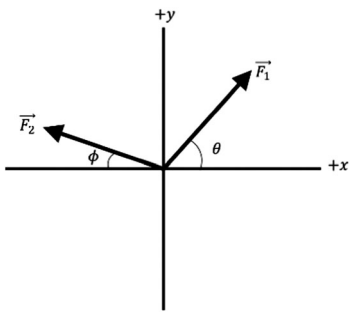

FIG. 12. Image to accompany the net force trig sample question.

[1] B. S. Bloom, Learning for mastery. Instruction and curriculum. Regional Education Laboratory for the Carolinas and Virginia, Topical Papers and Reprints, Number 1, Eval. Comment **1**, n2 (1968), https://files.eric.ed.gov/fulltext/ED053419.pdf.

[2] C. L. C. Kulik, J. A. Kulik, and R. L. Bangert-Drowns, Effectiveness of mastery learning programs: A meta-analysis, Rev. Educ. Res. **60**, 265 (1990).

[3] N. Schroeder, G. Gladding, B. Gutmann, and T. Stelzer, Narrated animated solution videos in a mastery setting, Phys. Rev. ST Phys. Educ. Res. **11**, 010103 (2015).

[4] G. Gladding, B. Gutmann, N. Schroeder, and T. Stelzer, Clinical study of student learning using mastery style versus immediate feedback online activities, Phys. Rev. ST Phys. Educ. Res. **11**, 010114 (2015).

[5] B. Gutmann, G. Gladding, M. Lundsgaard, and T. Stelzer, Mastery-style homework exercises in introductory physics courses: Implementation matters, Phys. Rev. Phys. Educ. Res. **14**, 010128 (2018).

[6] P. W. Wambugu and J. M. Changeiywo, Effects of mastery learning approach on secondary school students' physics achievement, Eurasia J. Math. Sci. Technol. Educ. **4**, 293 (2008).

[7] M. Guthrie and Z. Chen, Comparing student behavior in mastery and conventional style online physics homework, available at SSRN: https://ssrn.com/abstract=3522737 or 10.2139/ssrn.3522737 (2020).

[8] B. D. Mikula and A. F. Heckler, Framework and implementation for improving physics essential skills via computer-based practice: Vector math, Phys. Rev. Phys. Educ. Res. **13**, 010122 (2017).

[9] K. R. Koedinger, J. L. Booth, and D. Klahr, Instructional complexity and the science to constrain it, Science **342**, 935 (2013).

[10] P. C. Kyllonen and J. Zu, Use of response time for measuring cognitive ability, J. Intell. **4**, 14 (2016).

[11] Y. H. Lee and H. Chen, A review of recent response-time analyses in educational testing, Psychol. Test Assess. Model. **53**, 359 (2011), https://psycnet.apa.org/record/2011-28090-006.

[12] O. Wilhelm and R. Schulze, The relation of speeded and unspeeded reasoning with mental speed, Intelligence **30**, 537 (2002).

[13] S. L. Wise, D. A. Pastor, and X. J. Kong, Correlates of rapid-guessing behavior in low-stakes testing: Implications for test development and measurement practice, Appl. Meas. Educ. **22**, 185 (2009).

[14] D. J. Palazzo, Y. J. Lee, R. Warnakulasooriya, and D. E. Pritchard, Patterns, correlates, and reduction of homework copying, Phys. Rev. ST Phys. Educ. Res. **6**, 010104 (2010).

[15] Z. Chen, M. Xu, G. Garrido, and M. W. Guthrie, Relationship between students' online learning behavior and course performance: What contextual information matters?, Phys. Rev. Phys. Educ. Res. **16**, 010138 (2020).

[16] K. E. DeLeeuw and R. E. Mayer, A comparison of three measures of cognitive load: Evidence for separable measures of intrinsic, extraneous, and germane load, J. Educ. Psychol. **100**, 223 (2008).

[17] W. Huang, P. Eades, and S. H. Hong, Measuring effectiveness of graph visualizations: A cognitive load perspective, Inf. Vis. **8**, 139 (2009).

[18] K. R. Koedinger, A. T. Corbett, and C. Perfetti, The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning, Cogn. Sci. **36**, 757 (2012).

[19] C. Lin, S. Shen, and M. Chi, Incorporating student response time and tutor instructional interventions into student modeling, in *Proceedings of the 2016 Conference on user modeling adaptation and personalization* (2016), pp. 157–161, https://dl.acm.org/doi/pdf/10.1145/2930238.2930291.

[20] A. Hellas, P. Ihantola, A. Petersen, V. V. Ajanovski, M. Gutica, T. Hynninen *et al.*, Predicting academic performance: A systematic literature review, in *Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education* (2018), pp. 175–199, https://dl.acm.org/doi/pdf/10.1145/3293881.3295783.

[21] I. A. Chounta and P. Carvalho, Will time tell? Exploring the relationship between step duration and student performance, in *Proceedings of 13th International Conference of the Learning Sciences (ICLS), London, UK* (2018), https://repository.isls.org/bitstream/1/539/1/217.pdf.

[22] J. Scharfen, J. M. Peters, and H. Holling, Retest effects in cognitive ability tests: A meta-analysis, Intelligence **67**, 44 (2018).

[23] A. B. Simmons and A. F. Heckler, Grades, grade component weighting, and demographic disparities in introductory physics, Phys. Rev. Phys. Educ. Res. **16**(2), 020125 (2020).

[24] S. Salehi, E. Burkholder, G. P. Lepage, S. Pollock, and C. Wieman, Demographic gaps or preparation gaps?: The large impact of incoming preparation on performance of students in introductory physics, Phys. Rev. Phys. Educ. Res. **15**, 020114 (2019).

[25] L. E. Kost, S. J. Pollock, and N. D. Finkelstein, Characterizing the gender gap in introductory physics, Phys. Rev. ST Phys. Educ. Res. **5**, 010101 (2009).

[26] Z. Felker and Z. Chen, The impact of extra credit incentives on students' work habits when completing online homework assignments, presented at PER Conf. 2020, virtual conference, 10.1119/perc.2020.pr.Felker.

[27] M. Nieberding and A. F. Heckler, Patterns in assignment submission times: Procrastination, gender, grades, and grade components, Phys. Rev. Phys. Educ. Res. **17**, 013106 (2021).

[28] C. S. Dweck and D. S. Yeager, Mindsets: A view from two eras, Perspect. Psychol. Sci. **14**, 481 (2019).

[29] A. Rattan, K. Savani, D. Chugh, and C. S. Dweck, Leveraging mindsets to promote academic achievement: Policy recommendations, Perspect. Psychol. Sci. **10**, 721 (2015).

[30] A. P. Burgoyne, D. Z. Hambrick, and B. N. Macnamara, How firm are the foundations of mind-set theory? The claims appear stronger than the evidence, Psychol. Sci. **31**, 258 (2020).

[31] V. F. Sisk, A. P. Burgoyne, J. Sun, J. L. Butler, and B. N. Macnamara, To what extent and under which circumstances

are growth mind-sets important to academic achievement? Two meta-analyses, Psychol. Sci. **29**, 549 (2018).

[32] D. S. Yeager, P. Hanselman, G. M. Walton, J. S. Murray, R. Crosnoe, C. Muller *et al.*, A national experiment reveals where a growth mindset improves achievement, Nature (London) **573**, 364 (2019).

[33] H. J. Keselman, R. R. Wilcox, A. R. Othman, and K. Fradette, Trimming, transforming statistics, and bootstrapping: Circumventing the biasing effects of heterescedasticity and nonnormality, J. Mod. Appl. Stat. Methods **1**, 38 (2002).

[34] B. T. West, K. B. Welch, and A. T. Galecki, *Linear Mixed Models: A Practical Guide Using Statistical Software* (Chapman and Hall/CRC, Boca Raton, 2006).

[35] C. S. Dweck, *Self-Theories: Their Role in Motivation, Personality, and Development* (Psychology Press, New York, 1999).

[36] B. Settles and B. Meeder, A trainable spaced repetition model for language learning, in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2016), pp. 1848–1858, https://aclanthology.org/P16-1174.pdf.

[37] T. Nakata, Effects of expanding and equal spacing on second language vocabulary learning: Does gradually increasing spacing increase vocabulary learning?, Stud. Second Lang. Acquis. **37**, 677 (2015).

[38] S. K. Kim and S. Webb, The effects of spaced practice on second language learning: A meta-analysis, Lang. Learn. **72**, 269 (2022).

[39] D. Rohrer and M. K. Hartwig, Unanswered questions about spaced interleaved mathematics practice, J. Appl. Res. Mem. Cogn. **9**, 433 (2020).