Factors correlated with students' scientific reasoning ability in an introductory university physics course

David M. Harrison

University of Toronto

Author Note

David M. Harrison, Department of Physics, University of Toronto

We thank Vincent Coletta, Loyola Marymount University, for sharing the data for his students' answers on each question on the CTSR, and for helpful comments on this paper. We have benefited from discussions with Ruxandra Serbanescu and Sabine Stanley, both of the Department of Physics, University of Toronto. April Seeley, Department of Physics, University of Toronto, assisted with the collection and initial analysis of our CTSR data.

Correspondence concerning this article should be addressed to David M. Harrison, Dept. of Physics, Univ. of Toronto 60 St. George St., Toronto ON. M5S 1A7 CANADA

Contact: david.harrison@utoronto.ca

Last revision: August 14, 2015

Abstract

We have used Lawson's Classroom Test of Scientific Reasoning to test the cognitive level of students in an introductory physics course at the University of Toronto. We also collected information about the students, their background, and their reason for taking the course, and looked for correlations with their ability to think scientifically. We found correlations in CTSR performance for gender and for whether or not the student took a senior-level high school physics course, but not for any other of the factors we examined. We argue that both of the correlations that we found are not an indicator of causation. We examine efforts to aid students' in learning to think scientifically, and call for greater testing of this capacity using the Pre-Course/Post-Course protocol that has already been widely used with instruments such as the Force Concept Inventory.

Keywords: scientific reasoning, Piaget, gender, high school physics

Factors correlated with students' scientific reasoning ability

in an introductory university physics course

Introduction

Inhelder and Piaget (1958) described the cognitive development of young people as consisting of four stages:

1. Sensorimotor (birth – 24 months). Learns that he/she is separate from the external world. Learns about object permanence.
2. Pre-operational (2 – 7 years). Can represent objects as symbols which can be thought of separately from the object. Can "make believe." Wants the knowledge of knowing everything.
3. Concrete Operational (7 – 11 years). Can reason logically about concrete events or objects. Acquires concepts of conservation of number, area, volume, and orientation.
4. Formal Operational (11 – 17 years and onwards). Can reason logically about abstract formal concepts. Can reason with ratios. Can do separation and control of variables. Can think about different points of view or reference frames. Can think about thinking.

The ability to use the ways of thinking, the operations, associated with Formal Operations is clearly necessary to do science in general and physics in particular. However, as Arnett (2010) wrote: "research has shown that not all persons in all cultures reach formal operations, and most people do not use formal operations in all aspects of their lives" (p. 89).

Lawson (1978) has developed a 24-question Classroom Test of Scientific Reasoning (CTSR) to probe whether students are at a Formal Operational stage of development. We gave the CTSR to the students in our 240-student introductory physics course at the University of Toronto. The course is intended primarily for students in the life sciences and is calculus based. We gave the test during the first week of classes of the summer 2014 session of the course, and the students were given one grade out of 100 for answering all the questions on the test, regardless of what they answered.

Seeing some of the questions from the CTSR may help clarify what is meant by Formal Operations. Appendix A shows 2 questions from the CTSR. Question 5 requires knowing about the constancy of the volume of a fixed quantity of liquid, and especially the ability to reason about ratios. Only 59% of our students answered this correctly. Question 11 probes the ability of the student to separate and control variables; 59% of our students answered this correctly. All the individual question results for our students are consistent with those for students at Loyola Marymount University in Los Angeles (V.P. Coletta, personal communication, January 21, 2014) and, as will be shown in the Discussion section, are probably similar to 1st year university students worldwide.

According to orthodox Piagetian analysis, the different aspects of Formal Operations are all consolidated together at more-or-less the same time. We see little significant sign of this in our results. For example, for the two questions in Appendix A, 59% of the class got each question correct. However for the 141 students who answered Question 5 correctly, 63% answered Question 11 correctly, hardly greater than the overall result of 59% correct. Similarly, 141 students answered Question 11 correctly, and 63% of them answered Question 5 correctly, again only slightly greater than the 59% for the whole class. Put another way, the intersection of the set of 141 students who answered Question 5 correctly and the set of 141 students who answered Question 11 correctly is 89 students. A true Piagetian would predict that students who could answer one of these questions correctly would also be able to answer the other one correctly. In a critique of the CTSR, Pratt and Hacker (1984), amongst other criticisms of the instrument, claim that the orthodox Piagetian view, that Formal Operations are unitary, will not be shown by the CTSR because of flaws in the test instrument itself. Nonetheless, we have found that even if one is not a true Piagetian, being aware of the types of thinking associated with Formal Operations allows us to understand with more sensitivity the difficulties that some of our students are having in their physics course and that the CTSR is a useful way of concentrating our attention on the issues that our students are struggling with.

In addition to the 24 questions on the CTSR, we appended 7 questions at the end asking the students about themselves, their background, and their reason for taking our course. The questions and percentage of students in each category are given in Appendix B.

Results

244 students wrote the CTSR, but 6 of them did not answer all seven questions about themselves and their background: these 6 students did not receive credit for the assignment and are dropped from all analysis below. The range of scores was 25 – 100%. Figure 1 shows the results of the CTSR. The data are displayed in two ways: as a vertical bar histogram and a series of dots and lines. Below when we compare two sets of data in the same plot, the second form is easier to see and is what will be used. The displayed

uncertainties are $\sqrt{N}$ where N is the number of students in each bin. The bin values are (20-29), (30-39), etc. except for the last bin, which are scores from 90 to 100 inclusive.

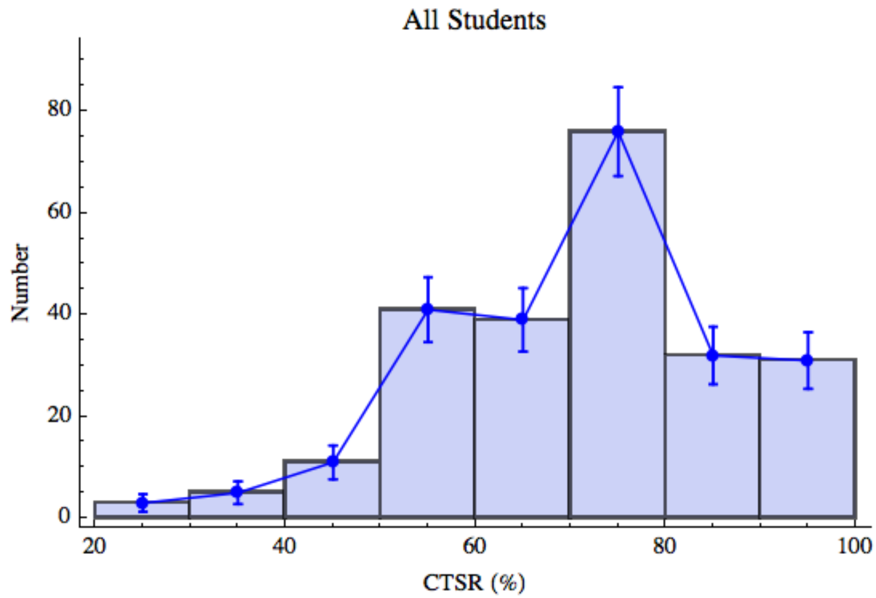


Figure 1

The data are not well modeled as a Gaussian, so the mean is perhaps not the most appropriate way of characterizing it. Therefore, throughout this work although we will give the value the mean since it is standard in the literature, we will also give the value of the median. The result for all students is (mean, median) = $(70.8 \pm 1.0, 70.8 \pm 1.6)$. In this case the two values ended up the same, although below we will see examples where this is not true. The uncertainty in the mean is $\sigma / \sqrt{N}$, where σ is the standard deviation of the sample and N is the number of students in the sample. The uncertainty in the median is calculated as the inter-quartile range divided by the square root of the number of students in the sample, which is similar but simpler than a suggestion by Iglewicz (1983).

The reliability of the CTSR with our students, the Cronbach α coefficient, is 0.75, which is fairly good for an instrument with only 24 questions. The standard uncertainty, which is the uncertainty in each individual student's score on the CTSR, is $\pm 8\%$. Harrison (2014) discusses this sort of calculation in more detail.

The remainder of this section discusses CTSR results for the various factors about the student and their background and interest that we examined.

Gender

Almost exactly 2/3 of our students were female and 1/3 male. For the purposes of this study the 2 students who reported that neither "female" nor "male" were appropriate for them are outliers, and are ignored in our analysis. The range of scores on the CTSR for both female and male students was 25 – 100%. Table I shows the median and mean values on the CTSR. The means differ by 5.9 ± 2.1 and the medians by 4.2 ± 2.5.

Table I

Gender	**(mean, median)**
Female	$(69.1 \pm 1.3, 70.8 \pm 1.7)$
Male	$(75.0 \pm 1.7, 75.0 \pm 1.9)$

Figure 2 shows the scores. Note that the data are normalised by the total number of students in each sample. The figure seems to be consistent with the values of Table I.

Compared to the female students there were more high-performing males and fewer low-performing ones, although there were many high-performing and low-performing students for both genders.

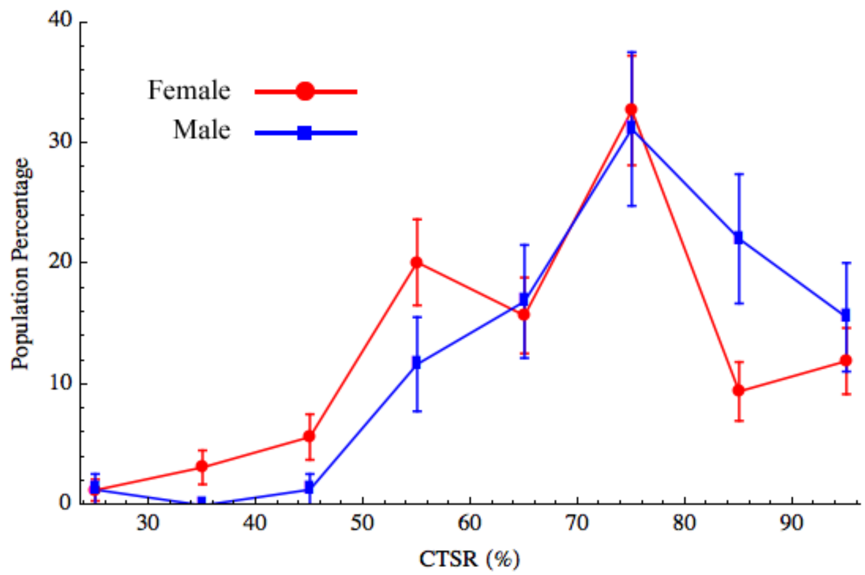


Figure 2

Another way to visually compare two or more datasets is the boxplot. Figure 3 shows the boxplots for the female and male students. The “waist” on the box plot 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 of 1.5 times the inter-quartile range extending from the upper and lower quartiles, as suggested by Emerson and Strenio (1983). The dots represent data points that are considered to be “outliers.” Also shown in the figure are the statistical uncertainties in the value of the medians. The male students show an overall small upward shift in values compared to the female ones.

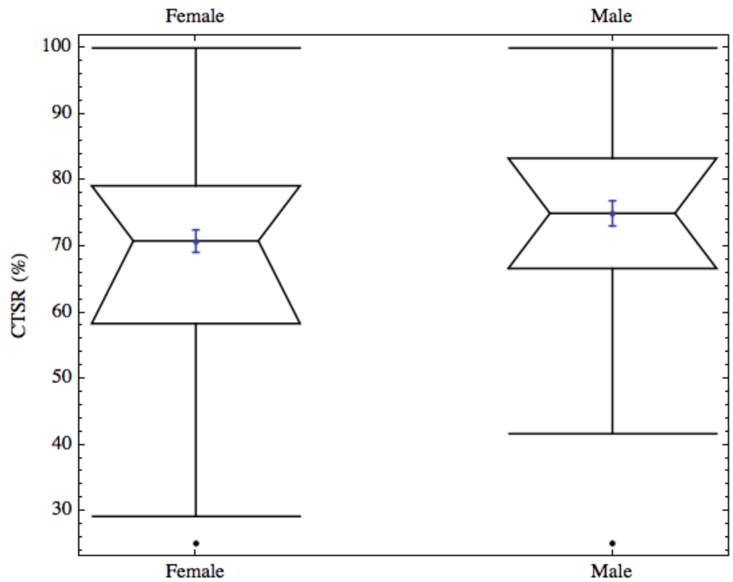


Figure 3

Student's T-Test is well known for testing whether or not two distributions are the same. (For example see Pugh and Winslow, 1966). It typically returns the probability that the two distributions are statistically the same, the *p-value*, which is sometimes referred to just as *p*. By convention, if *p* is < 0.05 then the two distributions are considered to be different. However, the test assumes that the two distributions are both Gaussian, which is not really the case for data of Figure 2. An alternative is the Mann-Whitney U-Test, also known as the Mann-Whitney-Wilcoxon test, which is based on the median, not the mean (Mann and Whitney, 1947). It returns a p-value, which is interpreted identically to the p-value of Student's T-Test. However, the Mann-Whitney test assumes that the distributions have the same shape, which is not really the case for our data. We are not aware of better alternatives to these ways of calculating p-values for our data, although neither are perfect. Table II shows the p-values calculated using both methods. For both calculations $p \ll 0.05$, indicating that the distributions are different.

Table II

Method	***p***
T-Test	0.0062
Mann-Whitney	0.0054

Although reporting p-values is fairly common in Physics Education Research, the p-value by itself doesn't address a crucial question, which is how large the difference is between the two samples. The *effect size* is a measure of the size of the difference. A discussion of effect sizes in the context of medical research is Sullivan and Feinn (2012). For distributions that are Gaussian, a common effect size parameter is the Cohen *d* (Cohen, 1992). For non-normal distributions like our CTSR data, Cliff's δ provides a somewhat similar measure (Cliff, 1993). The Cliff δ for 2 samples is the probability that a value randomly selected from the first group is greater than a randomly selected value from the second group minus the probability that a randomly selected value from the first group is less than a randomly selected value from the second group. It is calculated as:

$$\delta = \frac{\#(x_1 > x_2) - \#(x_1 < x_2)}{N_1 N_2} \tag{4}$$

where # indicates counting. The values of δ can range from -1, when all the values of the first sample are less than the values of the second, to +1, where all the values of the first sample are greater than the values of the second. A value of 0 indicates samples whose distributions completely overlap. By convention, $|\delta| < 0.147$ is a "negligible" difference between the two samples, $|\delta| < 0.33$ indicates a "small" difference, $|\delta| < 0.474$ a "medium" difference, and otherwise a "large" one (Romano, Kromrey, Coraggio & Showronik, 2006). For our gender data $\delta = 0.22$, indicating a small difference. The 95% confidence interval range of δ is 0.069 – 0.366; since this range does not include 0, the difference is statistically significant.

Another way of looking at the data is to examine the percentage of the students who missed 5 or more of the 24 questions on the test. Table III shows the result. The stated uncertainties are from assuming that the uncertainty in the number of students *N* is $\sqrt{N}$. The values in the table differ by 16 ± 11 .

Table III

Gender	**Percent Missing 5 or More Questions**
Female	78.6 ± 7.0
Male	62.3 ± 9.0

Senior-Level High School Physics

In Ontario, the senior-level high school physics course is called *Grade 12 Physics*. For our course it is recommended but not required. 64% of our students took

Grade 12 Physics. The range of scores on the CTSR was 25 – 100% for students with Grade 12 Physics, and 25 – 95.8% for students without. Table IV shows the median and mean values on the CTSR. The mean values differ by 4.5 ± 1.6 while the median values are equal.

Table IV

Grade 12 Physics?	**(mean, median)**
Yes	$(72.5 \pm 1.3, 70.8 \pm 1.7)$
No	$(68.0 \pm 1.0, 70.8 \pm 2.3)$

The suggestion from the value of the means that the two groups are different seems to be confirmed by the plots in Figure 4. For students with high school physics there were more high-performing ones and somewhat fewer low-performing ones compared to the students without high school physics.

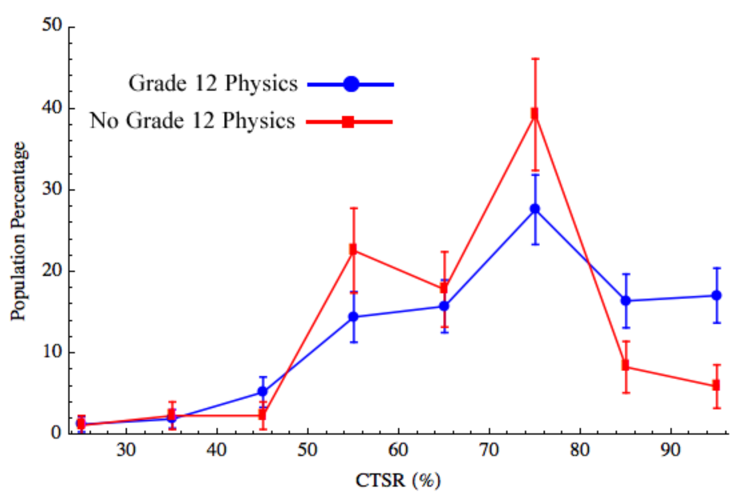


Figure 4

Figure 5 shows the boxplots.

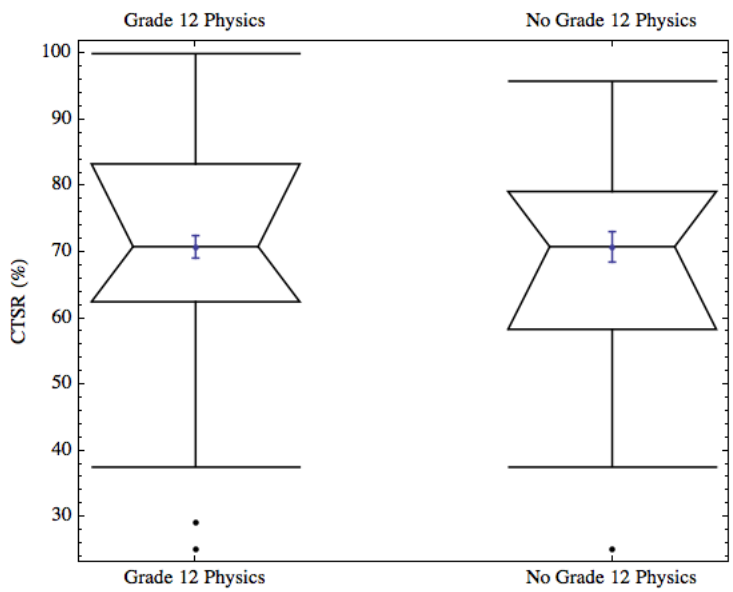


Figure 5.

The scores for students with Grade 12 Physics show an asymmetry, a *quartile skewness*, since the median is closer to the lower quartile than the higher one; the median for students without Grade 12 Physics is closer to the upper quartile than the lower one. Conventionally the quartile skewness *qs* is defined as:

$$qs \equiv \frac{(Q3 - M) - (M - Q1)}{Q3 - Q1} \quad (1)$$

where *Q3* is the upper quartile, *M* is the median, and *Q1* is the lower quartile. For the students with Grade 12 Physics *qs* = 0.2, while for students without Grade 12 Physics *qs* = -0.2. This explains why the medians of these two groups are the same but the means are different. By comparison, for the data for gender in the previous sub-section, female students had *qs* = -0.2, while for male students *qs* = 0.0.

The differences for students with and without senior-level high school physics are also weakly confirmed by the p-values shown in Table V. However, for the T-Test our software, *Mathematica*, complained about the non-Gaussian nature of the data.

Table V.

Method	***p***
T-Test	0.034
Mann-Whitney	0.027

Cliff's δ tells a similar story: $\delta = 0.17$ with a 95% confidence interval range of 0.023 – 0.314. Therefore the difference is "small" but statistically significant.

The percentage of students who missed 5 or more of the 24 questions on the test was also different, as seen in Table VI. The values in the table differ by $(19 \pm 12)\%$.

Table VI

Grade 12 Physics	**Percent Missing 5 or More Questions**
Yes	66.4 ± 6.6
No	85.7 ± 10.1

Other Factors

There are five other student characteristics for which we examined correlations with CTSR performance. None of them showed statistically significant differences using any of the ways of analyzing the data that we could devise.

For the 4 factors with more than two possibilities, we used a Kruskal-Wallis one-way analysis of variance (Kruskal & Wallace 1952). This is an extension of Mann-Whitney U-Test, can deal with more than two samples, but assumes that the distributions have the same shape and differ only in the value of the medians. The Kruskal-Wallis also returns a p-value, which is interpreted identically to p-values returned by the T-Test and Mann-Whitney U-Test. As with the T-Test and Mann-Whitney ones used earlier, Kruskal-Wallis is not completely appropriate for our data although we are unaware of better alternatives. Table VII shows some of the p-values for these other factors.

Table VII

Factor	**Test**	***p***
Program of Study	Kruskal-Wallis	0.26
Reason for Taking the Course	Kruskal-Wallis	0.53
When Graduated from High School	Kruskal-Wallis	0.26
When Took the Calculus Co-Requisite	Kruskal-Wallis	0.64
Previously Dropped this Course	Mann-Whitney	0.61

The Cliff δ can only be calculated for the factor of whether or not the student had previously started but dropped the course, since it is the only one with only two groups. The result is $\delta = 0.053$ with a 95% confidence interval of -0.14 – 0.24. Therefore, the difference is "negligible" and not statistically significant.

Discussion

The Force Concept Inventory (FCI) tests students' conceptual understanding of classical mechanics. The FCI was introduced by Hestenes, Wells and Swackhammer (1992) and was updated in 1995 (available from http://modeling.asu.edu/R&E/Research.html). It has 30 questions. Performance on the FCI is taken to be a measure of the quality of previous instruction. In the context of a single course, the FCI is often given at the beginning of the term, the Pre-Course, and again at the end of the term, the Post-Course. The standard way of measuring student gains on the FCI is from a seminal paper by Hake (1998). It is defined as the gain divided by the maximum possible gain, often called the normalized gain *G*:

$$G = \frac{(\mathrm{PostCourse}\% - \mathrm{PreCourse}\%)}{(100 - \mathrm{PreCourse}\%)} \qquad (2)$$

Clearly, *G* cannot be calculated for students whose PreCourse% score was 100.

To assess the quality of instruction in a course, Hake (1998) also defined the average normalised gain. It is:

$$< g > = \frac{< \text{PostCourse\%} > - < \text{PreCourse\%} >}{100 - < \text{PreCourse\%} >} \tag{3}$$

where the angle brackets indicate means. The FCI is probably the single most-used diagnostic instrument in Physics Education Research (PER).

Bao et al. collected data on Pre-Course scores on the FCI, the Brief Assessment of Electricity and Magnetism (BEMA), and the CTSR by students at four U.S. and three Chinese universities (Bao, Cai, Fang, Han, Wang, Liu, … & Wu, 2009). All the universities were chosen to be of medium ranking. On the FCI and BEMA Chinese students outperformed their U.S. counterparts by a large margin, while on the CTSR performance was essentially identical. This indicates that the instruction that leads to good FCI outcomes is not necessarily correlated with giving the students an increased ability to reason scientifically.

We did not give the FCI to students in this course, but did the previous year (Harlow, Harrison & Honig, 2015) and believe that those results would have been comparable for this course. Figure 6 shows the Bao data for the FCI, taken from a scan of the figure in the Science article, plus the data for last year's University of Toronto summer course. Evidently China is doing better than North America in preparing their students with a good conceptual understanding of classical mechanics. This data should not be used to compare Canadian and U.S. preparation, in part because the demographics of students taking our summer course are different from the demographics of students taking the fall version of the course. Presumably the fall version is a closer match to the U.S. students in this data. The mean FCI scores for our students was 43% while for the students in the fall course it was 54%, which was higher than the 49% for U.S. students.

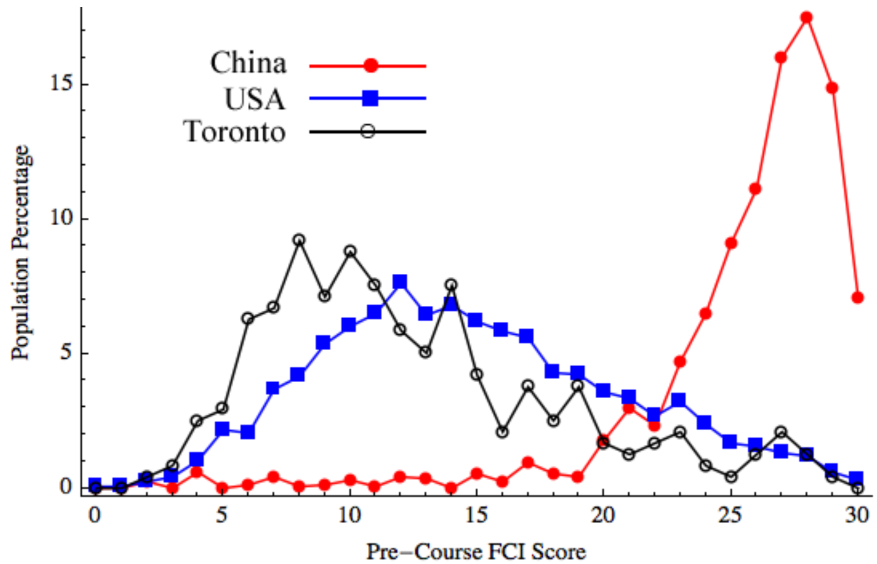


Figure 6

Figure 7 shows the Bao data for the CTSR plus the data for our course. All three curves are almost identical. Evidently the teaching that did well for FCI outcomes is not strongly related to giving the students the ability to reason in a scientific way.

The remainder of this section discusses the results on the CTSR for the factors that we examined in the Results section above.

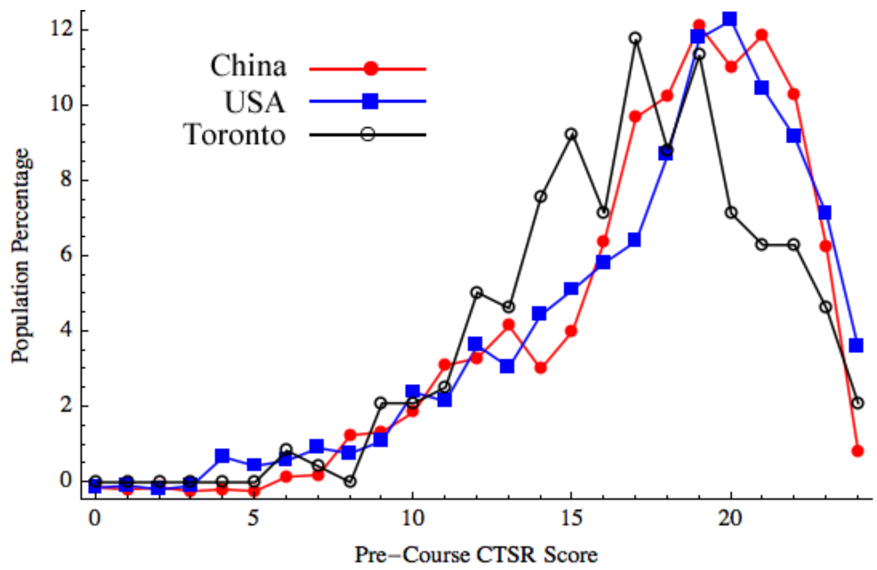


Figure 7

Gender

In previous years in the course that is the subject of this study, we have asked the students about themselves and their background. The questions were identical to Questions 24 – 30 of Appendix B, except the year of graduation from high school (Question 27) was different for different years that the information was collected. How these factors correlated with performance in the course has been investigated (Harlow, Harrison & Meyertholen, 2014). This data has also been used in a study comparing the effectiveness of a compressed-format summer version of the course to the regular-format version in the fall term (Harlow, Harrison & Honig, 2015). Question 31 on gender was not asked in these studies, in part because the issue of gender and physics performance is an important and much-studied topic, but which was not the focus of the work being reported on in those papers.

However, when we took our data we were not aware of any studies of correlations between gender and performance on the CTSR. There have been many studies of gender and Formal Operations using other probes of cognitive level (for example Overton & Meehan, 1982, and references cited there). The results of those studies are ambiguous: some studies find differences and some do not. Therefore for completeness we added the question on gender in this study, although we didn't expect to see any statistically significant differences, in part because the course is relatively small; only 238 students wrote the CTSR.

Our expectation does not seem to be correct. Although the differences in CTSR performance by gender are fairly small, they appear to be statistically significant. We are aware of the controversial nature of this conclusion, and are certainly willing to share the spreadsheet of our raw data, but with student identification information such as name etc. redacted.

Since taking our data we have discovered that Coletta (2015) has also found a strong correlation between gender and CTSR performance, and also for gender and normalized gains on the FCI. This work is consistent with our results that males tend to outperform females on the CTSR.

Of course, correlation is not causation, and in the next sub-section on CTSR performance and its correlation with whether or not the student took a senior-level high school physics course, we discuss data that indicate that for that factor causation is probably not in play. For the factor of gender, certainly cultural influences are at work.

Bem (1974) developed a Sex Role Inventory in which people self-reported on socially desirable stereotypically masculine and feminine personality characteristics. Both males and females associated being analytical with masculinity. Of course, our society has been undergoing large changes in its attitudes towards gender. However a 1998 reinvestigation of attitudes towards being analytical showed that it was still considered to be masculine both by males and females (Hold & Ellis, 1998). Particularly for students in adolescence and early adulthood, identifying oneself by gender is particularly important. Therefore, as Bem (1981) wrote earlier, “cultural myths become self-fulfilling prophecies.” Steele and Arsonson (1995) introduced the phrase *stereotype threat* to describe these cultural myths in a different context than gender.

Some relevant information is provided by a recent study by Leslie, Cimpian, Meyer, & Freeland (2015). They found that in U.S. post-secondary institutions, different disciplines are perceived as requiring different levels of raw intellectual talent. Those perceptions are negatively correlated with the percentage of female PhD students in the disciplines: the greater the perception of required raw talent, the fewer females in the discipline. This was found to be true not only in the STEM fields of science, technology, engineering and mathematics, but also in the social sciences and humanities. A similar correlation was found in the percentage of African-American PhD students, but not Asian-American PhD students. Although there is no data on whether or not the perception that some disciplines require more raw talent than others is actually correct, the authors argue that in either case stereotype threat is a factor in participation rates.

Senior-Level High School Physics

Coletta & Phillips (2005) studied the correlation of CTSR performance with the average normalized gain *G* (not <*g*>) on the FCI and found a positive correlation for students at Loyola Marymount University, but in an indirect argument propose that there is no such correlation for students at Harvard because a higher fraction of Harvard students are formal operational thinkers. Coletta, Phillips, & Steinert (2007) added data on a positive correlation for students at Edward Little High School, Diff & Tache (2007) found a positive correlation for students at Santa Fe Community College, and Nieminen, Savinainen & Viiri (2012) found a positive correlation for high school students in Finland.

Harlow, Harrison & Meyertholen (2014) found a correlation in student performance in the Fall version of this course, measured both by the average normalised gain on the FCI and by performance on the Final Examination, and whether or not the student took a senior-level high school physics courses. In that paper it was cautioned that ascribing a causal relationship to the correlation was dangerous and probably wrong. Similarly, especially in light of Bao et al.'s work (2009) we think it is highly implausible that taking a senior-level high school physics course *caused* an increase in student ability to think scientifically. Rather, it seems more likely that students who can't or believe that they can't think "that way" tend to avoid high school physics.

A final example of Piagetian taxonomy may be useful. Here are two math problems.

Problem C

$$x = y + 3$$
$$x + y = 17$$

Solve for x and y.

Problem F

> Xavier is three years older than Yolanda. The sum of Xavier and Yolanda's ages is 17. How old are Xavier and Yolanda?

The manipulations to solve Problem C, little more than pushing symbols around on a piece of paper with a pencil, require only Concrete Operations. However, casting Problem F into the form of Problem C requires the type of abstraction that is a characteristic of Formal Operations. Of course, many if not most physics problems involve the same type of abstract thinking when casting a physical situation into a set of equations.

Some years ago, with A.W. Key, we did some investigations and interventions with students in serious difficulty in their physics course. The course was the equivalent of the one that is the subject of this study. We were guided in this work by Piagetian taxonomy, although we were unaware of the CTSR; perhaps when we did it the CTSR was not yet available. Although we didn't manage to help these students very much if at all, we did hear the same story from them, in one form or another, many times.

> When the young person was about 13 or 14 years old they were considered to be good student. In their algebra class they confronted "word problems" such as Problem F above for the first time. The student couldn't make any sense of how to do the problem, which would be expected for a student not yet capable of Formal Operations. After the teacher repeatedly urges the student to just try harder, the

> teacher finally loses patience and tells the student they are too stupid for this kind of work! The student believes the teacher, and is now traumatised.

Later, even if in principle the young person has become capable of Formal Operations, their trauma makes them incapable of this way of thinking, at least in situations that trigger their trauma. Thus at least in some circumstances their belief in what their teacher told them has become a self-fulfilling prophecy. On the other hand, a 13 or 14 year old who was capable of Formal Operations would have little difficulty with word problems in algebra class, and might later choose to do high-school physics because they are confident in their ability to do well.

Of course, the trauma due to an incompetent teacher is an extreme case. Just the fact that physics is considered to be hard by a large fraction of the general population could be sufficient to lead many students to believe that this sort of thing is beyond them.

Other Factors

One reason why we studied the summer version of this course instead of using the better statistics from the much larger 1200-student fall term version is that in the fall course almost 80% of the students were fresh out of high school (Harlow, Harrison & Meyertholen, 2014), while in the summer version the students had graduated from high school longer ago. Since Formal Operational ability is correlated with age, and when a student graduated from high school is roughly a measure of their age, we chose to investigate CTSR performance in the summer course to see if we could see a correlation of CTSR performance with age. We did not see a statistically significant correlation. Perhaps for the age of our students, typically greater than or equal to 18 years, age is not a factor in Formal Operations.

Future Work: Can We Aid Stage Promotion?

All teachers need to be sensitive to the cognitive abilities of their students in the way we structure our courses, in our interactions with our students, and in the things we ask the students to do on test questions. However, the important issue is: can we organize our courses to aid students in becoming Formal Operational, i.e. in learning to "think like a physicist"? There are some studies that indicate that the answer is yes.

Lawson, Aklhoury Benford, Clark & Falconer (2000) demonstrated a normalised gain on the CTSR in a biology course for non-science majors ($p < 0.001$). Traditional courses begin with the theoretical concepts and then progress to more descriptive and hypothetical concepts. Lawson's course reversed the order: they start with the descriptive contents, progress to hypothetical concepts, and then finally to theoretical concepts.

In the United Kingdom a program called Cognitive Acceleration in Science Education (CASE) has had considerable success in stage promotion with students between ages 11 – 14 years (Adey, 1999). CASE rests on five pillars:

1. Cognitive conflict. This occurs when a student encounters a problem that forces them to confront their misconceptions. Structured help from a teacher or particularly through interactions with other students helps the student gain at least an understanding of the source of the conflict.
2. Construction. The student must actively construct new ways of thinking.
3. Metacognition. The student is encouraged to think about his or her own thinking.

4. Concrete preparation. Just giving a student a cognitively challenging task is not enough. First there must be a phase of preparation in which the language and any apparatus to be used are introduced.
5. Bridging. The ways of thinking developed in a particular context must be linked to other contexts in science and experiences in real life.

A video of CASE in action that nicely demonstrates how it is implemented is available at: http://archive.teachfind.com/ttv/www.teachers.tv/videos/cognitive-acceleration.html.

Physics Education Research has led many courses to reform their pedagogy. One of the pioneers of this reform is Lillian McDermott, who has "raised putting students into a state of cognitive dissonance into an art form" (Taylor, 2006). This is, of course, the first pillar of CASE. Her tutorials in introductory physics (McDermott, Shafer & the Physics Education Group, 2002) have been widely implemented and adapted to different educational contexts. McDermott's tutorials are often used conjunction with instruction in large lecture halls. A variation, called variously *Studio Physics*, *SCALE-UP*, *TEAL*, etc., replaces the lecture halls entirely with guided-discovery pedagogy with students working in small groups.

Coletta (2015) describes in detail courses that, in addition to PER-based Interactive Engagement pedagogy, have introduced reforms to explicitly aid students in learning to think scientifically. At least in the context of these courses, their results are spectacular. Not only did they increase the performance of their students on the CTSR, they also increased the normalised gains on the FCI and greatly reduced the gender gap. CASE was one of the major sources of their reforms.

A key result of PER is that students learn best by interacting with their peers: They do not learn best by being lectured to. There is an aspect to these peer interactions that may resonate with pillar 2 of CASE, that the students must construct new conceptual frameworks. This construction is inherently difficult and often actually frightening for the student. This is not surprising, since we are asking the students to take down previous ways of thinking which are based on a lifetime of experience. Often we see students actively resisting this process. However, we think that if the process is centered on student-student interactions, it is probably much less threatening for the students than if an instructor is involved in more than a Socratic role.

It is also possible that the nature of these peer interactions are an ideal way to address the small but troubling differences in scientific reasoning due to cultural preconceptions or math-trauma that we have seen in our data and in our own classrooms.

In any case it seems that reformed physics pedagogy is already partially implementing some of the pillars advocated by CASE. Explicitly adding the other pillars to our instruction and assessing the results using a Pre-Course/Post-Course method could be very interesting. However, the CTSR is perhaps not an ideal instrument for this. One problem is that the mean scores are fairly high, in the low 70% range. Using a normalised gain instead of just the gain helps alleviate this somewhat, but a test instrument with lower mean scores, using more questions of the type shown in Appendix A, could be better. Also some of the questions on the CTSR are testing for Concrete Operations that virtually all of our students get these correct, and some of the questions on the CTSR take the student longer to read than is ideal. In addition, adding some probes of whether the student is capable of thinking from different points of view or frames of reference could be useful.

Conclusions

An acausal correlation between 2 factors usually indicates a third “hidden variable” with a causal relationship to both. We have argued that a student’s beliefs about the types of cognition that they are capable of or feel are appropriate for them become self-fulfilling prophecies which impacts their ability to think scientifically. We have examined gender and the belief that Physics is hard and beyond the student’s reach, although there are certainly other factors such as socio-economic background and family that have similar subtle but real influences.

Physics Education Research applied to introductory courses has for some years been using assessment instruments such as the FCI and the BEMA in a Pre-Course/Post-Course protocol. This work has been vital in discovering what pedagogy works in building conceptual understanding in our students. Although still important tools for instructors to assess their own success with their students, we may have learned about all that we can about general principles of effective pedagogy from them. Therefore we second and somewhat extend the advice of Coletta, Phillips, and Steinhart (2007) : we should be concentrating more on testing the ability of our students to think scientifically using instruments such as the CTSR, or perhaps a modified version, in a Pre-Course/Post-Course protocol.

References

Adey, P. (1999). The Science *Of* Thinking, and Science *For* Thinking: A Description of Cognitive Acceleration Through Science Education. Innodata Monographs – 2, The International Bureau of Education, UNESCO.

Arnett, J.J. (2010). *Adolescence and Emerging Adulthood: A Cultural Approach* (3rd ed.). New York, NY: Pearson.

Bao, L., Cai, T., Fang, K.K.K., Han, J., Want, J., Liu, Q., … & Wu, N. (2009). Learning and Scientific Reasoning. *Science* 323, 586. The paper and some supplementary materials are available at: http://www.physics.ohio-state.edu/%7Elbao/Papers/Bao_Learning-Scientific-Reasoning.pdf (Retrieved June 15, 2014).

Bem, S. (1974). A measure of androgyny and gender role. *Journal of Consulting and Clinical Psychology* 42(2), 155.

Bem, S. (1981). Gender Schema Theory: A Cognitive Account of Sex Typing. *Psychological Review* 88(4), 354.

Cliff, N. (1993). Ordinal analysis to answer ordinal questions. *Psychological Bulletin* 114, 494.

Cohen, J. (1992). A Power Primer. *Psychological Bulletin* 112, 155.

Coletta, V.P. (2015). *Thinking in Physics*. San Francisco, CA: Pearson.

Coletta, V.P., & Phillips, J.A. (2005). Interpreting FCI scores: Normalized gain, preinstruction scores, and scientific reasoning ability. *The American Journal of Physics* 13, 1172.

Coletta, V.P., Phillips, J.A., & Steinert, J.J. (2007). Why You Should Measure Your Students' Reasoning Ability. *The Physics Teacher* 45, 235.

Diff, K., & Tache, N. (2007). From FCI to CSEM to Lawson Test: A Report on Data Collected at a Community College. *2007 Physics Education Research Conference Proceedings* 951, 85. Available at: http://www.compadre.org/portal/items/detail.cfm?ID=9054&Relations=1

Emerson, J.D., & Strenio, J. (1983). Boxplots and Batch Comparison. In D.C. Hoaglin, F. Mosteller, and J.W. Tukey (Eds.), *Understanding Robust and Exploratory Data Analysis* (pg. 424). Toronto, ON: Wiley-Interscience.

Hake, R.R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *The American Journal of Physics* 66, 64.

Harlow, J.J.B., Harrison, D.M. & Honig, E. (2015). Compressed-format compared to regular-format in a first year university physics course. *The American Journal of Physics* 83(3), 2015.

Harlow, J.J.B., Harrison, D.M. & Meyertholen, A. (2014). Correlating student interest and high school preparation with learning and performance in an introductory university physics course. *Physical Review Special Topics: Physics Education Research* 10, 010112.

Harrison, D.M. (2014). The uncertainty of grades in physics courses is surprisingly large. *Physics in Canada* 70(2), 75. Available at: http://www.upscale.utoronto.ca/PVB/Harrison/UncertaintyOfGrades/UncertaintyOfGrades.pdf

Hestenes, D., Wells, M. & Swackhammer, G. (1992). Force Concept Inventory. *The Physics Teacher* 30, 141.

Hold, C.L., & Ellis, J.B. (1998). Assessing the Current Validity of the Bem Sex-Role Inventory. *Sex Roles* 39(11/12), 929.

Inhelder, B., & Piaget, J. (1958). *The Growth of Logical Thinking From Childhood to Adolescence; An Essay On The Construction of Formal Operational Structures*. New York, NY: Basic Books.

Iglewicz, B. (1983). Robust Scale Estimators and Confidence Intervals for Location. In D.C. Hoaglin, F. Mosteller, and J.W. Tukey (Eds.), *Understanding Robust and Exploratory Data Analysis* (pg. 424). Toronto, ON: Wiley-Interscience.

Kruskal, W.H., & Wallis, W.A. (1952). Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association* 47. 583.

Lawson, A.E. (1978). The development and validation of a classroom test of formal reasoning. *Journal of Research in Science Teaching* 15, 11. Available from https://modelinginstruction.org/wp-content/uploads/2013/06/LawsonTest_4-2006.pdf

Lawson, A.E., Aklhoury, S., Benford, B., Clark, B.R. & Falconer, K.A. (2000). What Kinds of Scientific Concepts Exist? Concept Construction and Intellectual Development in College Biology. *Journal of Research in Science Teaching* 37(9), 996.

Leslie, S.J., Cimpian, A., Meyer, M. & Freeland, E. (2015). Expectations of brilliance underlie gender distributions across academic disciplines. *Science* 347(6219), 262.

Mann, H.B., & Whitney, D.R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Annals of Mathematical Statistics* 18, 50.

McDermott, L.C., Shaffer, P.S. & the Physics Education Group, Dept. of Physics, Univ. of Washington (2002). *Tutorials in Introductory Physics*. Upper Saddle River, NJ: Prentice-Hall.

Nieminen, P., Savinainen, A., & Viiri, J. (2012). Relations between representational consistency, conceptual understanding of the force concept, and scientific reasoning. *Physical Review Special Topics: Physics Education Research* 8, 101223.

Overton, W.F., & Meehan, A.M. (1982). Individual Differences in Formal Operational Thought: Sex Role and Learned Helplessness. *Child Development* 53(6), 1536.

Pratt, C., & Hacker, R.G. (1984). Is Lawson's Classroom Test of Formal Reasoning Valid? Educational and Psychological Measurement 44, 441.

Pugh, E.M., & Winslow, G.H. (1966). *The Analysis of Physical Measurements* (pg. 172 ff.). Don Mills, ON: Addison-Wesley.

Romano, J., Kromrey, J.D., Coraggio, J. & Showronik, J. (2006). Exploring methods for evaluating group differences on the NSSE and other surveys: Are the t-test and Cohen's d indices the most appropriate choices? Presented at the annual meeting of the Southern Association for Institutional Research, October, 14- 17, 2006, Arlington, Virginia. Available at: http://www.coedu.usf.edu/main/departments/me/documents/MethodsforEvaluatingGroup.pdf

Steele, C.M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology* 69, 797.

Sullivan, G.M., & Feinn, R. (2012). Using Effect Size – or Why the *P* Value is Not Enough. *Journal of Graduate Medical Education* 4, 279.

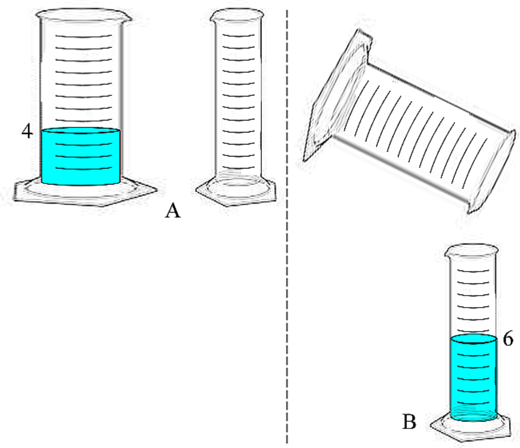
Taylor, E.F. (2006). A remark at the American Association of Physics Teachers Topical Workshop on Teaching General Relativity to Undergraduates. Syracuse, NY.

Appendix A

Two Questions from the CTSR

QUESTION 5

To the right are drawings of a wide and a narrow cylinder. The cylinders have equally spaced marks on them. Water is poured into the wide cylinder up to the 4th mark (see A). This water rises to the 6th mark when poured into the narrow cylinder (see B).



Both cylinders are emptied (not shown) and water is poured into the wide cylinder up to the 6th mark. *How high would this water rise if it were poured into the empty narrow cylinder?*

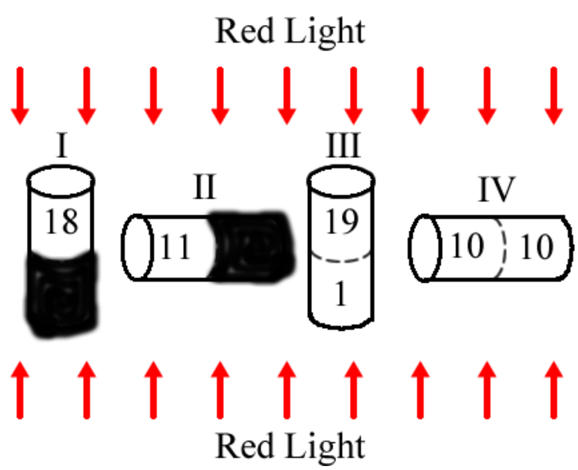
A. to about 8
B. to about 9
C. to about 10
D. to about 12
E. none of these answers is correct

Results

The correct answer is B. 59% of the class got this correct, i.e. 41% of them missed it. This indicates that a significant fraction of the class is having difficulty with ratios.

QUESTION 11

Twenty fruit flies are placed in each of four glass tubes. The tubes are sealed. Tubes I and II are partially covered with black paper; Tubes III and IV are not covered. The tubes are placed as shown. Then they are exposed to red light for five minutes. The number of flies in the uncovered part of each tube is shown in the drawing.



This experiment shows that flies respond to **(respond means move to or away from):**

A. red light but not gravity

B. gravity but not red light

C. both red light and gravity

D. neither red light nor gravity

Results

Answer B. 59% got this correct. This question probes the ability of students to separate and control variables.

Appendix B

Questions About the Student and Their Background

25. “What is your intended or current Program of Study (PoST)?”

Answer	**Percent**
Life Sciences	73
Physical and Mathematical Sciences	8
I haven’t decided yet	3
Other / NA	17

26. “What is the main reason you are taking PHY131?”

Answer	**Percent**
Because it is required	55
For my own interest	10
Both because it is required and for my own interest	22
Other / NA	10

27. "When did you graduate from high school?"

Answer	Percent
2013	39
2012	21
2011	22
2010	22
Other/NA	10

28. "Did you take Grade 12 Physics or an equivalent course elsewhere?"

Answer	Percent
Yes	64
No	36

29. "MAT135 or an equivalent calculus course is a co-requisite for PHY131. When did you take the math course?"

Answer	Percent
I am taking it now	8
Last year	49
Two or more years ago	37
Other / NA	8

30. “Have your previously started but did not finish PHY131?”

Answer	**Percent**
Yes	16
No	84

31. “What is your gender?

Answer	**Percent**
Female	67
Male	32
Neither of these are appropriate for me	1