Using cueing from question pairs to engage students in reflective thinking: An exploratory study

Joss Ives* and Jared B. Stang
Dept. of Physics & Astronomy, University of British Columbia, 6224 Agricultural Road, Vancouver, BC, V6T 1Z1

In this exploratory study, guided by dual process theories of reasoning, we used a low-stakes diagnostic test in a large introductory calculus-based physics course to test the effectiveness of using multiple-choice question pairs to improve student performance on conceptual multiple-choice questions. As part of this study, we measured students' tendency to engage analytic reasoning via the Cognitive Reflection Test, a three-item questionnaire embedded in a start-of-term diagnostic. These pairs of questions used a common question stem to ask about different but related concepts that students often conflate, such as acceleration and force in the context of a collision. Focusing on three questions from previously piloted question pairs, and controlling for measures of student knowledge and tendency to engage analytic reasoning, we used mixed-effects logistic regression techniques to observe that students who received the question as part of a pair were 7.2 times (95% confidence interval [4.8, 10.9], $p < .001$) more likely to answer the question correctly relative to having the question alone. Furthermore, the intervention was more impactful for students with a lower tendency to engage analytic reasoning. These results have implications for the design of short-answer physics questions in learning and assessment situations.

2019 PERC Proceedings edited by Cao, Wolf, and Bennett; Peer-reviewed, doi.org/10.1119/perc.2019.pr.Ives

I. INTRODUCTION

This study was motivated by an anecdotal teaching observation in which students seemed to be able to answer a clicker question about forces in a collision between unequal masses correctly more often if they were cued to consider acceleration by an earlier question. This motivated us to try an intervention where we would use *Question Pairs*—pairs of questions on related topics in the same situation—to help students avoid incorrectly applying the reasoning from one topic to the other, such as applying reasoning about acceleration to a question about force. An example Question Pair from the study presented in this manuscript is shown in Figure 1.

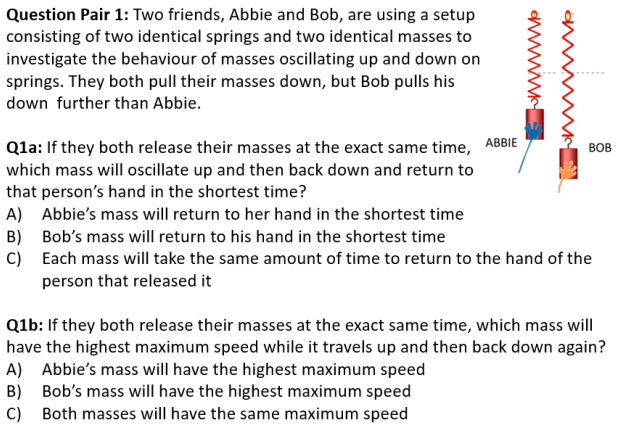

FIG. 1. An example Question Pair: A common stem is shared by a pair of questions about related topics that could be conflated by students. Q1a asks about the period of oscillation while Q1b asks about the maximum speed.

A theoretical framework that can be used to describe this phenomenon is that of dual process theories of reasoning (DPTOR) [1, 2]. In the context of conceptual physics problems, DPTOR describe a model of reasoning in which people can use fast (heuristic, intuitive or automatic) or slow (analytic, conscious or deliberate) reasoning to determine their answer to a question. From the perspective of DPTOR, the reasoning that leads students to an incorrect answer can come from a lack of relevant conceptual knowledge or from choosing an incorrect intuitively appealing answer without engaging in analytic reasoning. Our view of how a DPTOR lens applies to the context of conceptual physics questions is informed and enriched by previous work by Heckler [3]; Wood, Galloway, and Hardy [4]; and Gette and Kryjevskaia [5].

As discussed by Gette and Kryjevsjaia [5], "the analytic process could be engaged either by imposed external conditions (e.g., requirements to provide an explanation) or in the presence of a strong 'red flag.'" However, they go on to remind us that reasoners engaging in analytic thinking might still maintain their heuristic model due to confirmation bias.

The DPTOR literature refers to those that maintain a heuristic answer to a question, instead of engaging in the appropriate analytic thinking, as a cognitive misers [6]. As an example of an external condition designed to encourage reasoners to move beyond cognitive miserliness, Heckler [3] showed that the correct answer rate on a conceptual question could be increased dramatically by imposing a 3-second time delay between when participants saw a question and when they were allowed to answer in a computer-administrated testing environment. These questions required using the slope of a graph to answer correctly, but the graphs were designed to have an incorrect answer cued by noticeable differences in height, what Heckler calls a "salient, irrelevant dimension." Other examples of imposing external conditions include asking students to explain their answer to a question [7] and the Question Pair intervention upon which this study is based.

We hypothesize that our Question Pair intervention would be most beneficial to those most prone to cognitive miserliness, so we used the three-item Cognitive Reflection Test (CRT) [8] to measure a person's tendency to engage in analytic reasoning when an intuitive or heuristic answer exists.

II. METHODS

A. Experimental Design

Participants were students from an introductory calculus-based course on fluids, waves and energy, offered at the University of British Columbia - Vancouver (a large public research university in Canada) during the Fall, 2016 term. The data set consists of 698 student records, including those that withdrew or dropped. The filters applied to these data will be described after the main experimental design and data streams are described.

This manuscript focuses on an intervention where Question Pairs were administered across three versions of an end-of-course posttest using a randomized crossover design. The posttest was administered during each student's final laboratory session of the term. This same test was also used as a pretest at the start of the term, administered in the first lab. However, the pretest also had the three Cognitive Reflection Test questions embedded approximately one third of the way into the pretest. The pretest and posttest were each worth a 0.5% participation bonus toward their final grade. Participants were not required to do anything more than fill in their names and student numbers to earn these participation bonuses.

The posttest question bank consisted of nine Question Pairs: Three pairs were re-used from two pilot studies and six new pairs were developed. From these nine Question Pairs, three versions of the posttest were created and distributed randomly to students, thus each student was randomly assigned to condition A, B, or C (see Fig. 2). For a given Question Pair,

* joss@phas.ubc.ca

participants from one condition would receive both questions from the pair and thus be the treatment group. Participants from the other two conditions would each receive either the first or second questions of the pair and thus each be the control group for one of the questions from the pair.

	Pair 1		Pair 2		Pair 3		...	Pair 9	
	Q1a	Q1b	Q2a	Q2b	Q3a	Q3b		Q9a	Q9b
Condition A	Trt	Trt		Ctrl	Ctrl			Ctrl	
Condition B	Ctrl		Trt	Trt		Ctrl			Ctrl
Condition C		Ctrl	Ctrl		Trt	Trt		Trt	Trt

FIG. 2. The experimental design shows how the Question Pair treatment (Trt) and control (Ctrl) conditions are distributed across the three versions of the posttest (Conditions A-C).

Although the posttest consisted of questions drawn from a set of eighteen possible questions, from nine Question Pairs, this study focuses on only the three questions for which we saw statistically significant improvements in performance in our two pilot studies. These three questions will be referred to as *Intervention Questions*, whether they are used in the Treatment or Control conditions.

B. Predictor variables

This study uses a logistic regression model to predict student success at answering questions correctly at the level of the individual posttest Intervention Question. This model controls for pretest score, and Table I summarizes and describes all predictor variables considered for the model.

In addition to the variables emerging from the experimental design (see Section II A), we consider gender as a predictor in our models. It is important to include gender since score differences between men and women have been observed on both the CRT [8, 9] and physics concept inventories [10], where researchers have recently begun to deconstruct the underlying reasons for differences in the latter [11]. This literature raises the likelihood of effects within our measures that are related to student identities and not to their conceptual or analytical capabilities. Including gender as a predictor allows us to partially account for these effects, which we do not otherwise directly measure.

We recognize that gender identity exists on a spectrum and that considering it as binary is a simplification which does not fully represent the complexity which with individuals experience gender [12]. However, to operationalize gender, we relied on data from university records; at this institution, students are currently forced to choose between radio button options for "male" and "female" to indicate their "gender."

We observe statistically significant differences in favour of males on many of the assessment-based measures used in this study, including final exam (1.9%), CRT (14.5%), pretest (6.3%), posttest (7.5%), and the posttest score after removing Intervention Question performance (7.0%).

C. Data Filtering

The initial data set consists of 698 records for students who had registered for the course, which includes withdrawals and drops. Although we use Multiple Imputation techniques (described in the next section) to account for missing data, we do this only for predictor variables and not for the outcome variable, which is how each student performed on each of the Intervention Questions on the posttest. As a result, our first filter was to remove all records for which we did not have posttest data. This removed 100 records, of which 35 were drop or withdrawal students for which we have no pretest, posttest or final exam data. Another 27 of these filtered records were drop or withdrawal students for which we have pretest, but not posttest or final exam data. Next we removed students for which we had incomplete registration records. This removed one student, leaving 597 records after these first two filters.

We then removed pretest or posttest scores, but not records, for which students answered less than 80% of the questions, following the reasoning by Nissen *et al.* [13] that these tests would not represent students' actual knowledge. Thirteen pretests and no posttests were removed by this filter. After these filters were applied, 597 student records remained, with a 55.1% female population. Some records were missing partial data as will be described in the next section.

Finally, we had a further reduction in our data set as a result of study design issues, which became obvious only in hindsight. As shown in Fig. 2, each question had a version of the posttest for which it served as neither a treatment nor a control question. Unfortunately, our third version of the posttest served as neither treatment nor control for all three of the Intervention Questions, resulting in only 399 of the 597 post-filter student records being relevant for the analysis. We used all 597 records for the data imputation step since that did not focus on the individual Intervention Questions, but then had only 399 records for the logistic regression analyses that followed.

D. Multiple Imputation

To deal with missing pretest, CRT and final exam data, Multiple Imputation by Chained Equations was performed in R [14] using the MICE package [15]. Imputation is a statistically robust method of using the existing data to fill in the missing data over multiple data sets and then using Rubin's rules [16] to pool the averages and variances from analyses of the resulting imputed data sets.

The variables being imputed are as follows, with the amount of missing data being indicated in parentheses: pretest score (4.7%); pretest version (2.8%); CRT Question 1 (2.5%), 2 (2.8%) and 3 (5.2%); and final exam grade (1.0%). Based on a maximum of 5.2% missing data, we performed 10

TABLE I. Variables used and considered for the mixed-effects logistic regression model. The three Intervention Questions are those posttest questions for which we saw statistically significant effects due to the Question Pair intervention during previous pilot studies.

Variable	In final model	Role	Data type	Description
QCorrect	Y	Outcome	Binary	If the given Intervention Question was answered by the given student (ID) correctly on the posttest
Intervention	Y	Fixed effect	Binary	Treatment or Control
Gender	Y	Fixed effect	Binary	Forced-binary gender data taken from university records
QNum	Y	Fixed effect	Categorical	Labels for each of the three Intervention Questions
CRTscore	Y	Fixed effect	Integer [0,3]	Score on the three Cognitive Reflection Test questions
PostOther	Y	Fixed effect	Continuous [0,1]	Grade on the posttest questions other than the three Intervention Questions
Pre	Y	Fixed effect	Continuous [0,1]	Grade on the pretest
FinalExam	N	Fixed effect	Continuous [0,1]	Grade on final exam, taken after the posttest
ID	Y	Random effect	Categorical	Unique anonymous identifier for each student
PostVersion	Y	Random effect	Categorical	Which of three versions of the posttest the student took
PreVersion	Y	Random effect	Categorical	Which of three versions of the pretest the student took
Section	N	Random effect	Categorical	In which of the three lecture sections the student was registered

imputations, considered conservative for this amount of missing data [17]. In addition to the variables being imputed, the following additional variables were used in the imputation: gender, posttest version and overall score on the posttest.

III. ANALYSIS AND RESULTS

The following represents an exploratory data analysis. We focus our analysis on our Intervention Questions, which are the three questions for which we had previously seen statistically significant effects from our Question Pair intervention during two years of pilot studies. Thus, after using posttest data for the multiple imputation, we split the posttest results into two quantities: performance on each of the three Intervention Questions, and PostOther, the performance on the non-Intervention-Questions from the posttest.

A. Mixed-effects logistic regression

We used the GLMER function in R's LME4 package [18] to perform a mixed-effects logistic regression to predict if a student will answer an Intervention Question correctly, a binary outcome, based on a combination of various predictors. Mixed-effects refers to a type of regression model which includes fixed effects, which are fitting parameters of interest (e.g., Intervention, score on the CRT), and random effects, which are categorical variables that control for the data being nested or for the presence of repeated measures. Random effects variables are used when their variance needs to be accounted for in the model, but comparisons of outcome means between the levels of these categorical variables is not necessary.

The general form of our regression model is

$$\ln\left[\frac{P}{1-P}\right] = \beta_0 + \sum_i \beta_i X_i + \sum_{j,k} \beta_{jk} X_j X_k + \sum_l \epsilon_l,$$

where P is the probability that a given student has answered a given Intervention Question correctly. The β terms are the model fitting coefficients. The X_i variables represent the fixed effect variables. The double summation runs over indices (j and k) that are a subset of the i index and these terms represent possible interaction terms such as the interaction between Intervention and CRT score, which is a term motivated by our theoretical framework and required to test our hypothesis. Finally, the ϵ_l terms represent the random effect variables.

B. Model selection

Model selection is the process of determining which of the variables being considered should be included in the model. The relevant variables are summarized in Table I. We have used the model selection methods described by Theobald [19] to inform our process, which recommend selecting the most parsimonious model among the best fitting models, using Akaike information criterion (AIC) to compare competing models. AIC is an estimator of the relative quality of the fit of model to data, and rewards simpler models by accounting for the number of variables in the model. When choosing between competing models, the one with the lowest AIC and fewest parameters should be considered the preferred model with the additional caveat that models with $\Delta\text{AIC} \leq 2$ should be considered equivalent.

The first step in this process was to determine, for each variable being considered for the model, if it should be a fixed or random effect. We then built a fixed-effects-only model which tests our hypothesis explicitly, where an interaction term between Intervention and CRTscore was also included to test our hypothesis that the intervention would be most beneficial to those scoring low on the CRT. The random and fixed effects variables are summarized in Table I.

In looking at the results toward the end of our model selection process, we noticed that Gender was statistically sig-

nificant ($p < .001$) with results from the fixed-effects-only model being that females were 0.67 times (95% CI = [0.49, 0.89]) as likely as males overall to answer Intervention Questions correctly, independent of intervention condition. To investigate why large gender effects might be present despite controlling for PostOther, we created individual models which each included an additional interaction term of Gender and one of the other fixed effects. Through this process, we discovered that the interaction term between Gender and Pre was statistically significant and including it improved the overall model. Thus, we included Pre and this interaction term in all further steps of the model selection process. The improvement to the model due to this term persisted into the final model.

The next step was to include all possible combinations of random effects (intercepts) under consideration and to use the best model, as determined by AIC, to determine which random effects to keep in the model. It was decided that ID, the identifier for each student, should be kept in the model to account for the repeated measure of each student being asked up to three Intervention Questions. Additionally, to account for differences between versions of the diagnostic, PreVersion would always be kept when Pre was in the model, and PostVersion would always be kept when PostOther was in the model. Thus, only Section was tested for inclusion and it was found that it could be excluded.

The final step in the model selection was to select the appropriate fixed effects. A backward selection technique was used by which all fixed effects variables and interaction terms were first included in the model and then removed singularly, using AIC to test for the best model at each step. From this process it was determined that FinalExam was not needed in the model.

C. Results and discussion

Table II shows the results from our main model, with the three most relevant findings summarized below.

First, consistent with our hypothesis, the Question Pair intervention helped those with low CRT scores, as measured by three questions embedded in the pretest, more than those with high CRT scores, which is how the negative β coefficient for the interaction term Intervention×CRTscore (-0.30, SE = 0.15, $p = .043$) can be interpretted.

Second, the intervention had an overall beneficial effect across the Intervention Questions, which we quantified using the Intervention coefficient extracted after re-running the logistic regression without the Intervention×CRTscore term. Those in the Treatment condition were 7.2 (95% CI = [4.8, 10.9], $p < .001$) times more likely to answer an Intervention Question correctly than in the Control condition.

Third, we observe a gender difference in the predictive power of prettest score. Males answered the Intervention Questions correctly at a rate 7.7 (95% CI = [1.7, 36.0], $p = .0093$) times higher when scoring 100% on the Pre as when scoring 0%. In contrast, females answered these questions correctly at a rate which is independent of pretest score (Odds Ratio = 0.66, [0.19, 2.27], $p = .51$). This second result was determined by re-running the model with female, instead of male, as the baseline gender, but this same result could also have been determined through combining the β terms for Pre and Gender×Pre and then converting that to an odds ratio.

Additionally, we note that the Question Pair intervention can be beneficial if the Intervention Question is asked first or second as evidenced by statistically significant impacts for questions which are first in their pair (QNUM2) or second in their pair (QNUM1 and QNUM3).

Overall, we see that our exploratory analysis using previously piloted questions has provided strong evidence that the Question Pair intervention can benefit student performance. A reproduction study is needed to determine how confident we can be that this intervention benefits those with low CRT scores more than high CRT scores. Further investigations are also needed to study the observed gender effects, including exploring to what degree the Intervention Questions may be gendered.

ACKNOWLEDGEMENTS

We would like to thank Jim Carolan for running diagnostic question validation interviews and Jonathan Massey-Allard for valuable feedback on this manuscript. We would also like to thank Andrew Boudreaux for bringing the CRT to our attention and Mila Kryjevskaia for valuable discussions related to thinking about DPTOR in the context of physics.

TABLE II. Results from the main mixed-effects logistic regression model. The Gender coefficient represents the performance of female with respect to male for Pre held at 0. The Intervention coefficient represents the performance of Treatment with respect to Control for CRTscore held at 0.

	β (SE)	p	Odds [95% CI]
(Intercept)	-1.84 (0.62)	.0029	0.16 [0.05, 0.53]
Intervention	2.48 (0.33)	$< .001$	12.0 [6.2, 23.1]
CRTscore	0.29 (0.09)	.0013	1.33 [1.12, 1.58]
Intervention×CRTscore	-0.30 (0.15)	.043	0.74 [0.56, 0.99]
PostOther	1.06 (0.39)	.0063	2.89 [1.35, 6.20]
Gender	1.00 (0.57)	.080	2.72 [0.89, 8.35]
Pre	2.04 (0.78)	.0093	7.73 [1.66, 36.0]
Gender×Pre	-2.45 (0.90)	.013	0.09 [0.01, 0.59]
QNUM2	0.31 (0.23)	.17	1.36 [0.87, 2.15]
QNUM3	-0.93 (0.17)	$< .001$	0.39 [0.28, 0.55]

[1] J.S.B.T. Evans, The heuristic-analytic theory of reasoning: Extension and evaluation, Psychon. Bull. Rev. **13**, 378 (2006).
[2] D. Kahneman, *Thinking, Fast and Slow* (Anchor Canada, 2013).
[3] A.F. Heckler, The Role of Automatic, Bottom-Up Processes: In the Ubiquitous Patterns of Incorrect Answers to Science Questions, in *Psychology of Learning and Motivation* (Academic Press, 2011), Vol. 55, pp. 227-267.
[4] A.K. Wood, R.K. Galloway & J. Hardy, Can dual processing theory explain physics students' performance on the Force Concept Inventory? Phys. Rev. Phys. Educ. Res. **12**, 023101 (2016).
[5] C.R. Gette & M. Kryjevskaia, Establishing a relationship between student cognitive reflection skills and performance on physics questions that elicit strong intuitive responses, Phys. Rev. Phys. Educ. Res. **15**, 010118 (2019).
[6] M.E. Toplak, R.F. West & K.E. Stanovich, Assessing miserly information processing: An expansion of the Cognitive Reflection Test, Thinking and Reasoning **20**, 147-168 (2014).
[7] J. Ives & J.B. Stang, Engaging reflective thinking during exam-like situations: slowing students down on short-answer questions increases performance. Poster presented at: GIREP-MPTL conference 2018; 2018 Jul 9-13; San Sebastian, Spain.
[8] S. Frederick, Cognitive reflection and decision making, Journal of Economic Perspectives **19**, 25-42 (2005).
[9] D.C. Zhang, S. Highhouse, & T.B. Rada, Explaining sex differences on the Cognitive Reflection Test, Personality and Individual Differences **101**, 425-427 (2016).
[10] A. Madsen, S.B. McKagan, & E.C. Sayre, Gender gap on concept inventories in physics: What is consistent, what is inconsistent, and what factors influence the gap? Phys. Rev. ST Phys. Educ. Res. **9**, 020121 (2013).
[11] R. Henderson, J. Stewart, & A. Traxler, Partitioning the gender gap in physics conceptual inventories: Force Concept Inventory, Force and Motion Conceptual Evaluation, and Conceptual Survey of Electricity and Magnetism, Phys. Rev. Phys. Educ. Res. **15**, 010131 (2019).
[12] A.L. Traxler *et al.*, Enriching gender in physics education research: A binary past and a complex future, Phys. Rev. Phys. Educ. Res. **12**, 020114 (2016).
[13] J.M. Nissen *et al.*, Comparison of normalized gain and Cohen's d for analyzing gains on concept inventories, Phys. Rev. Phys. Educ. Res. **14**, 10115 (2018).
[14] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2019. https://www.R-project.org
[15] S. van Buuren, K. Groothuis-Oudshoorn, mice: Multivariate Imputation by Chained Equations, in R. J. of Stat. Softw. **45**, 3 (2011).
[16] D.B. Rubin, *Multiple Imputation for Nonresponse in Surveys* (New York: John Wiley and Sons, 1987).
[17] T.E. Bodner, What improves with increased missing data imputations? Structural Equation Modeling: A Multidisciplinary J. **15**, 651 (2008).
[18] D. Bates, M. Mäachler, B. Bolker, and S. Walker, Fitting Linear Mixed-Effects Models Using lme4, Journal of Statistical Software **67**, 1 (2015).
[19] E. Theobald, Students are rarely independent: When, why, and how to use random effects in discipline-based education research, CBE Life Sci. Educ. **17**, 1 (2018).