Cognitive reflection and physics student performance

David M. Harrison
Dept. of Physics, Univ. of Toronto, Toronto, ON M5S 1A7, Canada
david.harrison@utoronto.ca

Frederick's Cognitive Reflection Test (CRT) assesses an individual's ability to suppress an intuitive and spontaneous wrong answer in favor of a reflective and deliberative right answer. We asked questions from the CRT on the mid-term test and on the final examination in an introductory physics course, and looked at correlations with student performance. We measured performance by grades on the test and examination, by scores on the Force Concept Inventory (FCI), and by normalised gains on the FCI. CRT performance was correlated with test and examination grades by over a full letter grade. We found a correlation with FCI scores. The correlation with gains on the FCI was hardly statistically significant.

I. INTRODUCTION

The Cognitive Reflection Test (CRT) was developed by Frederick to assess an individual's ability to suppress an intuitive and spontaneous wrong answer in favor of a reflective and deliberative right answer.[1] It consists of three items:

1. A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost? _____ cents
2. If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets? _____ minutes
3. In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake? _____ days

Psychologists often use a taxonomy of cognition using the terms *System 1* and *System 2*.[2] System 1 is fast, automatic, effortless, and largely unconscious. System 2 is slow, logical, effortful and conscious. System 2 is often taken to be evolutionarily recent.[3] We live most of our lives under the control of System 1, which is capable of allowing us to drive a car on an empty road, recognize whether another person is frowning or smiling, understand simple sentences, etc. An expert physics problem solver can often intuitively solve or at least outline how to solve a physics problem using System 1. When a situation is too complex for System 1, it invokes System 2. Examples include driving a car in a blizzard, looking for a woman with white hair in a crowd, or multiplying 17×24 in your head. System 1 is often called a "miserly" information processing system.

System 1 will give incorrect answers to the three questions on the CRT (10 cents, 100 minutes, and 24 days respectively), while invoking System 2 will give correct answers (5 cents, 5 minutes, and 47 days). In Ref. 1 Frederick reports that when given the CRT, undergraduate students do surprisingly poorly. The best result was students at MIT, where 48% answered all 3 questions correctly, and the worst was at the University of Toledo, where 5% answered all 3 questions correctly. These results have been widely replicated.[4]

Some circumstances will "trigger" an invocation of System 2. For example, the CRT was administered to 40 Princeton undergraduates. One half of the students were given the test in an easy-to-read black Myriad Web 12-point font, while the other half were given a difficult-to-read 10% gray italicized Myriad Web 10-point font. About 90% of the students who saw the easy-to read version made at least one mistake on the test, but the proportion dropped to 35% when the font was barely legible. The cognitive strain of reading the difficult-to-read version apparently invoked System 2 which, when invoked, then went on to correctly answer the questions.[5]

In Physics Education Research (PER), the Force Concept Inventory (FCI) is a well-known instrument to evaluate the quality of teaching. The FCI was introduced by Hestenes, Wells and Swackhammer in 1992,[6] and was updated in 1995.[7] The instrument measures the conceptual understanding of Newtonian mechanics. A common methodology is to administer the instrument at the beginning of a course, the "precourse", and again at the end, the "postcourse", and to examine the gain. Both the CRT and the FCI were given to 148 students at the University of Edinburgh in Scotland. The CRT and the precourse FCI were administered separately on the web, and for the CRT the students were informed that they were being given "three quick questions to try out the numerical response capability of the new system." A positive correlation was found between the CRT and FCI scores.[8]

In this study, we asked two questions from the CRT, one on a mid-term test and the other on the final examination, in an introductory physics course. We were interested in student performance on the questions in the context of a test, where hopefully the students have been using System 2 to answer the questions. Thus System 2 has perhaps been "triggered". In addition, the question counts for a very small but non-zero percentage of the grade on the test or exam, which might influence the number of students who answer correctly. We also examine performance on the CRT question compared to performance on the test and exam. Finally, we replicate the results of the correlation between CRT and FCI scores of Ref. 8.

II. METHODS

The course studied here is the first of a two-semester sequence on introductory physics intended primarily for life science students. There is a separate course for students intending to go on in physics. Our course is calculus based, and the textbook is Wolfson.[9] Clickers, Peer Instruction,[10] and Interactive Lecture Demonstrations[11] are used

extensively in the classes. The session that is studied here was held in the summer of 2016. Although the summer version of the course has a compressed 6-week format compared to a normal 12-week term for the fall version of the course, and the student demographics for the two sessions are somewhat different, previously we have shown that the two versions are roughly comparable in terms of student performance.[12]

In addition to the classes, traditional tutorials and laboratories have been combined into a single active learning environment, which we call *Practicals*.[13] In the Practicals students work in small teams on conceptually based activities using a guided discovery model of instruction, and whenever possible the activities use a physical apparatus or a simulation. Most of the activities are similar to those of McDermott[14] and Laws[15], although we also spend some time on uncertainty analysis and on experimental technique such as is found in traditional laboratories.

A third major component of the course is a weekly homework assignment. We use *MasteringPhysics*[16] and the typical weekly assignment takes the students about two hours to complete. Although we use some of the tutorials provided by the software to help student's conceptual understanding, the principle focus of most homework assignments is traditional problem solving, both algebraic and numeric. We expect most students do these assignments as individuals, although we do not discourage the students from working on them together in a study team.

We gave a very slightly modified version of CRT Question 3 on the 90-minute mid-term test. The modified question, with the change underlined, was:

> In a lake, there is a <u>very small</u> patch of lily pads. But every day the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long did it take for the patch to cover half the lake? _____ days

This question was the very last one asked, and was graded correct/incorrect, with correct responses given 1 point. The test was marked out of a total of 50 points. The other 49 points were divided between 10 multiple-choice questions worth 4 points each, and a long-answer question of 4 parts worth a total of 9 points. The breakdown of the types of questions on the test, excluding the CRT one, was:

- 4 algebraic problems
- 4 numeric problems
- 6 conceptual questions with no or only minimal use of formulae or calculations

Although the CRT question counted towards the students' grade on the test, below we present the grades with the CRT question excluded. This grade is normalised by dividing by 0.49, so the values are percentages.

We gave a modified version of CRT Question 2 on the 2-hour final examination. The question was:

> 5 identical gas-fired electric generators produce 5 MJ of energy in 5 seconds.
> How long would it take 100 such generators to produce 100 MJ? _____ seconds

As with the CRT question on the mid-term test, this question was also the last one asked, and was graded correct/incorrect, with correct responses given 1 point. The exam was marked out of a total of 100 points. The other 99 points were divided between 14 multiple-choice questions worth 5 points each, and a long-answer question of 7 parts worth a total of 29 points. The breakdown of the types of questions on the exam, excluding the CRT one, was:

- 5 algebraic problems
- 4 numeric problems
- 13 conceptual questions with no or only minimal use of formulae or calculations

As with the test grades, below we present the examination grades with the CRT question excluded and normalised by dividing by 0.99, so the values are percentages.

We also gave the Force Concept Inventory (FCI) to students in PHY131 in a precourse/postcourse protocol. The students were given one-half a point, 0.5%, towards their final grade in the course for answering all questions on the precourse FCI, regardless of what they answered, and given another one-half point for answering all questions on the postcourse FCI also regardless of what they answered. The precourse was given on the day of the first class of the term, and the postcourse on the last day of the term. Below all FCI scores are in percent.

On the precourse FCI, in addition to the standard 30 questions we asked some non-graded questions about the students, their background, and their reasons for taking the course. One question and the percent responding was:

> What is your gender?
> A. male (35%)
> B. female (65%)
> C. neither of these are appropriate for me (0%)

It is important to realize that the mid-term test and final examination measure somewhat different things than the FCI. We try to avoid "plug and chug" problems on our tests, so we hope that some conceptual understanding as measured by the FCI is a pre-requisite. But the problems require skills beyond just conceptual understanding. Also, the "conceptual" questions on the tests are of quite a different character than the questions on the FCI. They are typically focused much more narrowly on particular content. In this respect they resemble the ConcepTests that are used in the Peer Instruction process more than they resemble the FCI questions.[17] Often our conceptual test questions are based on ConcepTest ones we have used in class.

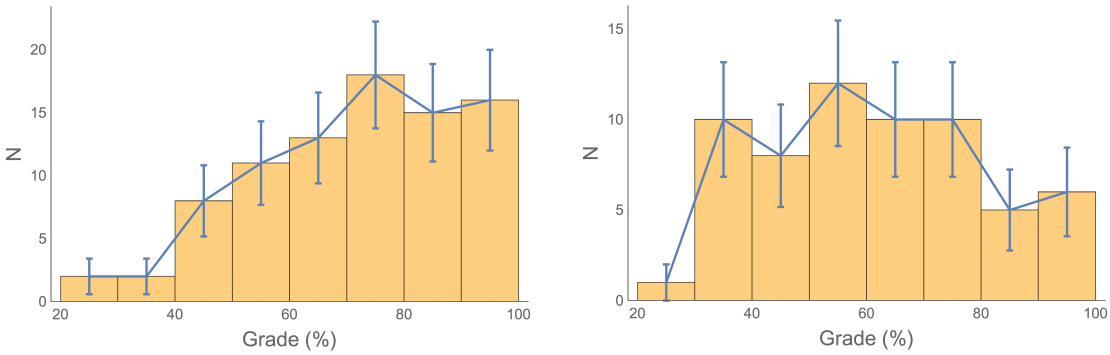
III. RESULTS

First we discuss the results of the mid-term test, then the final examination, and finally the FCI scores. As part of the analysis of the final examination we will re-visit the mid-term test results.

A. Mid-term Test

147 students wrote the test. 85 students = $(58 \pm 6)\%$ answered the CRT question correctly, and 62 students = $(42 \pm 5)\%$ answered it incorrectly; the stated uncertainties are $100\sqrt{N}\,/\,N_{\text{tot}} = 100\sqrt{N}\,/147$ where N is the number of students in the sample. Five students had perfect grades of 100% excluding the CRT question, and four of them answered the CRT question correctly. The intuitive answer to the CRT question is 24 days, but slightly less than one-half of our students gave this answer. For example, the one student who scored 100% of the test not counting the CRT question but gave a wrong answer to the CRT question tried to use formulae like $A_L = A_P(48)^2$ but ended up answering 34 days.

Figure 1a shows the histogram of test grades for students who answered the CRT question correctly, and Figure 1b is for students who answered it incorrectly. The bins of the histograms are for grades of 20 – 29, 30 – 39, etc. except for the last one, which is for grades of 90 – 100. The displayed uncertainties are the square root of the number of students in each bin.



(a) (b)

Figure 1. Test grades. (a): students who answered CRT question 3 correctly. (b): students who answered it incorrectly.

There is an issue as to how to best characterize distributions such as in Fig. 1. The Shapiro-Wilk test is used to determine if a distribution is Gaussian, i.e. "normal".[18] The test returns a p statistic, with higher values meaning the distribution is more Gaussian. For the distribution of Fig. 1(a), $p = 0.01$, and for Fig. 1(b) $p = 0.15$. The low value for the data of Fig. 1(a) is largely due to the fact that there are so many high grades, so the

distribution does not "turn over" for higher values. We will assume that a Gaussian distribution is appropriate for both, and will use the mean values to characterize the data.[19]

Table I shows the mean grade on the mid-term test, excluding the CRT question, for all students, for students who answered the CRT question correctly, and for students who answered the CRT question incorrectly. The uncertainties are the "standard errors of the mean,"[20] $\sigma_m = \sigma / \sqrt{N}$. At the University of Toronto, grades between 60% and 69% are defined as C, and grades between 70% and 79% are defined as B. Thus the overall mean on the test excluding the CRT question, 66.9%, is consistent with other courses at the university.

Table I. Mid-term test grades

	Mean Test Grade (%)
All students	66.9 ± 1.6
CRT question 3 correct	71.3 ± 2.0
CRT question 3 incorrect	60.7 ± 2.4

Figure 2 shows the boxplot of test grades for students who answered the CRT question correctly and incorrectly. The "waist" on the boxplot is the median, the "shoulder" is the upper quartile, and the "hip" is the lower quartile. The vertical lines extend to the largest/smallest value less/greater than a heuristically defined outlier cutoff.[21] The "notch" around the median value represents the statistical uncertainty in the value of the median, which is $1.58 \times \mathrm{IQR} / \sqrt{N}$, where IQR is the interquartile range and N is the number of students in the sample.[22] This uncertainty is roughly taken to indicate a 95% confidence interval, i.e. it is equivalent to $2 \times \sigma_m$ for a normal distribution.

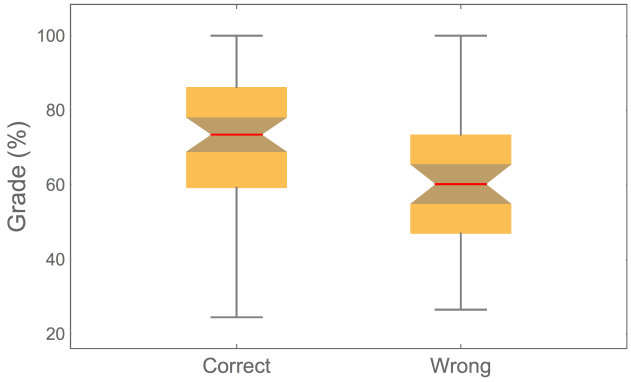


Figure 2. Boxplot of test grades for students who answered CRT question 3 correctly and incorrectly

We examined the effect size of the difference between the test grades for students getting or missing the CRT question using Cohen's d.[23] It is defined as:

$$d = \frac{|\text{mean}_1 - \text{mean}_2|}{\sigma_{\text{pooled}}} \quad (1)$$

where:

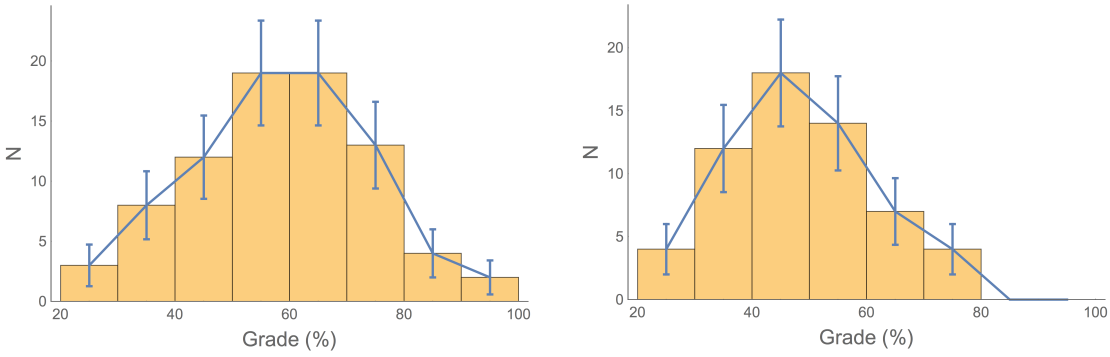
$$\sigma_{pooled} = \sqrt{\left(\sigma_1^{\,2} + \sigma_2^{\,2}\right)/2} \quad (2)$$

The result is $d = 0.57$, which is heuristically characterized as a "medium" effect. The 95% confidence interval range for d is 0.23 – 0.91; since this range does not include zero, the difference is statistically significant.

B. Final Examination and "Matched" Mid-Term Tests

139 students wrote the final examination. 80 students = $(58 \pm 6)\%$ answered the CRT question correctly, and 59 students = $(42 \pm 6)\%$ answered it incorrectly. Two students had exam grades > 90%, and both of them answered the CRT question correctly. The intuitive answer to the CRT question is 100 sec., and about 25% of the students who answered the question incorrectly gave this answer.

Figure 3a shows the histogram of examination grades for students who answered the CRT question correctly, and Figure 3b is for students who answered it incorrectly.



(a) (b)

Figure 3. Examination grades. (a): students who answered CRT question 2 correctly. (b): students who answered it incorrectly.

The Shapiro-Wilk test for the distribution of Fig 3(a) gives $p = 0.73$, and Fig. 3(b) gives $p = 0.55$. Thus we use the means to characterize the distributions.

Table II shows the mean grade on the final examination, excluding the CRT question, for all students, for students who answered the CRT question correctly, and for students who

answered the CRT question incorrectly. The overall mean of 54% was somewhat lower than we intended, and was adjusted in calculating final grades in the course.

Table II. Final examination grades

	Mean Exam Grade (%)
All students	54.1±1.4
CRT question 2 correct	58.6±1.8
CRT question 2 incorrect	47.9±1.8

Figure 4 shows the boxplot of final examination grades for students who answered the CRT question correctly and incorrectly.

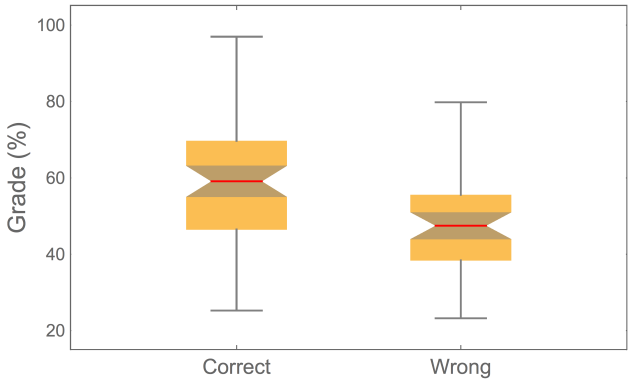


Figure 4. Boxplot of exam grades for students who answered CRT question 3 correctly and incorrectly

The Cohen $d = 0.71$, which is a "medium" difference. The 95% confidence interval was 0.39 – 1.06; since this does not include zero the difference is statistically significant.

134 students wrote both the mid-term test and the final examination. We formed a "matched" dataset of these students. For each student we calculated a CRT Score, the number of CRT questions answered correctly on the test and final exam, and also calculated the Test and Exam Average Grade. Table III summarises.

Table III. CRT Score and mean of the Test and Exam Average Grade for "matched" students

CRT Score	**N**	**Mean of the Test and Exam Average Grade (%)**
All students	134	61.5±1.4
0	31	54.2±2.9
1	47	60.0±1.9
2	56	67.0±2.2

Fitting the means and their uncertainties versus CRT scores to a straight line gave a slope $m = 6.5 \pm 1.8$ with $\chi^2 = 0.052$ for 1 degree of freedom. Figure 5 is a boxplot of the Test and Exam Average Grades for different CRT Scores.

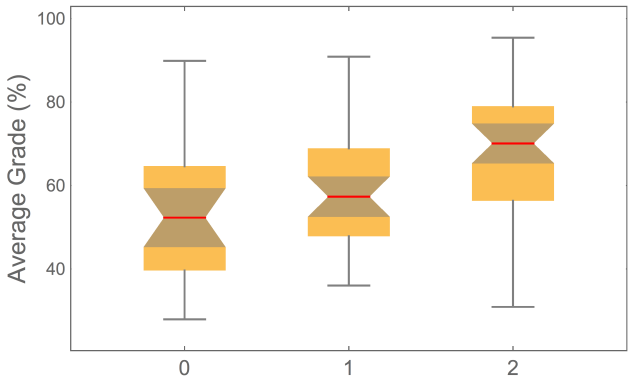


Figure 5. Test and Exam Average Grades by CRT Score

As discussed, the test and exam both have problems and topic-focused conceptual questions. We examined these two types of questions separately. Table IV summarises.

Table IV. Problem and conceptual question average grades for matched students

CRT Score	Mean of the Test and Exam Average Grade (%)	
	Problems	Conceptual Questions
All students	56.1 ± 1.4	68.4 ± 1.5
0	47.0 ± 2.5	62.9 ± 2.9
1	55.7 ± 2.4	67.7 ± 2.2
2	62.8 ± 2.2	72.1 ± 2.5

Although the grades on the problems are lower than for the conceptual questions, we have no reason to believe this is anything but an artifact of the difficulty of the questions that we asked. Fitting the problem grades vs. CRT Score to a straight line gave a slope $m = 7.9 \pm 1.7$ with $\chi^2 = 0.075$ for 1 degree of freedom; fitting the conceptual questions gave a slope $m = 4.6 \pm 1.9$ with $\chi^2 = 0.0047$ for 1 degree of freedom. The difference in the slopes is $(7.9 \pm 1.7) - (4.6 \pm 1.9) = 3.3 \pm 2.5$. Perhaps the CRT score correlates more strongly with the performance on the problems, but the uncertainties are large.

D. FCI

156 students wrote the precourse FCI, which was almost all students registered at that time, and 129 students wrote the postcourse FCI, which was almost all students still registered. The dropout rate of 17% is typical for this course.

The histograms of the FCI scores, which are not shown, look almost identical to the results for the 2013 session of the course shown in Ref. 12. As discussed in Ref. 12, those distributions are far from Gaussian, so the median is a better way of summarising the

scores than the mean. For the data of this paper, the Shapiro-Wilk test gave very low values: $p = 1.9\times10^{-9}$ for the precourse scores and $p = 2.8\times10^{-4}$ for the postcourse ones.

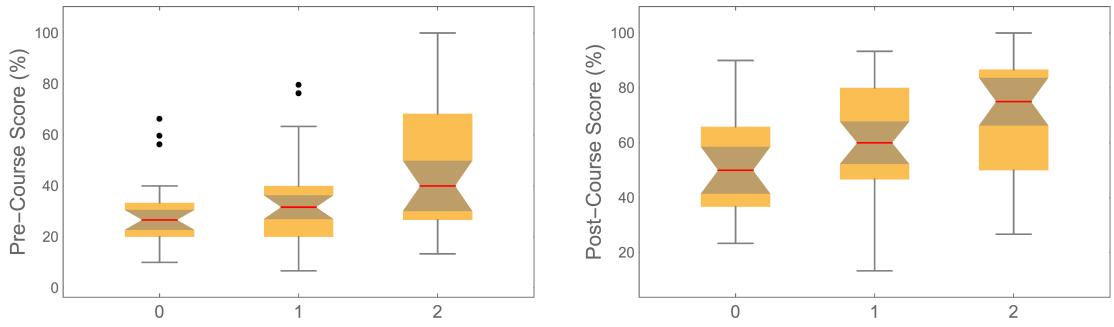
For the precourse FCI the median score was $(33.3\pm4.2)\%$, and the postcourse FCI median was $(63.3\pm5.6)\%$. These values are both similar to those for the 2013 session. As already mentioned, the uncertainty in the median is taken to be $1.58\times\mathrm{IQR}/\sqrt{N}$, where IQR is the interquartile range and N is the number of students in the sample; this uncertainty is taken to be indicate roughly a 95% confidence interval, i.e. the equivalent of $2\times\sigma_m$ for a normal distribution.

We formed a "super-matched" dataset of the 119 students who did the precourse FCI, the postcourse FCI, the mid-term test, and the final examination. Table V shows the median FCI scores for all students and for different values of the CRT Score..

Table V. FCI median scores for super-matched students

CRT Score	N	Precourse FCI (%)	Postcourse FCI (%)
All	119	33.3 ± 4.6	60.0 ± 5.3
0	29	26.7 ± 3.9	50.0 ± 7.8
1	46	31.7 ± 4.7	60.0 ± 7.3
2	44	40.0 ± 9.7	75.0 ± 8.7

Figure 6a shows the boxplot for the precourse FCI scores for different values of the CRT Score. The dots represent scores that are outside the cutoffs, and are therefore considered to be outliers; fitting the median values and their uncertainties versus CRT Scores to a straight line gave a slope $m = 6.0\pm4.4$ with a $\chi^2 = 0.055$ for 1 degree of freedom. Figure 6b is the boxplot for the postcourse FCI scores; fitting the median values to a straight line gave a slope $m = 12.4\pm5.8$ with a $\chi^2 = 0.071$ for 1 degree of freedom. The difference in the slopes is $(12.4\pm5.8)-(6.0\pm4.4) = 6.4\pm7.3$ which is zero within uncertainties. The large uncertainties in the slopes arise because of the large uncertainties in some of the median values, which in turn arise because of the large interquartile ranges for some distributions.



(a) (b)

Figure 6. Boxplots of FCI scores for different CRT Scores for super-matched students. (a) Precourse (b) Postcourse

The standard way of measuring student gains from the precourse to the postcourse FCI is from a seminal paper by Hake.[24] It is defined as the gain normalised by the maximum possible gain:

$$G = \frac{\mathrm{PostCourse\%} - \mathrm{PreCourse\%}}{100 - \mathrm{PreCourse\%}} \tag{3}$$

Clearly G cannot be calculated for precourse scores = 100. This was 1 student in our course.

One hopes that the students' performance on the FCI is higher at the end of a course than at the beginning. As also discussed more fully in Ref. 12, somewhat similar to Hake in Ref. 24 we define the median normalised gain:

$$< g >_{\mathrm{median}} = \frac{< \mathrm{PostCourse\%} > - < \mathrm{PreCourse\%} >}{100 - < \mathrm{PreCourse\%} >} \tag{4}$$

where the angle brackets on the right-hand side indicate medians. Table VI shows the values of $< g >_{\mathrm{median}}$ for all students and for different values of the CRT Scores. The difference between CRT Scores of 2 and 0 is $(0.58 \pm 0.16) - (0.32 \pm 0.11) = 0.26 \pm 0.19$, which is barely greater than zero within uncertainties.

Table VI. FCI gains for super-matched students

CRT Score	$< g >_{\mathrm{median}}$
All	0.40 ± 0.09
0	0.32 ± 0.11
1	0.41 ± 0.12
2	0.58 ± 0.16

IV. DISCUSSION

We have examined the correlation between CRT Scores and test and exam grades in a number of ways, and in all cases it is just over a full letter grade. We see no significant indication that the grade – CRT correlation is different for the problems than for the conceptual questions.

Most of our students answer test questions more-or-less in the order in which they appear on the test.[25] Since the CRT questions were the last ones given, cognitive fatigue due to answering the previous questions on the test could lead to difficulties in exerting the self-control necessary to suppress System 1 and invoke System 2 processing on these questions. It seems that self-control requires glucose in the brain as an energy source, and cognitive fatigue is related to a deficiency of glucose due to previous acts of self-control.[26]

Thus, there are at least two competing factors in the CRT questions. The fact that they were part of a physics test and count for a small grade might trigger the students to give a reflective and correct answer, while the fact that they occur last on the test might mean that cognitive fatigue will suppress the reflective process.

As discussed, in Ref. 8 Wood, Galloway, and Hardy report on CRT results given to introductory physics students at the Univ. of Edinburgh in a "non-threatening" environment: the results did not count for student grades in the course, and they obfuscated the intent by telling the students they were just testing the capabilities of a new system. Table VII compares the percentage of students answering CRT questions 2 and 3 correctly at Edinburgh to the modified versions of those questions given to our students.

Table VII. Comparing CRT performance at Edinburgh and Toronto

	Edinburgh	**Toronto**
CRT Question 2 Correct	71%	$(58 \pm 6)\%$
CRT Question 3 Correct	86%	$(58 \pm 6)\%$

The Toronto results are consistent with those reported in the meta-study of Ref. 4, while the Edinburgh students did much better. For CRT Question 3 the percentage of students who gave the intuitive and wrong answer were 45% at Edinburgh and slightly under 50% at Toronto. But for CRT Question 2 it was 49% at Edinburgh but only about 25% at Toronto. Perhaps our modification of CRT Question 2, making it look more like a physics question than the original form, is the reason for the discrepancy between the Edinburgh and Toronto wrong answers.

There are other factors to be aware of here.

1. The meta-study of Ref. 4 shows that male students perform better than females on the CRT. The Edinburgh students were about 80% male and 20% female; for the Toronto course, as discussed, the genders were 35% male and 65% female. For our super-matched students, the mean CRT Score for females was 1.00 ± 0.09, and for males it was 1.38 ± 0.12. Thus, the males outperformed the females by $(1.38 \pm 0.12) - (1.00 \pm 0.09) = 0.38 \pm 0.15$.
2. The meta-study shows that performance on the CRT is negatively correlated with whether the questions were given at the beginning or the end of a longer experiment; at Edinburgh the CRT was essentially stand-alone, while in Toronto the questions were at the end of a long test or exam.
3. As discussed in, for example, Ref. 2, the efficacy of System 2 is impaired by time pressure. Although we designed our tests to try to minimize time pressure and about 10% of our students left early on both the mid-term test and the final examination, students will still feel under some pressure to answer all the questions.
4. If one informally asks one of the CRT questions to friends, the badly math-phobic ones will reject even trying to think about it. These are also people who sometimes say, "I don't do Sudoku puzzles because I don't do math," although that puzzle has nothing to do with math. Although this level of math-phobia is an extreme case, math anxiety in general has been shown to have a negative correlation with CRT performance.[27] When we interview students in serious difficulty in an introductory physics course as measured by test performance, we discover that at least in the context of physics that many suffer from various degrees of math-anxiety.[28] Perhaps the correlation of CRT scores with test and exam grades is partly related to math-anxiety of our poorer students.

Despite the differences in student populations between the Edinburgh and Toronto students, and especially the difference in the way the CRT questions were delivered to the students, our data confirms the correlation between CRT performance and FCI scores. For our data, the correlation was roughly the same for the precourse and the postcourse FCI.

Although our data on the values of $< g >_{\text{median}}$ as a function of CRT Score shown in Table VI is perhaps suggestive of a correlation, the large uncertainties mean that the differences are hardly statistically significant; this result is slightly different than the Edinburgh one which concluded there was no correlation. Previously, we have shown that the normalised gains on the FCI are independent of many factors about the students, their background, and their motivation for taking the course.[29] However, the Edinburgh data on normalised gains used the mean of the FCI scores with uncertainties σ_m, while in Table V we used the median and the propagated uncertainties in the median. Calculating the normalised gains as Edinburgh did gives numbers whose correlation looks more like theirs. For a CRT Score of 0 the average normalised gain was 0.32 ± 0.06; for a CRT Score of 1 it was 0.40 ± 0.05; for a CRT Score of 2 it was 0.39 ± 0.08.

V. CONCLUSIONS

Any effect from administering the CRT questions on the test and exam instead of as a stand-alone instrument was not measurable. There is a strong correlation between how students perform on CRT questions and their grades on tests and exams: the difference is over a full letter grade. There is also a strong correlation with FCI performance. Normalised gains on the FCI are not strongly correlated with CRT performance.

To become better teachers, first we must understand the difficulties of our students. As Arnold Arons constantly reminded us, if we listen carefully our students will often tell us what those difficulties are.[30] The precourse FCI allows us to listen to an entire class about student misconceptions about Newtonian mechanics; the FCI has been one of the most important tools leading to the adoption of *interactive engagement* pedagogy. Lawson's Classroom Test of Scientific Reasoning[31] allows us to listen to a class about the students' Piagetian stage of cognitive development. The results have led Coletta[32] to further insights into effective pedagogy; in Ref. 28 we have also discussed this issue. Regarding the study of this paper, we believe that the CRT gives another perspective on psychological factors that are relevant for learning physics; we hope that sensitivity to the issue of intuitive versus reflective cognition will in future lead to ideas about how to improve the learning of our students.

ACKNOWLEDGEMENTS

We thank Eli Honig, Dept. of Physics, Univ. of Toronto, for useful discussions. We also benefited from discussions with Nicholas Rule, Dept. of Psychology, Univ. of Toronto. Zaheen Sadeq, Dept. of Physics, Univ. of Toronto, marked the CRT question on the mid-term test and supplied the data on the wrong answers by the students. Olinka Bedroya, Dept. of Physics, Univ. of Toronto, marked the CRT question on the final examination and supplied the data on the wrong answers by the students. April Seeley, Dept. of Physics, Univ. of Toronto, assisted in the data collection for this study.

REFERENCES

[1] S. Frederick, "Cognitive reflection and decision making," Jour. of Economic Perspectives **14**(2), 24 (2005).
[2] D. Kahneman, *Thinking, Fast and Slow* (Farrar, Straus and Giroux, 2011).
[3] S. Mithen, *The Cognitive Basis of Science* (Cambridge Univ. Press, 2002), pg. 23 ff.
[4] A recent meta-study of 118 studies comprising of 44,558 participants in 21 countries is P. Branas-Garza, P. Kujal, and B. Lenkei, "Cognitive Reflection Test: Whom, how, when," Munich Personal RePEc Archive Paper No. 68049 (November 25, 2015). Online at: https://mpra.ub.uni-muenchen.de/68049/ (Retrieved June 20, 2016)
[5] A.L. Alter, D.M. Oppenheimer, N. Epley and R.N. Eyre, "Overcoming Intuition: Metacognitive Difficulty Activates Analytic Reasoning", Jour. of Exptl. Psych .: General

136(4), 569 (2007).
[6] D. Hestenes, M. Wells and G. Swackhammer, "Force Concept Inventory," The Physics Teacher **30**(3), 141 (1992).
[7] Available from http://modeling.asu.edu/R&E/Research.html
[8] A.K. Wood, R.K. Galloway, and J. Hardy, "Can Dual Processing Theory Explain Physics Students' Performance on the Force Concept Inventory?" Phys. Rev. PER, forthcoming.
[9] R. Wolfson, *Essential University Physics*, 3rd ed. (Pearson, 2016).
[10] E. Mazur, *Peer Instruction: A User's Manual* (Addison-Wesley, New York, 1996).
[11] D.R. Sokoloff and R.K. Thornton, "Using Interactive Lecture Demonstrations to Create an Active Learning Environment," The Physics Teacher **35**(6), 340 (1997).
[12] J.J.B. Harlow, D.M. Harrison, and E. Honig, "Compressed-format compared to regular-format in a first-year university physics course," Am. J. Phys. **83**(3), 272 (2015).
[13] The U of T Practicals web site is: http://www.upscale.utoronto.ca/Practicals/ .
[14] L.C. McDermott, P.S. Schaffer and the Physics Education Group, *Tutorials in Introductory Physics* (Prentice Hall, New Jersey, 2002).
[15] P.W. Laws, *Workshop Physics Activity Guide* (Wiley, 2004). This type of pedagogy is also sometimes called SCALE-UP, TEAL, and other names.
[16] http://www.pearsonmylabandmastering.com/northamerica/masteringphysics/
[17] ConcepTest questions are included in Ref. 10. There are also a growing number of web sites that provide further questions suitable for Peer Instruction.
[18] S.S. Shapiro and M.B. Wilk, "An analysis of variance test for normality (complete samples)," Biometrika **52**(3 – 4), 591 (1965).
[19] A rough estimate of the biases in this assumption can be found by fitting the distributions to a Gaussian. For the distribution of Fig 1a, this gave a value of $\mu = 81.0 \pm 6.9$ with $\chi^2 = 1.94$ for 5 degrees of freedom, and for Fig 1b the mean was $\mu = 62.9 \pm 3.3$ with $\chi^2 = 6.04$ for 5 degrees of freedom. It is tempting to use these values for μ instead of the arithmetic means used in Table I.
[20] Although the phrase "standard error of the mean" is standard, particularly for students the word *error* is misleading since it implies that some mistake has been made. A better phrase is "standard uncertainty of the mean".
[21] There are various conventions for the cutoff definition. We use 1.5 times the inter-quartile range extending from the upper and lower quartiles, which was proposed in J.D. Emerson and J. Strenio, "Boxplots and Batch Comparison," in D.C. Hoaglin, F. Mosteller, and J.W. Tukey eds., *Understanding Robust and Exploratory Data Analysis* (Wiley-Interscience, Toronto, 1983), p. 58. This cutoff definition is the usual one.
[22] R. McGill, J.W. Tukey, and W.A. Larsen, "Variations of box plots," American Statistician **32**, 12 (1978). http://amstat.tandfonline.com/doi/pdf/10.1080/00031305.1978.10479236 (Retrieved November 15, 2014). Note that in this article the multiplier is 1.57, not 1.58: since the uncertainty itself is largely heuristic, the difference in these values is trivial. Also, in Ref. 12 and Ref. 29 we incorrectly omitted the factor of 1.58 entirely.
[23] J. Cohen, "A power primer," Psychological Bulletin **112**(1), 155 (1992).

[24] R.R. Hake, "Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses," Am. J. Phys. **66** (1998), 64.
[25] Shortly after the mid-term test, we asked the students the following clicker question: "When taking a test such as the PHY131 mid-term, which statement best describes you?" The choices and percent responding were: [A] I answer the question on the test in the order that they appear on the test. (2%) [B] I answer the questions more-or-less in the order that they appear, but if I get stuck on a question I skip and it and go on. Later I go back to that question. (78%) [C] I go through the test questions in order, but the first time through only answer the ones that are fairly easy. Then I go back to the other ones. (10%) [D] I look over the entire test first, and then choose the easiest questions to answer first. (7%) [E] None of these are how I take a test. (3%)
[26] See, for example, M.T. Gailliot, R.F. Baumeister, C.N. DeWall, J.K. Maner, E.A. Plant, D.M. Tice, L.E. Brewer, and B.J. Schmeichel, "Self-control relies on glucose as a limited energy source: Willpower is more than a metaphor," Journal of Personality and Social Psychology **92**(2), 325 (2007).
[27] K. Morsanyi, C. Busdraghi, and C. Primi, "Mathematical anxiety is linked to reduced cognitive reflection: a potential road from discomfort in the mathematics classroom to susceptibility to biases," Behavioral and Brain Functions **10**, 41 (2014).
[28] D.M. Harrison, "Factors correlated with students' scientific reasoning ability in an introductory physics course," in K.A. MacLeod and T.G. Ryan, eds., *The Physics Educator: Tacit Praxes and Untold Stories* (Common Ground Publishing, Champaign IL, 2016), p. 186. Available from: http://www.upscale.utoronto.ca/PVB/Harrison/CTSR_Factors/CTSR_Factors.pdf
[29] J.J.B. Harlow, D.M. Harrison, and A. Meyertholen, "Correlating student interest and high school preparation with learning and performance in an introductory university physics course," Phys. Rev. ST PER **10**, 010112 (2014).
[30] See, for example, Arons' epic trilogy: A.B. Arons, "Student Patterns of Thinking and Reasoning, Part One," Phys. Teach. **21**, 576 (1983), A.B. Arons, "Student Patterns of Thinking and Reasoning, Part Two," Phys. Teach. **22**, 21 (1984), and A.B. Arons, "Student Patterns of Thinking and Reasoning, Part Three," Phys. Teach. **22**, 88 (1984).
[31] A.E. Lawson, "The development and validation of a classroom test of formal reasoning," Jour. Res. Sci. Teaching **15**, 11 (1978). Available from: https://modelinginstruction.org/wp-content/uploads/2013/06/LawsonTest_4-2006.pdf
[32] V.P. Coletta, *Thinking in Physics* (Pearson, San Francisco, 2015).

Last revision: July 3, 2016