Factors affecting learning of vector math from computer-based practice: Feedback complexity and prior knowledge

Andrew F. Heckler and Brendon D. Mikula

Department of Physics, Ohio State University, 191 W. Woodruff Avenue, Columbus, Ohio 43210, USA

(Received 29 June 2015; published 9 June 2016)

In experiments including over 450 university-level students, we studied the effectiveness and time efficiency of several levels of feedback complexity in simple, computer-based training utilizing static question sequences. The learning domain was simple vector math, an essential skill in introductory physics. In a unique full factorial design, we studied the relative effects of "knowledge of correct response" feedback and "elaborated feedback" (i.e., a general explanation) both separately and together. A number of other factors were analyzed, including training time, physics course grade, prior knowledge of vector math, and student beliefs about both their proficiency in and the importance of vector math. We hypothesize a simple model predicting how the effectiveness of feedback depends on prior knowledge, and the results confirm this knowledge-by-treatment interaction. Most notably, elaborated feedback is the most effective feedback, especially for students with low prior knowledge and low course grade. In contrast, knowledge of correct response feedback was less effective for low-performing students, and including both kinds of feedback did not significantly improve performance compared to elaborated feedback alone. Further, while elaborated feedback resulted in higher scores, the learning rate was at best only marginally higher because the training time was slightly longer. Training time data revealed that students spent significantly more time on the elaborated feedback after answering a training question incorrectly. Finally, we found that training improved student self-reported proficiency and that belief in the importance of the learned domain improved the effectiveness of training. Overall, we found that computer based training with static question sequences and immediate elaborated feedback in the form of simple and general explanations can be an effective way to improve student performance on a physics essential skill, especially for less prepared and low-performing students.

DOI: 10.1103/PhysRevPhysEducRes.12.010134

I. INTRODUCTION

Computer-based instruction has proven to be effective in a variety of contexts, with effect sizes typically ranging from 0.3 to 0.5 [1–4], yet there is still a wide variation in effectiveness, particularly in studies investigating different methods of feedback. Numerous reviews have noted—perhaps to be expected—that there is no single best prescription for feedback, rather the effectiveness of feedback depends on a number of potentially interacting factors. Examples of these factors include the type and level of knowledge or skill to be learned, the type (e.g., complexity) of feedback, timing of the feedback, prior knowledge of topic, student achievement, correctness of and confidence in responses, interest in topic, self-efficacy, and other attitudinal factors; for reviews, see Refs. [4–9].

The sheer complexity of numerous potentially interacting factors compels one to simplify the issue by focusing on specific cases that are general enough to be applicable to some important educational contexts, but constrained enough that results will be generalizable within those contexts. Therefore, in this study, we will focus on a specific but important physics learning domain, namely, basic vector math skills essential for success in an introductory physics course. We will also focus on factors important for practical instructional implementation.

Essentially, we are addressing three questions: (i) Which level of feedback complexity is most effective and time efficient for this domain? (ii) To what extent do the factors of prior knowledge, student achievement, and attitudinal factors such as perceived importance of the topic for the course interact with the effectiveness of different levels of feedback for this domain? (iii) What insights do the timing data (during training) tell us about how students are using the most effective feedback levels?

The format of the learning task was a self-contained online computer-based practice module on vector math skills for university-level introductory physics. Naturally, determining optimal feedback for this domain is relevant and applicable to common and current educational contexts. Furthermore, this study is consistent with several reviews which made calls for research expanding and exploring the space of possible factors influencing (and

improving) the effectiveness of feedback [8–10]. The factors we are investigating can help to gain insight into both practical instructional questions and provide empirical results to advance theoretical models of computer-based instruction.

A. Varying complexity of feedback

While there are numerous ways to characterize feedback complexity, we will focus on a framework commonly used by many researchers in which feedback can be broken into five types of increasing complexity (cf. Ref. [11]):

(1) No feedback.
(2) Knowledge of results (KR): user is informed of whether the answer is right or wrong.
(3) Knowledge of correct response (KCR): user is informed of the correct answer.
(4) Knowledge of correct response and elaborated feedback (KCR + EF): user is informed of the correct answer and further information such as an explanation is given.
(5) Knowledge of correct response and interactive teaching (KCR + IT).

A number of studies have investigated the relative effectiveness of various combinations of these feedback types, with most finding that higher complexity results in higher learning (see, e.g., Refs. [4,5,11–13]), while some others find no such differences [14,15]. Once again, these studies represent a significant variation in other factors, so it is difficult to generalize.

In a recent meta-analysis, Van der Kleij *et al.* [4] found that EF tends to be more beneficial than KR and KCR, especially for higher level learning outcomes (e.g., applying knowledge rather than recalling it). However, there are two issues with these results. First, the meta-analysis reveals a wide range of outcomes, both positive and negative depending on a variety of factors; and while one might reasonably expect to find similar results, it is not clear that they apply to the domain and context studied here. Second, rather than explicitly varying learning outcomes in this study, we will investigate the pedagogically important issue of how variations in prior knowledge may affect learning.

In this study, we are interested in learning via simple computer feedback (and not interactive teaching), and as such we will only consider complexity types 2, 3 and 4. In order to better understand any possible interaction between KCR and EF, we will study the effectiveness of KCR and EF separately and together—always providing KR—in a specific science, technology, engineering, and mathematics (STEM) essential skill learning domain. Note that in this study, EF will be in the form of general topic-contingent explanations describing the relevant procedural rule in the format of words, equations, and/or diagrams. The EF is general and topic contingent in the sense that it is not specific to the numerical values of a specific problem, but rather can be applied to a class of similar problems.

There are both practical and theoretical reasons for exploring a full factorial analysis of KCR and EF feedback methods, which—to our knowledge—has not been previously attempted. From the practical standpoint, when designing computer-based instruction modules for a particular learning domain, one must make decisions (either implicitly or explicitly) about the feedback provided to the learner. Naturally, an empirical determination of the most effective and/or efficient feedback is the most useful.

From a theoretical standpoint, one can test predictions of simple models, though in some cases existing models produce conflicting predictions. Empirical results can help to resolve ambiguities. First, let us start with predictions from the simple model that prior knowledge plays a critical role in determining the effectiveness of different levels of feedback; see, e.g., Refs. [10,16–18]. Specifically, learners with low prior knowledge tend to need more complex feedback and support while learners with high prior knowledge do not benefit from such extra feedback, e.g., Refs. [19,20]. In short, feedback must provide sufficient, useful information to the learner, but the notions of sufficiency and usefulness depend on the amount of prior knowledge of the learner.

One useful way to investigate the effects of prior knowledge is to compare the performance of participants with relatively high vs low levels of relevant prior knowledge after training with one of the four conditions in a full 2×2 (KCR $\times$ EF) design (once again noting that all training conditions also include KR). Table I summarizes the hypothesized relative benefits of each condition, following the arguments outlined below.

For the KR only (KCR and EF absent) condition, students only receive information about the correctness of the answer. If a participant has no prior knowledge of the rule, then KR

TABLE I. Hypothesized relative benefit from the four training conditions according to prior knowledge level.

	Relative prior knowledge	
Training condition	Low	High
KR	Minimal benefit	Moderate benefit
KR + KCR	Minimal benefit	Moderate benefit
KR + EF	Moderate benefit	Moderate benefit
KR + KCR + EF	Conflicting predictions	Conflicting predictions

alone would naturally seem to have little—if any—benefit, since this is likely not enough information to determine the correct response, rule, or procedure—even in a multiple choice format [11,12]. However, if the participant has significant prior knowledge of the rule, then providing KR may help the participant to reverse engineer the answer using a partially recalled rule, and repair the mistake.

In this framework, how would the benefits of KR + KCR vs KR + EF feedback compare? Since both present more complex feedback than KR, but they are different kinds of feedback, it is difficult to quantify the relative complexity of these feedbacks and predict which may be more beneficial. Nonetheless, one can argue the following for the KR + KCR condition: if the participant has low prior knowledge, then since the skill to be learned is not a trivial fact and cannot be deduced from the remaining multiple-choice options, knowing the correct answer is not likely to help them much more than KR alone. However, if the participant has relatively high prior knowledge, then KCR may help them to reconstruct the correct application of the rule. On the other hand, in the case of relatively high prior knowledge, this feedback may be of marginal additional benefit compared to knowing the correctness of the response alone.

The KR + EF condition is different; in this case the explanation in EF can help the participant with no prior knowledge of the rule to apply the rule and find the correct answer. In this condition even participants with no prior knowledge will have increased learning compared to KR. However, for participants with high prior knowledge, the benefit of this condition (albeit moderate) will not be different from the KR alone condition (since they already know the rule). Rather, the benefit for these high-performing students will accrue from the KR provided.

Finally, for the KR + KCR + EF condition, the level of benefit is not clear *a priori*. On the one hand, one could argue that the benefits of KCR and EF could add, resulting in higher learning than KCR or EF alone. On the other hand, several studies have shown that too much feedback information may inhibit learning (cf., Refs. [21–23]) or it may have little to no effect at all [5,24]. Therefore, it is an empirical question as to the extent to which these effects compete and which may be dominant in the learning domain of this study.

Therefore, this simple model of the effect of prior knowledge produces several testable predictions. Most notably, for participants with relatively low prior knowledge, KR and KR + KCR conditions will have the same (minimal) benefits and the KR + EF condition will have moderate benefits. For participants with relatively high prior knowledge, KR, KR + KCR, and KR + EF will all have similar moderate benefits, thus only the simplest feedback is needed. For the KR + KCR + EF condition it is an open question as to the benefits to either high- or low-performing students. Note that this model is somewhat consistent with Van der Kleij *et al.* [4], since they found the EF is best for higher level learning outcomes which, here, might be interpreted as relevant for students with low prior knowledge.

B. Learning domain: The physics essential skill of vector math

The learning domain in this study is a set of specific "essential skills" which are not highly complex. Rather, they are elementary procedural skills used as a necessary part of problem-solving tasks that instructors often assume students have mastered with high accuracy and in relatively little time (e.g., Refs. [25–29]). Often to the surprise or chagrin of the instructor, students—especially low-achieving students—typically do not have these skills or they are far from fluent in their use. As such, correcting these issues would serve to widen a critical bottleneck to student problem-solving success. Designing specific training to address these skills is in line with the theoretical frameworks of deliberate practice, in which expert performance in a domain is the result of focused efforts to improve performance [30], and reduction of cognitive load, [25,26]. While we stress that learning these basic procedural skills is not sufficient for succeeding in more complex problems, mastering such skills is presumed necessary.

The STEM essential skill used in this study is simple vector math, specifically the two kinds of vector products—dot and cross products—which are commonly used in university-level introductory physics, especially in the second semester. In fact most, if not all, relevant physics textbooks include an early chapter on vector math, but it is clear that this is not sufficient. A number of studies have documented significant and somewhat alarming student difficulties with vector operations, such as vector addition, subtraction, dot or cross products, as well as vector decomposition (e.g., Refs. [31–34]), showing that—even post-instruction—typically only 50%–70% of calculus-based physics students can correctly perform these basic vector arithmetic operations.

II. PARTICIPANTS AND DESIGN

A. Student populations and data collection

For both experiments in the study, a total of $N = 456$ (287 male, 169 female) students attending a large public university participated in the study, which took place over the course of two semesters. The participants were enrolled in introductory physics, either calculus-based mechanics (experiment 1, $N = 206$) in the first semester or calculus-based electromagnetism (experiment 2, $N = 250$) in the second semester. Participants were selected from a pool of all students enrolled in these courses ($N > 1000$ for each course, with 6 to 7 lecture sections per course), and were randomly assigned to this study and into its conditions, the remaining students participated in other physics education studies. No participants were in both experiments. A vast majority (444/456) of these students completed both

training (if applicable) and the assessment. A 1-way ANOVA of student grades in the course—normalized as z scores of all students within each lecture section (including ones who were not in this study)—showed no dependence of grade on condition [$F(4) = 1.352,\ p = 0.250$].

Students participated in the study as part of a special homework assignment (assigned to all students in the course) for a small amount of course credit. The option of an alternative written assignment was given to students if they chose not to participate in a PER study, though more than 95% of all students participated in the assignment. The special homework assignment consisted of completing various physics tasks, sometimes including those in this study, for a total time not exceeding 50 min. Full credit was given for participation, regardless of performance. Our observations and poststudy debriefing interviews with all students indicated that virtually all participants performed the tasks earnestly.

B. Procedure and design

All computer training for both experiments used the lab software EPrime, and participants sat in individual carrels—each with a computer—in a quiet room. Each training session consisted of brief instruction slides followed by a series of training questions. After each response, the system provided immediate feedback varying by condition. Timing data were collected, including response times, correct answer viewing times, and time spent viewing the EF. Participants were assigned to training conditions, summarized in Table II. The final assessment (described in Sec. II. D) was in paper-and-pencil format for all conditions.

The training conditions were control (no training), KR (only correct or incorrect feedback), KR + KCR (correct or incorrect + correct response), KR + EF (correct or incorrect + explanation), and KR + KCR + EF (correct or incorrect + correct response + explanation). Examples of the EF feedback are shown in Sec. II C. All computer training conditions used the same training questions and differed from each other only by the type of answer-based feedback given to the student. Examples of training and each type of feedback type can be found in the Supplemental Material [35].

Note that, as indicated in Table II, participants in experiment 1 had received very little (if any) explicit classroom instruction on dot and cross products, as they are not emphasized in the first semester, though some students may have had instruction in other previous physics or math courses in high school or college. However, the participants in experiment 2 did receive explicit dot and cross products instruction in the classroom, since it is emphasized much more in the second semester.

Immediately after training, participants completed unrelated physics tasks that lasted approximately 20 min. After this unrelated task, participants completed the final assessment with no feedback. Participants in the control condition also received the assessment after 20 min of unrelated tasks, but with no prior vector training. Assessment completion times were typically about 10–15 min.

C. Materials: Topic and question types

A correct determination of a vector product requires finding the magnitude and either the direction (cross product) or the sign (dot product). Because our own pilot research and previous research by others [31,32] found that students have difficulties with each of these subtasks, the training included practice on these subtasks (scores in control conditions in this study verify this issue). Specifically, training included six question types, three for each vector product type: dot product (determine sign, determine magnitude, compare magnitudes) and cross product (determine direction, determine magnitude, compare magnitudes). See Fig. 1 for examples of the training question types. Each of these question types varied the magnitude and/or direction of the vectors between questions. The training consisted of four blocks of six questions (one for each question type) for a total of 24 questions. The last block replaced the two compare magnitude questions with dot product direction and cross product sign because of the higher relative importance of the latter. The training typically took about 10–15 min (more details provided in Sec. IV).

The elaborated feedback for the EF condition consisted of explanations specific to a given question type. As an

TABLE II. Experimental design. Note that numbers are counts of participants.

					Conditions				
Training topic	n	Student population	Dot or cross classroom instruction	Attitudes & beliefs items	Control (no training)	KR	KR+KCR	KR+EF	KR+KCR+EF
Dot or cross products experiment 1	206	Calc-based mechanics	No	No	53	55	47	51	⋯
Dot or cross products experiment 2	250	Calc-based electromag.	Yes	Yes	50	49	49	51	51

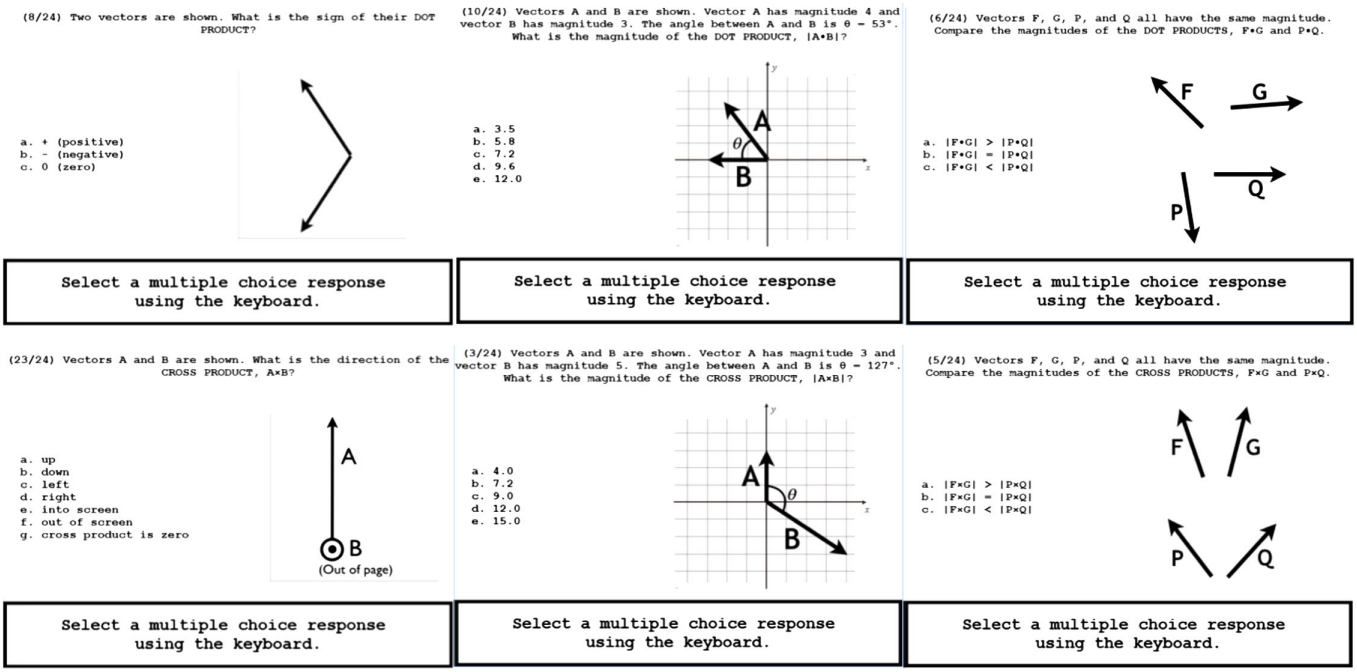

FIG. 1. Examples of each question type used during training. From left to right, top to bottom: Dot product determine sign, dot product determine magnitude, dot product compare magnitudes, cross product determine direction, cross product determine magnitude, and cross product compare magnitudes.

example, the EF for dot product sign and cross product compare magnitudes questions are shown in Fig. 2.

When KR or KCR feedback was used in conjunction with EF, the KR and/or KCR feedback was shown prior to the EF. The EF images used in training obscured (was placed over) only the multiple choice options, meaning the students could still see the problem statement and accompanying figure while studying the EF. To draw attention to the contrast between similar operations, all dot product EF images contained an orange border, and cross product EF images contained a purple border.

D. Materials: Assessment

Two assessments were used in this study. The first is an instrument we constructed in pilot studies, called the DotCross assessment, used as the content assessment for experiments 1 and 2. The DotCross assessment consisted of 14 items and measured participant ability to determine the sign and magnitude of the dot product of two vectors (8 items) and the direction and magnitude of the cross product of two vectors (6 items). See the Supplemental Material [35] for the full DotCross assessment. The assessment items were similar—though not identical to—the training questions, and covered all training question types. The reliability of this instrument is fairly high, with a Cronbach's $\alpha = 0.88$, and all items had mean scores in the range of about 0.3–0.7 (most around 0.5) and good point-biserial correlations in the range of 0.4–0.7, with the exception of one item which had a very high mean score.

The validity of the DotCross assessment is supported in four ways. First, the items were modeled after straightforward questions in textbooks and relevant items from other

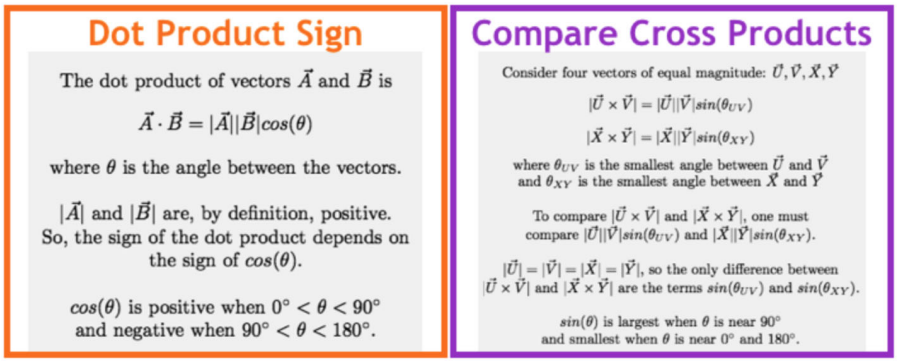

FIG. 2. Examples of EF images used in dot or cross product training.

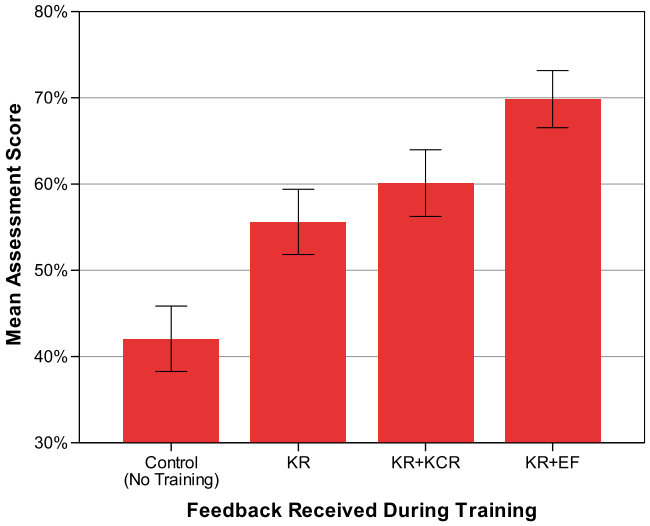

FIG. 3. Experiment 1 mean assessment scores for each condition.

validated vector assessment instruments [31,32]. Second, the items used were also based on over 5 years of pilot testing in our education research lab, including modifications based on student testing and postinterviews. Third, the instrument is significantly (albeit relatively weakly) correlated with the course grade ($r = 0.22$, $p = 0.03$). Finally, the items were constructed by two instructors (Heckler and Mikula) with, combined, over 20 years experience teaching the topic; during the research phase, other instructors were also involved in the process.

The second assessment used was a very brief assessment of student attitudes and beliefs, which was administered only in experiment 2. This assessment was not designed to be an extensive measurement of attitudes and beliefs. Rather the assessment consisted of several Likert-scale items intended to obtain a simple measure of the general sense of participant ratings of their own proficiency, importance of the topic, and beliefs about the how much they learned from the training (see Sec. IV. C for more details).

III. CONTENT ASSESSMENT SCORE RESULTS

We are interested in determining differences in scores for the various training conditions and how this may interact with prior knowledge. Specifically, we would like to test the predictions discussed earlier in Table I.

A. Experiment 1: Scores, feedback complexity, and prior knowledge

Experiment 1 mean scores for each condition (Fig. 3) reveal two important results. First, all training conditions resulted in significantly higher scores compared to control [ANOVA $F(3) = 9.8$, $p < 0.001$, *post hoc* comparisons against control using Dunnett's t test, $p \leq 0.04$ for all conditions]. Specifically, the effect sizes (gains over control) range from $d = 0.5$ for KR to $d = 1.0$ for KR + EF.

Second, the scores appear to increase with the increased complexity of training feedback. To gain more insight into this apparent signal, let us consider the factor of prior knowledge. In Sec. I. A we hypothesized that more feedback information is more beneficial for participants with relatively low levels of prior knowledge. Thus training with KR + EF should be the most effective, while KR and KR + KCR should be the least effective (and probably similar to each other). In contrast, for participants with relatively high levels of prior knowledge, only minimal feedback is needed to remind the learner of the correct method, thus KR, KR + KCR, and KR + EF would all be about equally and moderately effective. We can test these hypothesized effects of prior knowledge on training, by considering two different measures of (or proxies for) prior knowledge.

The first proxy we use for prior knowledge is course grade (though one might also argue that this is also a measure of aptitude). Figure 4 presents the scores broken down by students above and below the median grade, providing evidence for the claim that students with a higher course grade have higher prior knowledge. The scores for students in the control condition with course grades above the median are about $d = 0.7$ standard deviations higher than for students with grades below the median [$t(51) = 2.6$ $p = 0.013$].

A close examination of Fig. 4 reveals that the results support our prediction in Table I that the high and low grade students responded differently to the training feedback. Specifically, as expected, for students below the median (low prior knowledge), the KR + EF condition scores are higher than KR or KR + KCR. Both Fig. 4 and a *post hoc* analysis confirm this [ANOVA $F(2) = 5.2$, $p < 0.01$, *post hoc* comparing KR + EF against other two: Dunnett's t test, $p = 0.02$ for KR and $p = 0.03$ for KR + KCR]. Also as expected, for students above the median there are no significant differences in the scores for KR, KR + KCR, and KR + EF [$F(2) = 1.1$, $p = 0.34$].

The second proxy for prior knowledge is perhaps a more direct measure of vector prior knowledge. In the training conditions, the first 6 training questions are all different question types. Therefore, the score on these 6 questions ("prescore") could be used as a simple measure of the prior knowledge of the participant. In fact, the mean prescores were 36% and 49% for experiments 1 and 2, respectively—these scores were similar across conditions and were similar to the DotCross assessment means for the respective control conditions. Futhermore, perhaps as to be expected, there are only weak correlations between the prescore and the course grade ($r = 0.17$, $p < 0.05$ for both experiments), thus they are not measuring the same ability.

In order to allow for comparisons across experiments, we used the median pre score for the pooled populations of

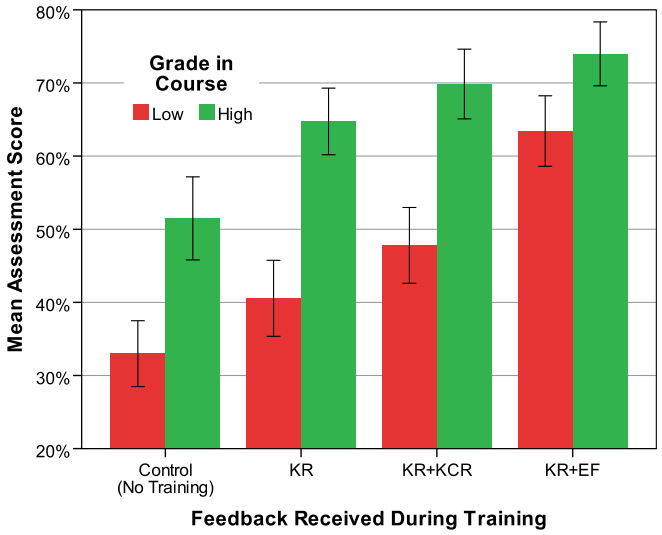

FIG. 4. Experiment 1 mean assessment scores for each condition, for students with high and low course grades.

experiments 1 and 2 to determine the cutoff between high and low pre scores. This resulted in 60% of experiment 1 and 40% of experiment 2 participants below the pooled median (low pre scores). This difference is also to be expected, since the students in Experiment 2 have had more practice with vector products.

The results using the prescores (Fig. 5) are similar to the results from the high-low grade student analysis. For students below the median on the first 6 training questions, the KR + EF scores are higher than the KR and KR + KCR scores [ANOVA $F(2) = 3.2$, $p = 0.046$, *post hoc* comparing KR + EF against other two: Dunnett's t test, $p = 0.037$ for KR and $p = 0.022$ for KR + KCR]. For students above the median, there were no significant differences in the scores for KR, KR + KCR, and KR + EF [$F(2) = 2.0$, $p = 0.14$].

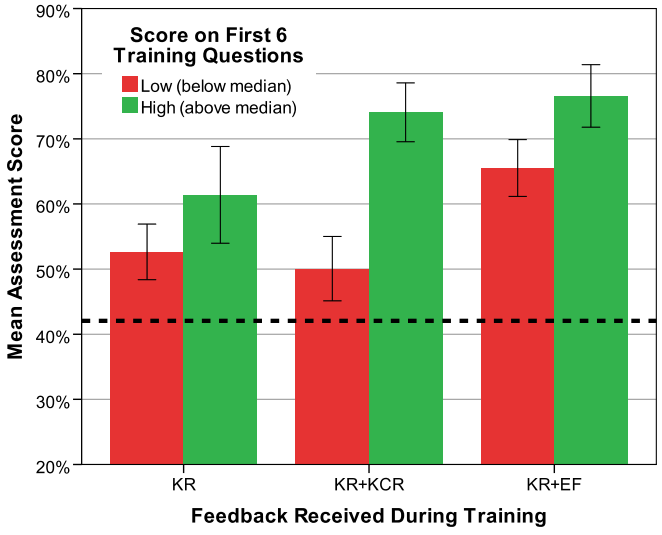

FIG. 5. Mean assessment scores for experiment 1 feedback conditions, split by high and low scores on the first six training questions. The dotted line represents the mean control score.

In sum, we used two different measures (or proxies) for prior knowledge, namely grade in course and score on the first 6 training questions. In both cases we found the same pattern: students with low prior knowledge benefitted significantly more from KR + EF feedback compared to KR or KR + KCR feedback, and students with high prior knowledge benefited equally from all three of these training conditions. These results confirm the model that, for this learning domain, increased feedback complexity—namely, providing explanations (EF)—helped students with low prior knowledge significantly more than simple feedback such as providing the correct answer.

B. Experiment 2: EF and KCR, a full factorial design

The results of experiment 1 confirmed that among the training types KR, KR + KCR, and KR + EF, the most complex feedback was most beneficial to participants with low prior knowledge. One could increase the complexity yet more with the feedback condition KR + KCR + EF, but as discussed in Sec. I. A, there are conflicting predictions as to whether such additional information will increase learning or will cause overload and decrease learning. In experiment 2, we implemented a 2 × 2 design (KCR × EF) (and KR always present) to determine whether any interactions occur between KCR and EF for this learning domain. Recall also that the participants in experiment 2 were enrolled in the semester 2 course (electromagnetism), in which dot and cross products are explicitly addressed in class.

The mean scores for each condition are presented in Fig. 6. Similar to experiment 1, all training conditions resulted in significantly higher scores compared to control

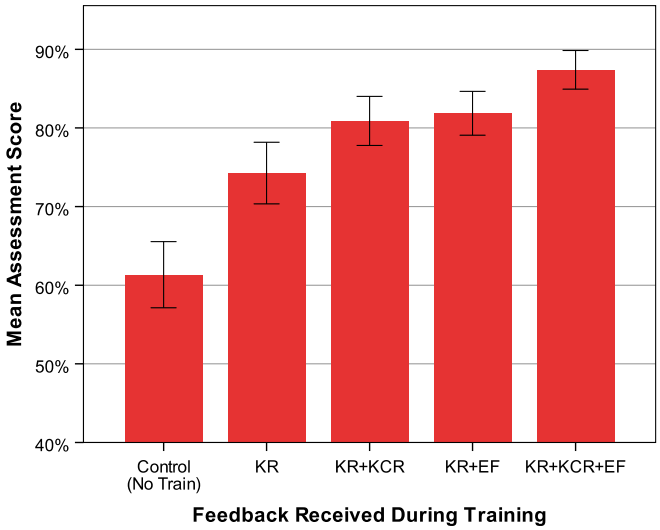

FIG. 6. Experiment 2 mean assessment scores for each condition.

[ANOVA $F(4) = 9.0$, $p < 0.001$, *post hoc* comparing against control Dunnett's t test, $p \leq 0.01$]. The effect sizes (gains over control) are also similar to experiment 1, ranging from $d = 0.46$ for KR to $d = 1.1$ for KR + KCR + EF.

However, there are some important differences from experiment 1. First, participants in experiment 2 (semester 2) scored higher than those in experiment 1 (semester 1) in all comparable conditions (t test $p \leq 0.006$). The effect size between control scores for experiment 1 and experiment 2 is $d = 0.7$. The higher score is to be expected since the participants in experiment 2 received more explicit course instruction and practice in dot and cross products in semester 2.

Perhaps the most important difference from experiment 1 is the pattern of the training condition means. To compare directly with experiment 1, let us first only consider KR, KR + KCR, and KR + EF for experiment 2. An ANOVA and *post hoc* analysis reveals that there were no significant differences in mean scores for these three training conditions [$F(2) = 1.575$, $p = 0.211$]. Furthermore, separating out high and low course grade students or high and low prescores reveals the same result, namely, that there were no significant differences in scores between KR, KR + KCR, and KR + EF feedback conditions for either high- or low-performing students ($p = 0.503$, 0.096 for low- and high-performing grades and prescores, respectively). The lack of difference in scores for these three conditions in experiment 2 is consistent with our hypothesis since, as discussed earlier, the participants in experiment 2 (semester 2) have relatively high prior knowledge on average, and there should be less added benefit of adding either KCR or EF to the feedback.

Nonetheless, this brings us to the question of whether including *both* KCR and EF (i.e., KR + KCR + EF) accrued additional benefit, or if the additional feedback impedes learning. Examination of Fig. 6 reveals that KR + KCR + EF certainly does not perform worse than KR + KCR or KR + EF, and, in fact, there is evidence for some added benefit to including both KCR and EF in the feedback for this learning domain. An ANOVA analysis on all four training conditions confirms that the means are different [$F(3) = 3.039$, $p = 0.030$] and a Tukey (*post hoc*) multiple comparison shows that the only significant pairwise difference between conditions is between KR + KCR + EF and KR ($p = 0.016$); all other pairwise comparisons show no significant differences.

We can gain insight into the nature of the benefit of KR + KCR + EF by comparing performance for participants with high or low scores on the prescore (i.e., high or low prior knowledge). Figure 7 shows the results. The results clearly show that students with high prior knowledge did not gain any added benefit from increased feedback complexity [ANOVA $F(3) = 1.24$, $p = 0.3$], but there was a difference between conditions for students with low prior knowledge [ANOVA $F(3) = 2.85$, $p = 0.04$], and the only significant pairwise difference is between KR + KCR + EF and KR (Tukey, $p < 0.05$). Therefore, once again, the benefit of the complex feedback stems from students with low prior knowledge.

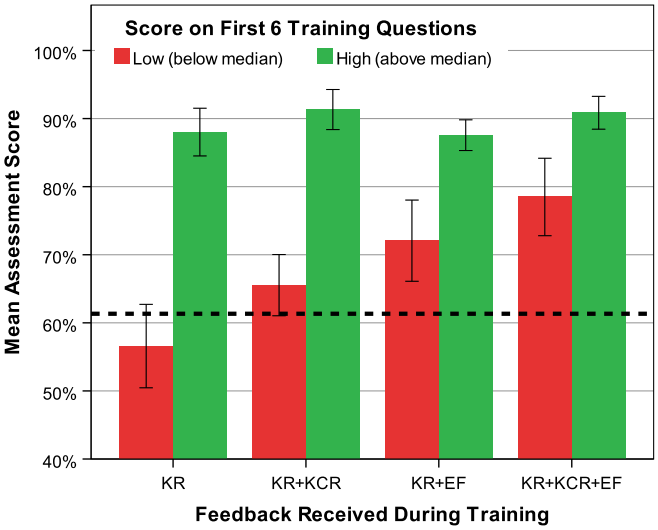

FIG. 7. Mean assessment scores for experiment 2 students for each training condition, split by high or low scores on the first six training questions (prescores). The dotted line represents the mean control score.

Finally, in order to gain more insight into the results, we completed a univariate ANOVA with two factors—KCR (present or absent), EF (present or absent)—and two covariates—course grade and prescore. We included all main effects and two-way interactions. All main effects and significant interactions are shown in Table III. All main effects were significant except for KCR, which had marginal benefits. There was no significant interaction between KCR and EF; rather, it appears that in this learning domain the two effects are simply additive.

The only significant interactions were between EF and a measure of prior knowledge—student performance on the first 6 questions of the training—and between KCR and course grade. Thus we have evidence of a knowledge-by-treatment interaction, namely, that students with low prior knowledge on average benefit from EF while high prior knowledge students on average do not benefit from EF,

TABLE III. Experiment 2 univariate ANOVA results, excluding control. Note that $R^2 = 0.39$.

Factor	$F(1)$	p	Partial η^2
KCR	2.17	0.14	0.012
EF	13.8	<0.001	0.070
Course grade	5.95	0.016	0.032
Prescore	67.2	<0.001	0.270
EF×Prescore	10.1	0.002	0.053
KCR×Course grade	6.14	0.014	0.033

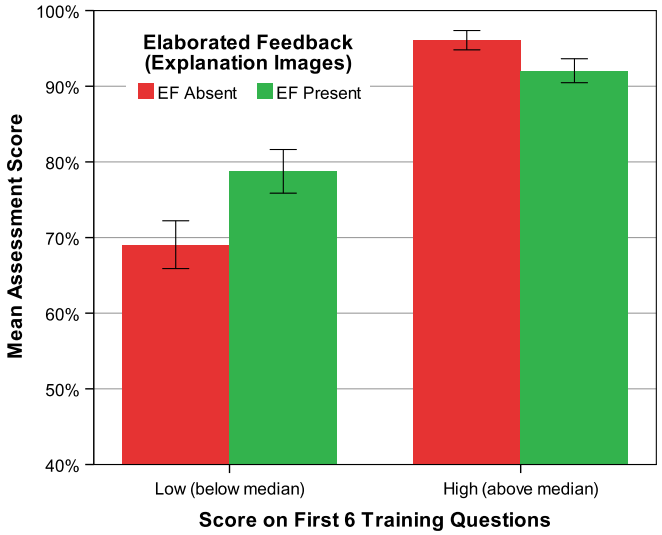

FIG. 8. Experiment 2 mean scores for EF present or absent and for students with high or low score on the first six questions of training.

perhaps because they are at ceiling (see Fig. 8). The KCR-course grade interaction reveals that students with low course grades tend to benefit from KCR more than students with high course grades.

IV. TRAINING AND TIMING DATA RESULTS

In this section we examine the amount of time learners spent on various parts of the training. The training times are analyzed for two reasons: to gain insight into the relative efficiency of the feedback conditions and to gain insight into how students are using the elaborated feedback in conditions containing that feedback, which were often the highest scoring conditions. The timing data considered here are the total training time and the time spent viewing the EF (explanation time). Note that timing data analyzed in this section only considers the training conditions and excludes the control (no training) condition.

A. Total training time and efficiency

There are several important observations about the total training time. First, the total training time was not correlated with either of our proxies of prior knowledge. However, the total time was weakly correlated with score for experiment 1 ($r = 0.25$, $p < 0.01$) and this correlation was about this same size for each condition. On the other hand, in experiment 2, the correlations were not significant ($r = 0.05$, $p > 0.4$). The correlation of score with training time for experiment 1 but not experiment 2 may indicate that the less prepared students tend to get more benefit by spending more time with training.

Second, in Sec. III we determined which feedback conditions were the most effective in terms of obtaining the highest scores. However, we would also like to determine which feedback conditions are the most *efficient*. Naturally, there are a variety of ways to define efficiency of an instructional intervention, usually including information on the amount of learning as well as time spent on learning. Here we will define efficiency ε as a rate of gain of score. That is, for the ith student, $\varepsilon_i = (\mathrm{Score}_i - \mathrm{MeanScore}_{\mathrm{Control}})/(\mathrm{StDev}_{\mathrm{Control}} \times \mathrm{TrainTime}_i)$.

It is important to note that the time unit is scaled to 1000 sec in order to make the values of ε easily readable. Thus, efficiency ε may be interpreted as the number of control standard deviations gained per thousand seconds spent training.

The total training times and efficiencies for each condition are shown in Table IV (mean and median times are shown because the distributions are right-skewed). The training times showed significant differences for experiment 1 (Kruskal-Wallis $K = 7.3$, $p = 0.03$) but not for experiment 2 (Kruskal-Wallis $K = 3.0$, $p = 0.39$). For both experiments, training time increased by about 10%–20% (1–2 min) when EF is included in the feedback. Interestingly, for experiment 2, the total training time for KR + KCR was slightly less than for KR alone, possibly because without KCR, students would have to take time to determine the correct answer.

As for efficiency, there were no significant differences between conditions for each experiment (Kruskall-Wallis $K = 3.3$, $p = .19$ for exp. 1, and $K = 5.6$, $p = .13$ for Exp. 2). However, taken together the trends are in the same direction, namely, that conditions with EF tended to have higher efficiencies but, again the significance is marginal at best (Mann-Whitney $U = 3031$, $p = 0.076$ for exp. 1, and $U = 5443$, $p = .22$ for Exp. 2), with an effect size of $d = 0.25$. Therefore, while the scores for EF training are

TABLE IV. Mean and median total training times and mean efficiencies for each condition in each experiment. Times are listed in seconds.

	Experiment 1			Experiment 2		
Feedback condition	Mean time (SD)	Median time	Mean efficiency (SE)	Mean time (SD)	Median time	Mean efficiency (SE)
KR	563 (249)	511	.70 (.28)	563 (233)	509	.91 (.29)
KR + KCR	655 (307)	585	.73 (.26)	481 (148)	454	1.24 (.27)
KR + EF	672 (224)	625	1.24 (.23)	530 (169)	542	1.20 (.24)
KR + KCR + EF	⋯	⋯	⋯	518 (154)	519	1.65 (.24)

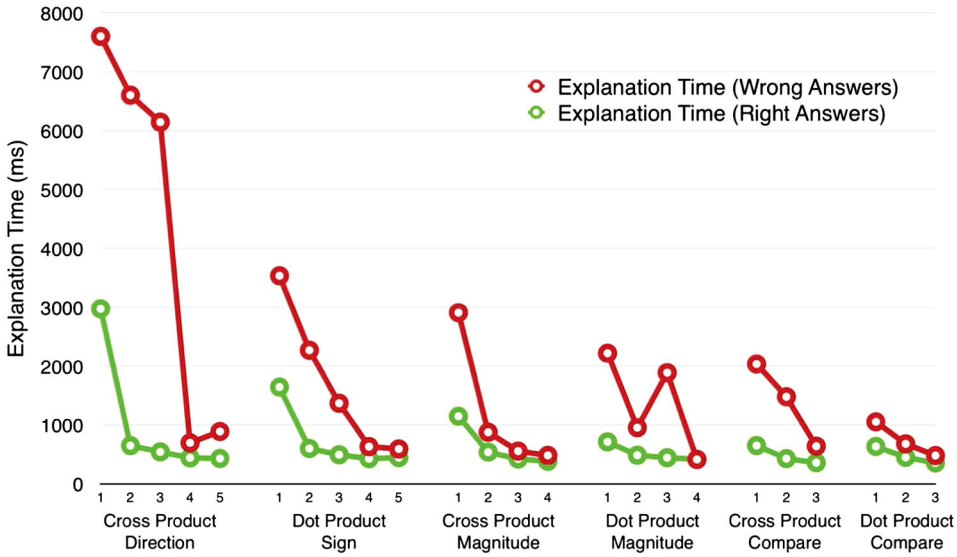

FIG. 9. Median explanation times for each training trial within a question type and for correct or incorrect responses in each trial.

significantly higher than KCR training, there is at best only marginal evidence that the efficiency (learning rates) of EF are higher than KCR. With such a relatively small effect size, a larger sample size is needed to determine whether this possible higher learning rate is statistically reliable.

Finally, while the above result confirms that EF training results in relatively higher scores, it is important to note that for the less effective feedback conditions, many students tended to finish them quickly with very little learning.

B. Explanation viewing time

We have shown that EF training resulted in the highest scores; therefore this kind of feedback is of high interest. In order to gain more insight into how students progress through the EF training and how they are using this feedback, we examined the amount of time spent viewing the explanations.

If providing explanations improves learning, then one would expect to see evidence for this causation in the explanation viewing times. Specifically, participants who answered a given training question incorrectly would be expected to view the explanations longer than those who answered correctly. Figure 9 presents the trial-by-trial median explanation viewing times for participants answering correctly or incorrectly for all six question types (medians are used to depict central tendency because distributions are right skewed). Recall that training consisted of several blocks, interleaving the six question types in each block. Note that, because of limitations for some question types, the last couple of blocks did not include all question types.

Figure 9 reveals two important patterns. First, it confirms the causal expectation that participants answering incorrectly tended to view the explanations significantly longer than those answering correctly. One might argue, however, that Fig. 9 confounds results, since it does not account for the longitudinal progress of the participants. For example, for a given question type, for participants answering the second trial incorrectly, it is not clear from the figure how many of these participants also answered the first question incorrectly or correctly.

To address this, in Fig. 10 we present an example for one of the question types in which student responses to the first three training trials of each question type were categorized according to their scoring pattern. That is, within each question type, students were categorized as "Right-Right-Right" if they got all three questions right, "Wrong-Right-Right" if they got the first question wrong and the second and third questions right, and so on. We include the four most-populated categories in the figure, which account for 81% of the participants. Explanation times for each category were then compared for the first three trials.

The results further support the expectation that explanation times are relatively high when the question was incorrectly answered and low when correctly answered. Figure 10 presents this result visually, and Wilcoxon signed rank tests showed that in all four answer patterns in this figure the explanation viewing time was significantly longer after incorrect answers than after correct answers ($p < 0.05$). For example, the Wrong-Right-Right category, the explanation time is largest for the first then smaller for the second ($Z = 2.4$, $p = 0.02$) and third ($Z = 3.0$, $p < 0.01$) trials, whereas for the Right-Wrong-Right category, the explanation viewing times are largest for the second trial and significantly smaller for the first ($Z = 3.0$, $p < 0.01$) and third ($Z = 3.0$, $p = 0.01$) trials, as expected. Further, for the Wrong-Wrong-Right category, the viewing times for the first and second trials were not significantly different ($Z = 1.4$, $p < 0.16$), but they were for the first and third ($Z = 2.1$, $p < 0.04$) and second and third

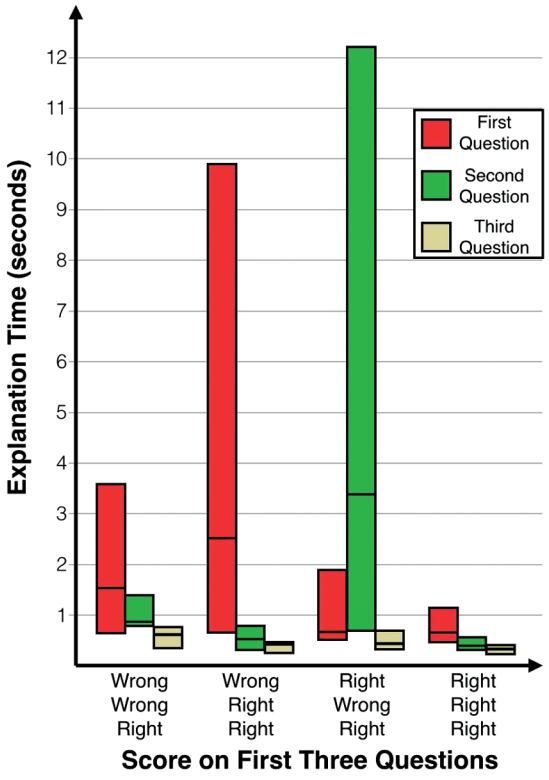

FIG. 10. Box plots indicating 1st, 2nd, and 3rd quartiles of explanation times on the first three cross product compare magnitudes questions, split by student response patterns on the first three questions.

($Z = 2.2$, $p < 0.03$) trials. This explicitly shows that students in the EF condition are spending time to read the explanations when they get a training question incorrect, and that is why the EF condition has high learning gains.

The second important pattern in Fig. 9 is that, within each question type, explanation viewing times for participants answering incorrectly typically decreased somewhat rapidly as the training trials advanced, and is approximately equal to the explanation viewing times for correct answerers by the third or fourth trial. It is important to keep in mind that typically only ~10% of the participants are answering incorrectly by the fourth trial, so this typically represents only a small portion of the population. However, this does indicate that after a certain number of trials, students answering incorrectly tended to ignore the EF. Possible explanations for this could include the following: (i) students have read the EF earlier, and found it unhelpful, (ii) students have read the EF earlier and found it helpful but no longer necessary to read again, and (iii) students lose (or never had) interest and/or motivation in how to determine the correct answer.

C. Data on beliefs of proficiency, learning, and importance of topic

After training and the content assessments, the participants in experiment 2 were given a brief survey including items pertaining to their beliefs about their proficiency in the topic, how much they learned in training, and the importance of the topic. The text of these items and summary of results are shown in Table V.

There are several important results from the combination of survey performance and content performance data. First, as seen in Table V, all training conditions but KR significantly raised self-reported proficiency compared to Control, with more effective training conditions—as measured by scores on the assessment—showing the largest increases (compared to control) in self-reported proficiency. Specifically, a Kruskal-Wallis test revealed significant differences between the conditions for the self-reported proficiency item [$K(4) = 20.5$, $p < 0.001$], with follow-up *post hoc* comparisons only showing significantly higher self-reported proficiency for KR + KCR + EF compared to both control and KR.

One possible explanation for highest self-reported proficiency for KR + KCR + EF is that this condition improved their performance the most, and this in turn improved their self-reported proficiency the most. This

TABLE V. Mean student scores (percent correct) and mean student responses to attitudes and beliefs survey questions. Note that numbers in parentheses represent standard deviations.

	Control	KR	KR + KCR	KR + EF	KR + KCR + EF
Average percent correct on assessment questions	61% (30%)	74% (27%)	81% (21%)	82% (20%)	87%(17%)
Survey question					
Rate your proficiency with using (and knowledge of) dot and cross products. (1 = low profic.; 5 = high profic.)	3.0 (1.0)	3.1 (1.0)	3.5 (1.0)	3.5 (1.0)	3.8 (0.9)
How much did you learn from the training on dot and cross products? (1 = nothing new; 5 = many new things)	n/a	2.6 (0.9)	2.6 (1.1)	3.1 (1.2)	2.8 (1.0)
How important do you think it is to understand how to use dot and cross products *for [this course]*? (1 = not at all important; 5 = very important)	4.4 (0.9)	4.4 (0.9)	4.3 (0.9)	4.1 (0.8)	4.4 (0.8)

TABLE VI. Experiment 2 univariate ANOVA results, including the factor of importance rating. *Note* that $R^2 = 0.46$.

Factor	$F(1)$	p	Partial η^2
KCR	3.09	0.08	0.019
EF	17.8	<0.001	0.099
Course grade	7.72	0.006	0.046
Prescore	57.0	<0.001	0.260
Importance rating	3.77	0.012	0.065
EF×prescore	13.2	<0.001	0.076
Importance rating×prescore	3.34	0.038	0.040

would imply that the students were fairly good judges of their own proficiency. In fact, this is supported by a high correlation between self-reported proficiency and performance ($r = 0.63$ in the control condition; other conditions had very similar correlations). This is in the high end of the range reported in a meta-analysis by Zell and Krizan [36], on a range of skills, but the high number may be expected since they report higher correlations between perceived ability and performance when tasks are well constrained and simple.

Second, there was no significant correlation between self-reported learning and assessment score within any of the training conditions. However, the conditions with EF had higher self-reported learning than conditions with no EF (mean ratings of 2.95 and 2.61 out of 5, respectively; Mann-Whitney $U = 5269$, $p = 0.034$, $d = 0.3$). Furthermore, for conditions with EF, there was a significant positive correlation between the self-reported learning and explanation viewing time ($r = 0.3$, $p = 0.003$).

Finally, while there was no significant difference in mean ratings on the importance of understanding vectors for the course among conditions, the self-reported importance rating appears to play a role in the effectiveness of the training. A univariate ANOVA similar to the analysis in Sec. III. B was performed, adding the importance rating as a factor. Table VI shows the results for significant main effects and interactions. The results are similar to the earlier results in Table III, additionally showing that student perception of importance of the topic influences the final score, and that there is an interaction between perception of importance and prescore on the first six training questions. Specifically, for students with low prescores, those who perceive the topic as important benefit more from the training than those students who do not report the topic as important (see Fig. 11). These results suggest that it is important that the students—particularly lower-performing students—believe in the importance of the skills to be trained in order to maximize learning gains.

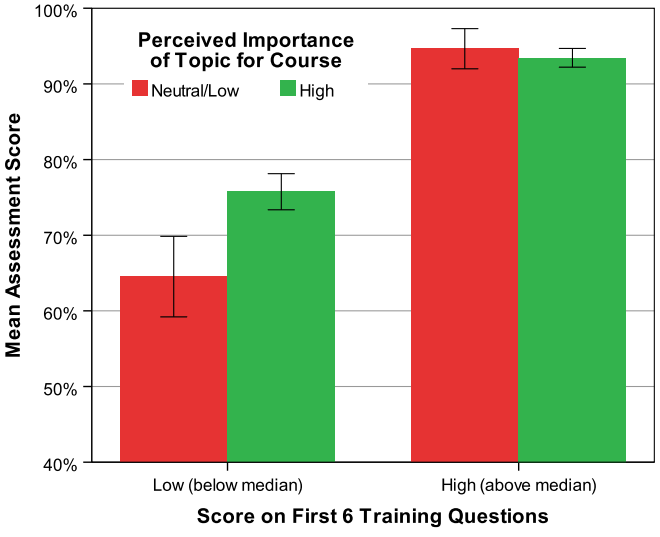

FIG. 11. Experiment 2 mean assessment scores by score on first 6 training questions (prescore), separated by perceived importance of the topic for the course.

V. SUMMARY AND GENERAL DISCUSSION

Even though there are currently considerable efforts to develop relatively complex intelligent tutoring systems, results from this study demonstrate that simple computer-based training utilizing a variety of relatively simple answer-based feedback methods can be effective in improving student accuracy with essential STEM procedural skills such as evaluating dot or cross vector products. Compared to research findings on computer-based instruction with college students [2,37,38], effect sizes from training in this study were large, ranging from 0.5 to 1.1.

As mentioned in the introduction, there are numerous studies and several reviews on the many factors affecting computer-based learning with feedback. Here, we help in the much-needed empirical and systematic exploration of the factor space to provide guidance for optimal training design. These results are helpful not only in the specific domain of the STEM essential skill of vector math, but also in providing evidence and arguments for a more general framework to guide the design of computer-based training with simple feedback in similar domains. Note that while this study investigates factors effecting learning of a simple and essential procedure skill for physics, there have been other relatively recent lab studies in the physics education domain for improving student knowledge of specific topics [39,40] and conceptual reasoning in problem solving [41] that also show promise for application in a course.

With a full factorial design studying KCR feedback and EF, a more detailed analysis reveals that EF has a significant impact above and beyond KR, while KCR had a more marginal impact beyond KR. An explanation of this lesser benefit of KCR (compared to EF) may be related to the findings of Narciss and Huth [22], who found that omitting KCR may compel students to think deeper (to find the correct answer).

Furthermore, in agreement with our hypothesis that the effectiveness of feedback improves with an increase in sufficiency and usefulness of feedback information, the

impact of training methods was modulated by prior knowledge. Namely, EF helped the most for students with the least amount of relevant prior knowledge of the topic—whether measured by scores on the first few training questions or by student grade in the course—while students with more prior knowledge achieved significant gains even from the lowest levels of feedback. Note also, as mentioned in Sec. I. A, that the explanations in the EF condition were not specific to the question, but were general to the question type. Our finding that EF results in the highest learning gains is somewhat consistent with the meta analysis of Van der Kleij *et al.* [4], but they found wide variations depending on numerous factors. Not only do our results directly validate the use of EF for an essential skills domain in introductory physics, but it also adds insight into the types of students benefiting the most.

It is interesting to note that, contrary to many previous studies mentioned in the introduction, in our study KR did often have a significant positive impact on performance. However, this improvement in performance is almost certainly not due to Skinnerian conditioning—which was tested and rejected by earlier studies (e.g., Refs. [11,12])—since it would be difficult to infer the abstract rules of vector math from KR feedback alone. Rather, consistent with our hypothesis of sufficient and useful information, it is more likely that KR improves performance because—even in the low-prior-knowledge cases—at least some subpopulation of students will still have some relevant prior knowledge of the rules (e.g., from previous courses). For these students, even the minimal KR feedback provides enough information to help them to recall the rule or repair any faulty or incomplete recollection.

We found that students tended to spend relatively more time viewing explanations after incorrect answers, but that overall, explanation viewing times decreased dramatically after the first one or two viewings within each question type. While these results helped to empirically confirm the assumption that students were attending to the explanations provided, it also raises question as to how to best sequence training questions and explanations, and why students who persist with incorrect answers ignore the explanations in the later training trials. For example, is it better to provide more difficult questions in which students answer incorrectly early rather than later? This would be consistent with models of productive failure [42] and impasse-driven learning [43], in which early struggles with difficult problems lead to stronger learning later on, despite the initial failure. Nonetheless, our study was not designed to answer such questions which must be left to future studies.

Notably, we found that the training not only improved performance but also improved students' beliefs about their own proficiency in the topic, with more effective training conditions showing higher self-reported proficiencies. Such improvements to self-efficacy-related beliefs can be an important part of STEM learning [44,45].

Furthermore, we found that student belief in the importance of the topic can play an important role in the effectiveness of training for lower performing students. That is, when low-performing students were neutral or negative about the importance of the topic, the positive effects of EF and KCR feedback vanish. However, it is not proven from these correlational results that this relation is causal, namely that increasing a student's perception of the importance of a topic will increase the effectiveness of training, though this is a compelling and interesting question to pursue.

Overall, results of this study suggests that computer-based practice to improve performance in a relatively simple STEM essential skill should include both KR and a brief general explanation (EF) relevant to the question type. One might also include information about the correct response in the feedback, but this extra information may not be necessary, and—as mentioned in other studies—may be distracting or otherwise inhibit improvement in some cases (e.g., Ref. [8]). Nonetheless, in our study, providing KCR in addition to EF did not appreciably increase or decrease learning.

Finally, it is important to consider that while students in this study had significant learning gains, they only completed one training session. In order for such gains in these essential skills to be retained over an academically relevant time interval, it is likely that repeated, distributed practice is necessary (e.g., Ref. [46]). We do have evidence in another STEM essential skill domain that successful retention occurs with such practice [28,29]. Retention of these essential skills is important, as these skills are necessary for success in solving more complex problems.

ACKNOWLEDGMENTS

Funding for this research was provided by the Center for Emergent Materials: an NSF MRSEC under Grant No. DMR-1420451.

[1] C.-L. C. Kulik and J. A. Kulik, Effectiveness of computer-based instruction: An updated analysis, Comput. Hum. Behav. **7**, 75 (1991).

[2] R. Niemiec and H. J. Walberg, Comparative effects of computer-assisted instruction: A synthesis of reviews, J. Educ. Comput. Res. **3**, 19 (1987).

[3] R. M. Tamim, R. M. Bernard, E. Borokhovski, P. C. Abrami, and R. F. Schmid, What forty years of research says about the impact of technology on learning a second-order meta-analysis and validation study, Rev. Educ. Res. **81**, 4 (2011).

[4] F. M. Van der Kleij, R. C. W. Feskens, and T. J. H. M. Eggen, Effects of feedback in a computer-based learning environment on students' learning outcomes a meta-analysis, Rev. Educ. Res. **85**, 475 (2015).

[5] R. L. Bangert-Drowns, C.-L. C. Kulik, J. A. Kulik, and M. Morgan, The instruc- tional effect of feedback in test-like events, Rev. Educ. Res. **61**, 213 (1991).

[6] J. Hattie and H. Timperley, The power of feedback, Rev. Educ. Res. **77**, 81 (2007).

[7] B. J. Mason and R. Bruning, Providing feedback in computer-based instruction: What the research tells us, http://dwb4.unl.edu/dwb/Research/MB/MasonBruning.html.

[8] E. H. Mory, Feedback research revisited, in *Handbook of Research on Educational Communications and Technology*, 2nd ed., edited by D. H. Jonassen and M. P. Driscoll (Taylor and Francis, London, 2004), pp. 745–783.

[9] V. J. Shute, Focus on formative feedback, Rev. Educ. Res. **78**, 153 (2008).

[10] J. Hattie and M. Gan, Instruction based on feedback, in *Handbook of Research on Learning and Instruction*, edited by R. A. Mayer and P. A. Alexander (Routledge, New York, NY, 2011), pp. 249–271.

[11] W. J. Roper, Feedback in computer assisted instruction, Programmed learning and educational technology **14**, 43 (1977).

[12] D. A. Gilman, A comparison of several feedback methods for correcting errors by computer-assisted instruction, ERIC Clearinghouse (1968).

[13] R. W. Kulhavy and W. Wager, Feedback in programmed instruction: Historical context and implications for practice, in Interactive Instruction and Feedback edited by J. Dempsey and G. Ales (Educational Technology Publications, Englewood Cliffs, 1993), pp. 3–20.

[14] J. Gordijn and W. J. Nijhof, Effects of complex feedback on computer-assisted modular instruction, Comput. Educ. **39**, 183 (2002).

[15] F. M. van der Kleij, T. J. H. M. Eggen, C. F. Timmers, and B. P. Veldkamp, Effects of feedback in a computer-based assessment for learning, Comput. Educ. **58**, 263 (2012).

[16] E. R. Fyfe, B. Rittle-Johnson, and M. S. DeCaro, The effects of feedback during exploratory mathematics problem solving: Prior knowledge matters, J. Educ. Psychol. **104**, 1094 (2012).

[17] M. J. Hannafin, K. M. Hannafin, and D. W. Dalton, Feedback and emerging instructional technologies, in *Interactive Instruction and Feedback*, edited by J. V. Demposey and G. C. Dales (Educational Technology, Englewood Cliffs, 1993), pp. 263–286.

[18] R. E. Snow and D. F. Lohman, Toward a theory of cognitive aptitude for learning from instruction, J. Educ. Psychol. **76**, 347 (1984).

[19] S. Kalyuga, Expertise reversal effect and its implications for learner-tailored instruction, Educ. Psychol. Rev. **19**, 509 (2007).

[20] S. Tobias, An eclectic appraisal of the success of failure of constructivist instruction, in *Constructivist Theory Applied to Education: Success of Failure?* edited by S. Tobias and T. D. Duffy (Routledge, Taylor and Francis, New York, 2009), pp. 335–350.

[21] R. W. Kulhavy, M. T. White, B. W. Topp, A. L. Chan, and J. Adams, Feedback complexity and corrective efficiency, Contemp. Educ. Psychol. **10**, 285 (1985).

[22] S. Narciss and K. Huth, How to design informative tutoring feedback for multimedia learning, in *Instructional Design for Multimedia Learning*, edited by H. M. Niegemann, D. Leutner, and R. Brunken (Waxman, Munster, NY, 2004), pp. 181–195.

[23] G. D. Phye and T. Bender, Feedback complexity and practice: Response pattern analysis in retention and transfer, Contemp. Educ. Psychol. **14**, 97 (1989).

[24] A. N. Kluger and A. DeNisi, The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory, Psychol. Bull. **119**, 254 (1996).

[25] J. J. Van Merriënboer and P. A. Kirschner, *Ten Steps to Complex Learning: A Systematic Approach to Four-Component Instructional Design* (Routledge, New York, 2012).

[26] J. J. Van Merrienboer and J. Sweller, Cognitive load theory and complex learning: Recent developments and future directions, Educ. Psychol. Rev. **17**, 147 (2005).

[27] J. R. Hartman and E. A. Nelson, "Do we need to memorize that?" or cognitive science for chemists, Found. Chem. **17**, 263 (2015).

[28] A. F. Heckler, B. Mikula, and R. Rosenblatt, Student accuracy in reading logarithmic plots: the problem and how to fix it, *Proceedings of the 2013 IEEE Frontiers in Education Conference*, pp. 1066–1071.

[29] B. D. Mikula and A. F. Heckler, The effectiveness of brief, spaced practice on student difficulties with basic and essential engineering skills, *Proceedings of the 2013 IEEE Frontiers in Education Conference*, pp. 1059–1065.

[30] K. A. Ericsson, R. T. Krampe, and C. Tesch-Römer, The role of deliberate practice in the acquisition of expert performance, Psychol. Rev. **100**, 363 (1993).

[31] P. Barniol and G. Zavala, Test of understanding of vectors: A reliable multiple- choice vector concept test, Phys. Rev. ST Phys. Educ. Res. **10**, 010121 (2014).

[32] R. D. Knight, The vector knowledge of beginning physics students, Phys. Teach. **33**, 74 (1995).

[33] B. D. Mikula and A. F. Heckler, Student difficulties with trigonometric vector components persist in multiple populations, *Proceedings of the 2013 Physics Education Research Conference, Portland, OR*, pp. 253–256.

[34] A. F. Heckler and T. M. Scaife, Adding and subtracting vectors: The problem with the arrow representation, Phys. Rev. ST Phys. Educ. Res. **11**, 010101 (2015).

[35] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.12.010134 for examples of training materials for vector essential skills and the DotCross vector assessment.

[36] E. Zell and Z. Krizan Do people have insight into their abilities? A metasynthesis, Perspect. Psychol. Sci. **9**, 111 (2014).

[37] J. A. Kulik Integrating findings from different levels of instruction, http://files.eric.ed.gov/fulltext/ED208040.pdf (1981).

[38] R. P. Niemiec and H. J. Walberg The effects of computers on learning, Int. J. Educ. Res. **17**, 99 (1992).

[39] 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).

[40] 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).

[41] J. L. Docktor, J. P. Mestre, and B. H. Ross, Impact of a short intervention on novices' categorization criteria, Phys. Rev. ST Phys. Educ. Res. **8**, 020102 (2012).

[42] M. Kapur, Productive failure, Cognit. Instr. **26**, 379 (2008).

[43] K. VanLehn, S. Siler, C. Murray, T. Yamauchi, and W. B. Baggett, Why do only some events cause learning during human tutoring?, Cognit. Instr. **21**, 209 (2003).

[44] S. Lau and R. W. Roeser, Cognitive abilities and motivational processes in high school students' situational engagement and achievement in science, Educ. Assess. **8**, 139 (2002).

[45] V. Sawtelle, E. Brewe, and L. H. Kramer, Exploring the relationship between self-efficacy and retention in introductory physics, J. Res. Sci. Teach. **49**, 1096 (2012).

[46] N. J. Cepeda, N. Coburn, D. Rohrer, J. T. Wixted, M. C. Mozer, and H. Pashler, Optimizing distributed practice, J. Exp. Psychol. **56**, 236 (2009).