Relativity concept inventory: Development, analysis, and results

J. S. Aslanides and C. M. Savage[*]

Physics Education Centre, Research School of Physics and Engineering, Australian National University, Canberra ACT 0200, Australia

(Received 28 February 2013; published 9 May 2013)

We report on a concept inventory for special relativity: the development process, data analysis methods, and results from an introductory relativity class. The Relativity Concept Inventory tests understanding of relativistic concepts. An unusual feature is confidence testing for each question. This can provide additional information; for example, high confidence correlated with incorrect answers suggests a misconception. A novel aspect of our data analysis is the use of Monte Carlo simulations to determine the significance of correlations. This approach is particularly useful for small sample sizes, such as ours. Our results show a gender bias that was not present in course assessment, similar to that reported for the Force Concept Inventory.

DOI: 10.1103/PhysRevSTPER.9.010118 PACS numbers: 01.40.Di, 01.40.Fk, 01.40.gf

I. INTRODUCTION

Concept inventories are used to assess learning in many areas of physics education [1]. When used to determine the effectiveness of educational innovations, they may contribute to the teaching development cycle. Since the literature on special relativity education research does not include a concept inventory, we have developed the Relativity Concept Inventory (RCI), available from the Supplemental Material [2].

Special relativity is interesting in a physics education research context because of its combination of deeply challenging concepts and simple mathematics. This is in contrast with quantum mechanics, which has a more complex mathematical structure. Nevertheless, the amount of physics education research on special relativity is small [3–16].

The relativity of motion is a central concept in both Galilean and special relativity. Both theories use inertial reference frames to describe motion. In the context of Galilean relativity student difficulties with inertial reference frames have been documented, in particular, a tendency to distinguish between "real" and "apparent" motion [10,11]. This observation has been reproduced in the special relativity context [8]. Another misconception reported in the literature is that special relativistic phenomena, such as time dilation and length contraction, are merely optical illusions [6,7].

Scherr *et al.* [12,13,15] have conducted in-depth studies of students' understanding of special relativity, and of the relativity of simultaneity, in particular. They found that students commonly attribute the relativity of simultaneity to signal travel time. One RCI question (question 21) addresses this misconception directly. However, the relativity of simultaneity involves such a complex set of concepts that we also phrase some questions to specifically discourage this common misconception (questions 11 and 12).

In our data analysis we have paid particular attention to assessing the statistical significance of our results. To this end we developed and applied Monte Carlo simulation techniques suitable for the analysis of the statistical significance of correlations in data with small sample size.

In the next section we describe the process used to develop and validate the RCI. In Sec. III we characterize the students the RCI was administered to. In Sec. IV we describe the methods used to analyze the collected data, including the use of item response theory to control for the effect of student ability on correlations between questions, and Monte Carlo modeling. In Sec. V we present misconceptions diagnosed by the RCI and evidence for its gender bias. Finally, in Sec. VI, we suggest revisions of the RCI. We also argue that understanding the gender bias in concept inventories is a problem that should be addressed by physics education research.

II. DEVELOPMENT AND VALIDATION

Our primary purpose in developing the RCI is to provide an instrument for measuring changes in students' conceptual understanding of special relativity. In this role it would be administered prior to instruction as a pre-test, and after instruction as a post-test. The change in students' understanding is quantified by the normalized gain, described at the end of Sec. IV A. A second role for the RCI, especially in conjunction with the confidence testing, is to identify students' misconceptions. This is discussed in Sec. V A.

The development of the RCI followed Adams and Wieman [17] insofar as our six month project schedule

[*] craig.savage@anu.edu.au

Published by the American Physical Society under the terms of the Creative Commons Attribution 3.0 License. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

allowed. In particular, we conducted fewer student interviews than suggested by them. The only previous attempt to develop a concept inventory for special relativity is reported by Gibson [18].

The RCI has been validated by feedback from discipline experts, by detailed analysis of individual students' responses, and by standard statistical methods [17,19]. Students' RCI responses were also benchmarked against assessment items such as homework and an exam.

We first formulated a list of concepts that captured the learning goals of the introductory relativity instruction in the Physics 2 course at The Australian National University (ANU), described in the next section. These concepts were also informed by relevant textbooks [20] and the physics education research literature [3–16].

The RCI validation was an iterative process of constructing questions that are demonstrably measuring students' understanding of important concepts in special relativity. There are two components to this: the experts' judgement and the students' interpretation. A group of experts in relativity determined that the RCI is asking the right questions, and students' responses to the RCI were analyzed to ensure that they were interpreting the questions as intended.

The process started with 14 draft concepts of introductory relativity on which expert feedback was obtained from 30 international respondents using an online survey [21]. Agreement with the appropriateness of the concepts in our list ranged from 100% to 50%. After individual consideration, concepts with agreement below 75% were removed from the list. The final list of nine concepts is given in Table I.

These concepts were used to develop 24 draft RCI multiple-choice questions, with up to four questions addressing each of the concepts. Where possible, distractors were chosen to represent common misconceptions documented in the literature. Expert feedback on the draft RCI questions was obtained from seven respondents using another online survey. In addition, a face-to-face interview was conducted with the ANU academic teaching advanced special relativity.

Following the expert validation, the draft RCI was administered to six fourth-year physics students. These students were also asked to write a sentence or two explaining their reasoning for each question. Next, the RCI was taken by three second-year students in think aloud format: students were asked to verbalize their thinking while answering the RCI questions. These students had taken the Physics 2 class the previous year. These sessions were recorded and transcribed for study.

The RCI was then administered online to the 2012 ANU Physics 2 class, prior to instruction as a pre-test, and after instruction as a post-test. Neither contributed to the course assessment. Although these administrations were part of the validation process, the data produced research results that are discussed in Sec. V. Because of the compressed project time line, the development and analysis phases of the project overlapped.

Students' RCI post-test responses were compared to their answers to the relativity questions in the Physics 2 midcourse exam, which included short answer conceptual questions. This enabled individual students' written reasoning to be compared with their RCI responses. Among the quantitative measures that demonstrate validity is the Pearson product-moment correlation coefficient [discussed in Sec. IVA and defined by Eq. (2)] $r = 0.39$ between the RCI normalized gains (discussed in Sec. IVA) and the relativity exam question marks.

TABLE I. The concepts tested by the RCI. In the questions column are the question numbers we classified as associated with each concept. Although some questions clearly test more than one concept, we have allocated each question to only one concept.

Concept	Description	Questions
First postulate	The laws of physics are the same in all inertial reference frames.	16, 18, 19, 20
Second postulate	The speed of light in a vacuum is the same in all reference frames.	3, 4
Time dilation	The time interval between two timelike separated events is shortest in the reference frame for which the two events are at the same position. The time between these events is greater in all other frames.	5, 6, 7, 8
Length contraction	The length of an object (defined as the space interval between two simultaneous events at either end of the object) is the longest in the frame in which the ends of the object are at rest, and is shorter in all other frames.	13, 14, 17
Relativity of simultaneity	If two events A and B are spacelike separated, then there exist inertial frames in which A precedes B, and others in which B precedes A.	11, 12, 15, 21
Inertial reference frame	A coordinate system in which a free particle will maintain constant velocity; in particular, the concept that all inertial frames are equivalent.	1, 2
Velocity addition	Velocities transform between frames such that no object can be observed traveling faster than the speed of light in a vacuum.	9, 10
Causality	If two events are timelike separated, then the ordering of the events is fixed for all reference frames.	22, 23
Mass-energy equivalence	Energy has inertia.	24

All this feedback was used to continuously improve the draft RCI. Wording was clarified when found to be ambiguous and questions were deleted and replaced when it was determined they were not adequately addressing desired concepts. In particular, questions 18, 19, and 21 were substantially changed between the pre-test and post-test. The final version of the RCI is available in the Supplemental Material [2]. It consists of 24 multiple-choice questions. Throughout this paper individual questions are referred to by their RCI question number.

Each question also has an associated confidence scale. This asks the student to rate how confident they are in their answer. One of five options could be selected from the online form: guessing, unconfident, neutral, confident, and certain. Confidence measures have occasionally been used before with concept inventories [22,23], including in association with the Force Concept Inventory (FCI) [24].

Confidence information is potentially useful for gauging the quality of students' understanding. For example, consider a question that most students answer correctly. If they also expressed confidence in their answers, this would suggest mastery had been achieved. This was the case for the pair of questions 3 and 4 concerning the constancy of the speed of light. For questions that have high proportions of incorrect answers, high confidence might indicate misconceptions that may be difficult to dispel. This was the case for question 7 concerning a twin paradox type scenario. Low confidence might suggest that instruction about correct concepts may more often be successful.

III. STUDENTS

The RCI data analyzed in this paper were obtained from the 2012 ANU Physics 2 class [25]. This is the second physics course taken by physics majors. It is usually taken in the second semester of their first year at university. The class enrolment was 99, from whom 70 responses were obtained for the pre-test and 63 responses for the post-test, with 53 individuals taking both tests.

The relativity instruction was a three week module of nine lectures, a three hour simulation laboratory using the Real Time Relativity software [26], and three small-group problem-solving tutorials, developed at the ANU. It was assessed by two sets of weekly homework, a prelab problem, a lab log book, and a midterm exam question. The lectures were held in a studio space to encourage interaction and included clicker questions and small-group discussion.

The RCI was administered online in 30 minutes of scheduled class time, although those absent from class were able to complete it outside of class time (12 of 70 for the pre-test and 25 of 63 for the post-test). No significant differences in responses were found between those two groups. All questions were of equal value, with no partial marks given. The mean RCI score on the pre-test was 56%, and on the post-test 71%. For comparison, the expected mean score if answers were chosen randomly is 36%, with a standard deviation of about 1% (see Sec. IV B 1 for further explanation). These high scores should be considered in the context of the class being high academic achievers, as indicated by their median Australian Tertiary Admission Rank (ATAR) score of 95, out of a possible 99.95 [27].

IV. DATA ANALYSIS METHODS

In this section we analyze the data obtained from administering the RCI to the Physics 2 class. In Sec. IV A we use classical test theory to investigate the discrimination and consistency of the RCI. In Sec. IV B we investigate the correlations between students' responses to different RCI questions.

Our analysis is based on the simplifying assumption that students either understand a concept or do not. An alternative analysis, motivated by cognitive science, is that students may simultaneously have different knowledge frameworks for what experts would consider to be the same concept. This has led to an alternative approach called model analysis that seeks to identify these different knowledge frameworks [28]. It has been applied to the FCI [28] and to a conceptual survey of waves [29]. We have not attempted a model analysis.

As our sample size is small, we paid particular attention to the statistical significance of correlations. Where possible, we calculated the probability that the observed correlations might arise by chance from sampling noise rather than from actual properties of the underlying population: so-called p values. In the language of physics and engineering, we attempted to distinguish the signal from the noise [30].

For approximately normally distributed data, statistical significance was determined using standard deviations from the mean. Otherwise, we used either the Kolmogorov-Smirnov test [31] or Monte Carlo simulations to calculate the probability that the correlation could have arisen by chance. The Kolmogorov-Smirnov test is preferred over the χ-squared test for small sample sizes [32].

The Kolmogorov-Smirnov test determines the probability that the two data sets being compared are drawn from the same distribution. It uses the maximum difference D_{data} between the cumulative probability distributions of the two data sets. The probability distribution of D is known in the case that they are drawn from the same distribution. Hence, one can determine the probability that D exceeds the observed D_{data}. This is the probability p that the two data sets are drawn from the same distribution.

A. Classical test theory

Classical test theory provides a set of statistics for estimating the discrimination and consistency of a test. Discrimination is the capability to quantify students' understanding of the subject of the inventory.

TABLE II. RCI post-test statistics. Sample size $N = 63$ students. The desired ranges are those suggested by Ding and Beichner [34].

Statistic	RCI value	Desired range
Mean item difficulty	0.71	[0.3, 0.9]
Mean discrimination index	0.24	≥ 0.3
Ferguson's delta	0.96	≥ 0.9
Mean point biserial coefficient	0.36	≥ 0.2
KR20 reliability	0.74	≥ 0.7

Consistency is the extent to which each question is measuring the same broad understanding. Overviews have been given by Ding *et al.* [33] and Ding and Beichner [34].

Table II reports some test statistics for the RCI post-test. The desired ranges are boundaries, according to Ding and Beichner [34], beyond which consideration should be given to possible problems with the inventory. The item difficulty of question number i is the fraction of correct answers, $P_i = N_{\text{correct}}/N_i$, where N_i is the total number of answers to the question. Figure 1 shows the item difficulties for each question. The post-test RCI item difficulty averaged over all questions, $\langle P \rangle = 0.71$, tells us that the test was rather easy. However, as noted in the previous section, the class was particularly accomplished.

The only RCI statistic in Table II falling outside the desired range is the mean discrimination index. This compares the number of students whose total RCI results were in the top quartile to those in the bottom quartile. The discrimination index for a question takes the difference between the fraction of correct answers to that question from students in the top quartile $N_{i,T}$ and from those in the bottom quartile $N_{i,B}$: $D_i = N_{i,T}/(0.25N_i) - N_{i,B}/(0.25N_i)$. The mean discrimination index is the mean of the discrimination indices for all questions. The low RCI value in Table II is partially due to the ease of the RCI, which reduces discrimination because the difference in student performance between the top and bottom quartiles is less than for a difficult test. Questions 12, 13, 14, 20, and 24 had discrimination indices $D_i \leq 0$. Their range of item difficulties was $0.98 \geq P_i \geq 0.52$ with a mean of 0.85. These questions should be reconsidered in any RCI revisions. Indeed, in Sec. IV B 2 we recommend removing question 24, concerning mass-energy equivalence. Hence, the low mean discrimination index suggests how the RCI might be improved. Nevertheless, we next show that another measure of discrimination, Ferguson's delta, is within the acceptable range.

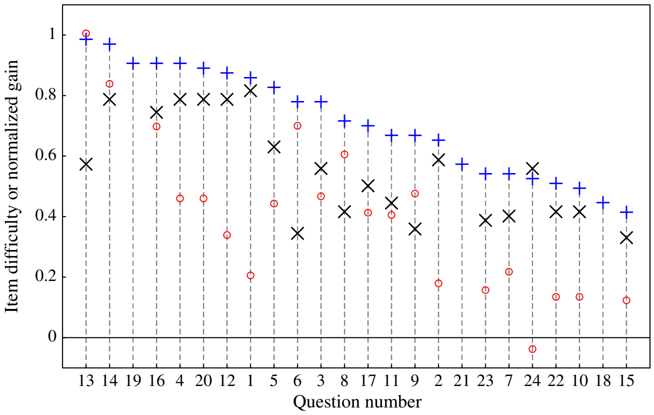

FIG. 1 (color online). RCI results by question for the Physics 2 class: the post-test item difficulties (blue +), pre-test item difficulties (black ×), and the normalized gain (red ○). The question number ordering is by postinstruction item difficulty. The sample sizes were 63 for the post-test and 70 for the pre-test, with 53 individuals doing both tests. Questions 18, 19, and 21 have no pre-test item difficulties or normalized gains as they were changed between the pre- and post-tests. The normalized gain is calculated for the students who took both the pre-test and the post-test. Hence, the normalized gain cannot be calculated using the plotted pre-test and post-test item difficulties, as they include additional students.

Ferguson's delta measures how the actual total scores are distributed in comparison to the possible range of scores. If only one particular score was ever achieved, then $\delta = 0$, while if all possible scores are achieved equally often, $\delta \approx 1$. Thus, Ferguson's delta measures the ability of the RCI to discriminate between students' understanding. It is defined to be [34]

$$\delta = \frac{N^2 - \sum_{i=1}^{K} f_i^2}{N^2 - N^2/(K+1)}, \tag{1}$$

where f_i is the number of times the total score was i, and $K = 24$ is the number of questions in the inventory. In contrast to the discrimination index, the RCI Ferguson's delta of $\delta = 0.96$ indicates that the RCI has adequate discrimination. We conclude that while the discrimination of the RCI might be improved, it is adequate.

The Pearson's r correlation between random variables X and Y is defined to be their covariance divided by the product of their standard deviations:

$$r_{XY} = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}, \tag{2}$$

where $\text{Cov}(X,Y) = \langle (X - \langle X \rangle)(Y - \langle Y \rangle) \rangle$ and $\text{Var}(X) = \langle (X - \langle X \rangle)^2 \rangle$. It ranges in value from $r = -1$ for perfectly anticorrelated variables, through $r = 0$ for uncorrelated variables, to $r = 1$ for perfectly correlated variables. For dichotomous variables, being right or wrong, Pearson's r may also be calculated using Eq. (6) given later in Sec. IV B 1.

In classical test theory the point biserial coefficient for a question is the Pearson r correlation between its item score and the total score for the inventory. Treating question answers as dichotomous variables, the point biserial coefficient for question number i can be expressed as [34]

$$r_{\mathrm{pbc},i} = (\langle X_{r,i}\rangle - \langle X_{w,i}\rangle)\sqrt{P_i(1-P_i)}/\sigma_X, \qquad (3)$$

where $\langle X_{r,i}\rangle$ is the mean total score for those who got the question right, $\langle X_{w,i}\rangle$ is the mean total score for those who got the question wrong, and σ_X is the standard deviation of the total score. The RCI mean point biserial coefficient over all post-test questions of $\langle r_{\mathrm{pbc}}\rangle = 0.36$ tells us that the RCI questions are consistent in what they measure.

The KR20 reliability statistic is another measure of the internal consistency of the inventory. It estimates the degree of correlation between the answers to questions. A value near 1 indicates that all questions are testing the same thing, while a value near 0 indicates that the answers are independent of each other. A value too close to 1 would be undesirable for the RCI, since it is intended to test a number of different concepts. However, as usual in physics, the concepts are interrelated, so that a deep understanding of relativity requires an understanding of all concepts, so a low value is also undesirable. The KR20 reliability statistic is defined to be [34]

$$r_{\mathrm{KR20}} = \frac{K}{K-1}\left(\sigma_X^2 - \sum_{i=1}^{K}\sqrt{P_i(1-P_i)}\right)\bigg/\sigma_X^2. \qquad (4)$$

The RCI reliability statistic of $r_{\mathrm{KR20}} = 0.74$ agrees with the mean point biserial coefficient that the RCI questions are consistent in what they measure.

Finally, we consider the changes from the pre-test to the post-test. Figure 1 shows the pre-test item difficulties and the normalized gain for those questions that did not change between the pre-test and post-test, namely, all except numbers 18, 19, and 21. The normalized gain for a question is defined to be the change in item difficulty divided by the maximum possible change in item difficulty, $g_i = (P_{i,\mathrm{post}} - P_{i,\mathrm{pre}})/(1 - P_{i,\mathrm{pre}})$ [35]. It is the fraction of the possible improvement that was achieved following instruction. The RCI normalized gain averaged over all questions was $\langle g\rangle = 0.40$. According to the Kolmogorov-Smirnov test, the probability that the pre-test and post-test results were sampled from the same population was $p = 4\times 10^{-6}$. Hence, we conclude that the normalized gain is statistically significant.

B. Question correlations

Correlations between students' responses to different questions can provide information on the reliability of concept inventories. They can also provide information about students' understanding, as we will show in Sec. V A.

As usual in statistical analysis, we assume that our sample, the Physics 2 class, is a subset of a larger population that we want to understand. This might be all students who have taken, or will take, a similar course. We assume that our sample of students is randomly chosen from the larger population and that its statistics estimate those of the larger population. However, in the particular sample, correlations can arise by chance even when no underlying correlation exists. Hence, it is important to calculate the statistical significance of correlations, especially with small sample sizes, such as ours. This tells us the probability that we might be misled by sample noise, and hence informs any action that might be taken based on the statistical evidence.

For example, for the 24 questions in the RCI there are $(24\times 23)/2 = 276$ possible correlations between question pairs. Using the post-test data to calculate these correlations, we find they have a distribution of values with mean $\langle r\rangle = 0.1$ and standard deviation of 0.15, consistent with a population mean of zero. The distribution is shown in Fig. 2. To understand why this distribution should alter our choice of statistical significance threshold, assume there was a hypothetical 5% chance of correlations above a certain strength occurring between any particular question pair, entirely due to random variation in the data. Then we would expect to find about $276\times 0.05 \approx 14$ so-correlated question pairs by chance. Choosing an acceptance threshold of $p < 1/276 \approx 4\times 10^{-3}$ ensures that in the long run less than one correlation is accepted due to sampling noise alone. Such care is required whenever there are many noisy channels in which a signal is being sought. However, it comes at the cost of an increased likelihood of missing correlations that in fact exist in the larger population.

A related problem is determining the significance of the absence of expected correlations. For example, consider two questions that were designed to test the same concept but that are not significantly correlated according to the student data. What strength of correlation can the data reliably rule out?

We have addressed such questions using Monte Carlo simulation. As this approach is not common in physics education research, we describe it in some detail in the next section.

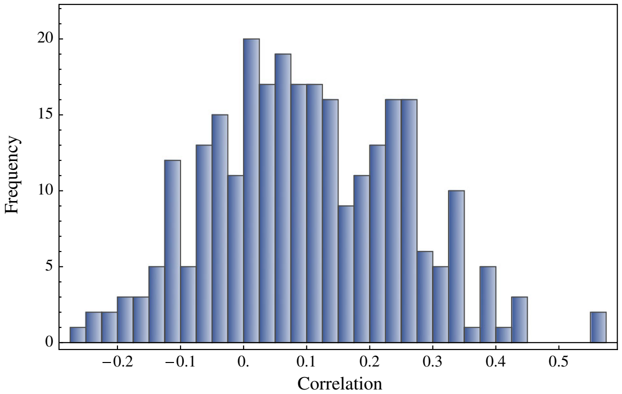

FIG. 2 (color online). Histogram of the Pearson's r correlations between all 276 question pairs from the post-test data. The correlations are calculated using Eq. (6), with the p_{XY} derived from the data.

1. Monte Carlo simulation

Our Monte Carlo simulations are based on stochastic models of the student population. Random samples are drawn from the model and their distributions used to estimate statistical significance. As models are simplified descriptions of students' responses, such estimates must be treated with care. Nevertheless, they help quantify the degree to which correlations in the data imply correlations in the larger population.

An example, concerning means rather than correlations, was given in Sec. III. The standard deviation in randomly answered mean scores was estimated from a model in which the answer to each question was chosen with uniform probability. The mean scores of samples of size $N = 70$ were approximately normally distributed with a mean of 36% and a standard deviation of about 1%. Since the pre-test mean of 56% is then about 20 standard deviations from the mean, we can conclude that the students are not guessing their answers.

More interesting is the estimation of the statistical significance of correlations between two questions. Let us call them Q1 and Q2. We code the question answers as correct (1) or incorrect (0). There are then four possible answers to the two questions: both correct, both incorrect, only Q1 correct, and only Q2 correct. Our model of the larger student population assumes that students' answers follow the multinomial distribution over these four possible outcomes.

Let p_{11} be the probability that both questions are answered correctly, p_{00} the probability that both are answered incorrectly, p_{10} the probability that only Q1 is answered correctly, and p_{01} the probability that only Q2 is answered correctly. The multinomial probability function is then [32]

$$\Pr(N_{11}, N_{00}, N_{10}, N_{01}) = \frac{N!}{N_{11}!N_{00}!N_{10}!N_{01}!} p_{11}^{N_{11}} p_{00}^{N_{00}} p_{10}^{N_{10}} p_{01}^{N_{01}}, \tag{5}$$

where N_{XY} is the number of XY outcomes from a sample of N answers. Three equations, in addition to the normalization, $p_{11} + p_{00} + p_{10} + p_{01} = 1$, specify the distribution. We take these to be the probability of a correct answer to Q1, $P_1 = p_{11} + p_{10}$, the probability of a correct answer to Q2, $P_2 = p_{11} + p_{01}$, and the Pearson's r correlation between the answers to Q1 and Q2,

$$r_{12} = \frac{p_{11}p_{00} - p_{10}p_{01}}{\sqrt{(p_{11} + p_{10})(p_{11} + p_{01})(p_{00} + p_{10})(p_{00} + p_{01})}}. \tag{6}$$

Hence, specifying P_1, P_2, and r_{12} determines the distribution. The first two are estimated by the item difficulties from the student data. In contrast, the correlation is chosen to test a significance hypothesis. For example, say the student data have a correlation of C, and we want to know whether this is significant. We then choose the model correlation to be $r_{12} = 0$. Taking Monte Carlo samples from the model [36], we can determine the probability that correlations equal to or larger than the observed correlation C arise from the model with zero correlation. If this probability is p, we would say that the observed correlation is statistically significant at the p level.

Monte Carlo significance testing of our post-test data found the seven correlations shown in Table III to be significant at the $p \le 10^{-3}$ level. From the argument at the beginning of Sec. IV B, these are unlikely to arise randomly. The first three are expected correlations between conceptually related questions. However, the others are unexpected. In the next section we explain the observed correlations between these conceptually unrelated questions using item response theory.

It is surprising that Table III does not contain more correlations between conceptually related questions. However, the fact that an observed correlation is not statistically significant does not, in itself, justify the conclusion that there is no correlation in the larger population. As far as the data alone are concerned, it leaves us uncertain either way.

One way of dealing with this problem is based on Bayes's theorem [30]. In our context, this approach assigns prior probabilities to correlations. These probabilities are then adjusted according to the statistical evidence from the data. This has the advantage that correlations that we have prior reason to believe exist, for example, between conceptually related RCI questions, are less likely to be rejected as noise than do correlations that we have no prior reason to believe exist. Although we will not use quantitative Bayesian statistics, the Bayesian framework helps explain the lack of expected correlations in Table III, as it takes no account of prior information.

Alternatively, further Monte Carlo simulations might show that sufficiently strong correlation values are unlikely. In cases for which we expected a correlation, this would justify a reconsideration of our reasons for that expectation. For example, we could select an assumed strong correlation C_A and set the model correlation equal to

TABLE III. Post-test correlations between questions statistically significant at the $p \le 10^{-3}$ level. The Pearson's r correlation is calculated using Eq. (6). The p values were obtained from 20 000 Monte Carlo samples for each question pair with zero correlations between questions.

Questions	Pearson's r	p value
1, 2	0.56	$<5 \times 10^{-5}$
5, 6	0.56	$<5 \times 10^{-5}$
11, 12	0.44	4×10^{-4}
3, 9	0.43	3×10^{-4}
15, 22	0.44	5×10^{-4}
2, 7	0.39	7×10^{-4}
9, 22	0.38	9×10^{-4}

it, $r_{12} = C_A$. From Monte Carlo simulations we could then determine the probability p that the simulated correlations are equal to or less than the observed correlation C, even though the model correlation is C_A. If this probability is sufficiently small, we may rule out the assumed correlation at the p level.

2. Item response theory

It is reasonable to assume that a major determinant of whether a student answers a question correctly is their academic ability. Given a question pair, strong students will tend to get both right and weak students will tend to get both wrong, strengthening the overall correlations. If this assumption is correct, then removing that part of students' performance due to academic ability may increase the correlations due to conceptual relations. (This idea was suggested to us by Dr. Paul Francis of ANU.) This may be achieved using item response theory [34].

Item response theory, sometimes called Rasch analysis [37], assumes that there is one parameter that describes the performance of student number j, their ability θ_j, and one parameter, b_i, that describes the difficulty of question number i. These are generated by a logistic regression algorithm [38] from the student data to provide a maximum likelihood estimate for the probability of student j getting question i correct from the model

$$P_{ij} = \frac{e^{(\theta_j - b_i)}}{1 + e^{(\theta_j - b_i)}}. \tag{7}$$

Let M_{ij} be the actual response of student j to question i, coded so 1 is correct and 0 incorrect. The residuals $R_{ij} = M_{ij} - P_{ij}$ measure the deviation of the particular student j and question i from the population of students and questions with the same respective ability and difficulty. According to item response theory, these residuals have the student ability and question difficulty factors removed. Hence, correlations between the residuals are due to factors other than student's ability and question difficulty.

We therefore calculated the correlations between the residuals for each question pair, averaged over all N students,

$$C_{ik} = \frac{1}{N} \sum_{j=1}^{N} R_{ij} R_{kj}. \tag{8}$$

These correlations were found to be approximately normally distributed with mean 0 (by construction of the P_{ij}) and standard deviation 0.02. We consider the statistically significant correlations to be those that are more than 3 standard deviations from the mean, that is, with a one-sided p value of $<2 \times 10^{-3}$. Table IV lists these.

The three positively correlated questions are precisely the conceptually related pairs in the raw scores correlation Table III. All the other correlations in Table III are absent.

TABLE IV. Item response theory residual correlations C_{ik}, statistically significant at the 3σ level, from the post-test data. The rightmost column is how many standard deviations C_{ik} is from the mean.

Questions	C_{ik}	σ
5, 6	0.08	4.0
1, 2	0.066	3.4
11, 12	0.066	3.4
7, 8	−0.083	3.6
23, 24	−0.086	3.7

Hence, student ability, as modeled by item response theory, explains the correlations between the raw scores of conceptually unrelated questions.

The last two rows in Table IV are anticorrelations, with one-sided p values of $\approx 3 \times 10^{-4}$. The first anticorrelation is surprising as both questions 7 and 8 were designed to test the concept of time dilation, and hence were expected to be positively correlated. However, as we shall see in Sec. VA, question 7 is unusual in being one of the two questions having an anticorrelation with confidence.

There is no obvious relation between the second anticorrelated pair, questions 23 (causality) and 24 (mass-energy). However, question 24 is unusual in being the only question with a negative normalized gain, as can be seen in Fig. 1. Hence, we recommend that question 24 be removed from the RCI.

V. RESULTS

The previous section focused on statistical methods and their application to establishing the consistency and reliability of the RCI. In this section the focus is on the implications of the RCI results for special relativity education. We first consider some of the misconceptions revealed by the RCI and then show that the RCI is gender biased.

A. Misconceptions

For our analysis we numerically coded the five confidence options as guessing (0), unconfident (0.25), neutral (0.5), confident (0.75), and certain (1). For the pre-test data the mean confidence over all questions and all students was 0.5 (neutral), and for the post-test it was 0.68 (neutral to confident). According to the Kolmogorov-Smirnov test the probability that the pre-test and post-test confidence data were sampled from the same population was $p < 10^{-5}$. Hence, there was a significant increase in confidence after instruction.

The average of the Pearson's r correlation, Eq. (2), between students' confidence and their score for each question was $\langle r_i \rangle = 0.11$ for the pre-test and $\langle r_i \rangle = 0.19$ for the post-test. These are different at the $p = 0.01$ level of significance. Hence, after instruction students not only became more confident but were also more likely to answer correctly *and* confidently.

Most individual questions in the post-test had a positive correlation between confidence and score, which indicates some mastery of the relevant concepts. However, two questions had negative correlations: question 7 ($r_7 = -0.3$) and question 23 ($r_{23} = -0.2$), significantly different from 0 with $p \lesssim 0.05$. These negative correlations suggest gaps in students' postinstruction mastery.

Question 7 had nearly equal numbers of correct and incorrect answers: item difficulty $P_7 = 0.54$. Of those students who rated their confidence as either certain or confident, nearly equal numbers answered correctly and incorrectly. This indicates a misconception about time dilation, which is not captured by the other time dilation questions 5, 6, and 8 that have positive correlations between confidence and score of $r = 0.2, 0.25, 0.4$, respectively. One difference between these questions is that the latter are phrased in terms of observations, whereas question 7 is about an experience: traveling across the galaxy. It may be that students are displaying the misconception that while time dilation applies to observations of things, it does not apply to the things themselves.

The other negatively correlated question is 23, concerning the concept of causality. Most of those who answered it correctly rated their confidence as either guessing or unconfident, suggesting a weak conceptual understanding.

Questions 5 and 6 of the RCI are a pair testing understanding of time dilation. They ask about the same situation from two different inertial reference frames, with each observer measuring the other's clock to run slow. Their pre-test item difficulties were $P_{5,\text{pre}} = 0.63$ and $P_{6,\text{pre}} = 0.34$, the difference being significant at the $p = 0.05$ level. Furthermore, their answers were anticorrelated, $r_{56,\text{pre}} = -0.25$, significant at the $p = 0.02$ level.

Correct relativistic thinking would recognize the symmetry between the two reference frames and hence lead to correlation between the answers. However, the anticorrelation suggests an asymmetry misconception in which A measuring B's clock to run slow implies B measuring A's clock to run fast. This is related to absolute motion misconceptions regarding Galilean relativity reported by Panse *et al.* [10]. The following student comment from a Real Time Relativity [26] lab session on time dilation is an example of both the absolute rest frame and asymmetry misconceptions:

The clocks are stationary, and I'm moving ... so my clock is running slow, which is why the clocks are running fast compared to mine ....

As Tables III and IV show, the post-test questions 5 and 6 were the most highly correlated of all pairs, with $r_{56,\text{post}} = 0.56$, significant at the $p \leq 5 \times 10^{-5}$ level. This indicates that relativistic thinking has been achieved after instruction and the asymmetry misconception reduced. The post-test item difficulties were $P_{5,\text{post}} = 0.83$ and $P_{6,\text{post}} = 0.78$, with corresponding normalized gains of $g_5 = 0.54$ and $g_6 = 0.67$.

Evidence from class assessment items indicated that the asymmetry misconception also occurred for length contraction. However, the RCI has no symmetrical pair of length contraction questions to test this. Hence, we recommend that a symmetrical partner question be added to the existing RCI length contraction question 13.

B. Gender differences

In the Physics 2 class we found statistically significant gender differences in the RCI results. The pre-test was taken by 19 females and 51 males, the post-test by 18 females and 45 males. Of those who took both tests 15 were female and 38 were male. As shown in Table V, males scored higher than females in the pre-test, post-test, normalized gain, and in confidence. All these differences are significant at the $p \leq 0.05$ level according to the Kolmogorov-Smirnov test.

In contrast, the gender groups were statistically identical for assessable homework and for the midterm exam relativity question. There was also no difference in prior achievement as measured by the ATAR score (discussed in Sec. III).

There were only four individual questions for which the gender difference was statistically significant ($p < 0.05$): questions 1 and 2 concerning inertial frames, question 9 concerning velocity addition, and question 17 concerning length contraction. In each of these cases the difference in item difficulty between males and females was ≥ 0.27. For more than half the questions the magnitude of this difference was ≤ 0.1.

Similar results have been reported for the Force Concept Inventory [39–42] and Brief Electricity and Magnetism Assessment (BEMA) [43]. There is a report of the FCI

TABLE V. RCI statistics by gender for the Physics 2 class. $\langle P \rangle$ is the mean item difficulty, $\langle g \rangle$ is the mean normalized gain, $\langle c \rangle$ is the mean confidence, $\langle x_{\text{exam}} \rangle$ is the mean exam score (fraction of possible score) for the students who did the post-test, and $\langle x_{\text{hw}} \rangle$ is the mean homework score (fraction of possible score). The ATAR is the university admission score discussed in Sec. III. p values are the probability that the female and male data were sampled from the same population, so that the observed difference is due to chance.

Statistic	Females	Males	p value
$\langle P_{\text{pre}} \rangle$	0.50	0.58	0.02
$\langle P_{\text{post}} \rangle$	0.63	0.72	0.003
$\langle g \rangle$	0.23	0.38	0.05
$\langle c_{\text{pre}} \rangle$	0.41	0.53	0.02
$\langle c_{\text{post}} \rangle$	0.64	0.70	0.04
$\langle x_{\text{exam}} \rangle$	0.66	0.67	0.95
$\langle x_{\text{hw}} \rangle$	0.75	0.75	1
$\langle \text{ATAR} \rangle$	94.2	93.5	0.96

gender gap being eliminated by high levels of interactive engagement [44], although this has not been found in other studies [45]. Other inventories have also been found to have gender differences [46,47].

Although some authors have claimed that multiple-choice tests are inherently gender biased, the largest studies have found no such effect [48,49].

VI. CONCLUSIONS

Classical test theory suggests that the RCI may be too easy and, perhaps consequently, insufficiently discriminating. However, we do not recommend revisions, other than those suggested below, until data from a wider range of students have been analyzed.

In Sec. IV B 2 we concluded that question 24, concerning the concept of mass-energy equivalence, should be removed from the RCI. It has zero discrimination, and is the only question having a negative normalized gain between the pre-test and post-test. It was also found to have a strong negative correlation with an apparently unrelated question. If removed, the concept of mass-energy equivalence would not be tested by the RCI.

In Sec. V A we concluded that a frame symmetrical pair of length contraction questions is desirable, mirroring the symmetrical pair of time dilation questions. Hence, we recommend that a partner question be added to the existing RCI length contraction question 13. However, any such question would require validation along the lines described in Sec. II.

The evidence presented in Sec. V B suggests that the RCI is gender biased. Previous work has shown similar biases in the Force Concept Inventory and in other concept inventories. Concept inventories are useful because they can help evaluate innovation and hence improve teaching. However, if their evaluations are biased with respect to certain student groups, there is a risk that improved learning for some comes at the expense of the learning of others. It is a task for future physics education research to investigate and understand this interesting and important problem.

ACKNOWLEDGMENTS

We would like to thank the academics and students who helped develop the RCI. We would especially like to acknowledge A. Wilson for her help with the student interviews, and P. Francis for his advice on statistical analysis.

[1] Assessment Instrument Information Page, North Carolina State University Physics Education R and D Group, http://www.ncsu.edu/per/TestInfo.html, accessed 2013.

[2] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevSTPER.9.010118 for the RCI post-test, correct answers, and response rates for each answer.

[3] M. Wegener, T. McIntyre, D. McGrath, C. Savage, and M. Williamson, Developing a virtual physics world, Australas. J. Educ. Tech. **28**, 504 (2012), special issue no. 3 [http://www.ascilite.org.au/ajet/ajet28/wegener.html].

[4] D. McGrath, M. Wegener, T. McIntyre, C. Savage, and M. Williamson, Student experiences of virtual reality—A case study in learning special relativity, Am. J. Phys. **78**, 862 (2010).

[5] C. M. Savage, A. Searle, and L. McCalman, Real Time Relativity: Exploratory learning of special relativity, Am. J. Phys. **75**, 791 (2007).

[6] Peter W. Hewson, A case study of conceptual change in special relativity: the influence of prior knowledge in learning, Eur. J. Sci. Educ. **4**, 61 (1982).

[7] George J. Posner, Kenneth A. Strike, Peter W. Hewson, and William A. Gertzog, Accommodation of a scientific conception: Toward a theory of conceptual change, Sci. Educ. **66**, 211 (1982).

[8] A. Villani and J. L. A. Pacca, Students' spontaneous ideas about the speed of light, Int. J. Sci. Educ. **9**, 55 (1987).

[9] N. D. Mermin, Lapses in relativistic pedagogy, Am. J. Phys. **62**, 11 (1994).

[10] Sudhir Panse, Jayashree Ramadas, and Arvind Kumar, Alternative conceptions in Galilean relativity: Frames of reference, Int. J. Sci. Educ. **16**, 63 (1994).

[11] Jayashree Ramadas, Shrish Barve, and Arvind Kumar, Alternative conceptions in Galilean relativity: Inertial and non-inertial observers, Int. J. Sci. Educ. **18**, 615 (1996).

[12] R. E. Scherr, P. S. Shaffer, and S. Vokos, Student understanding of time in special relativity: Simultaneity and reference frames, Am. J. Phys. **69**, S24 (2001).

[13] R. E. Scherr, P. S. Shaffer, and S. Vokos, The challenge of changing deeply held student beliefs about the relativity of simultaneity, Am. J. Phys. **70**, 1238 (2002).

[14] M. Belloni, W. Christian, and M. H. Darcy, Teaching special relativity using Physlets®, Phys. Teach. **42**, 284 (2004).

[15] R. E. Scherr, Modeling student thinking: An example from special relativity, Am. J. Phys. **75**, 272 (2007).

[16] Sébastian Cormier and Richard Steinberg, The twin twin paradox: Exploring student approaches to understanding relativistic concepts, Phys. Teach. **48**, 598 (2010).

[17] Wendy K. Adams and Carl E. Wieman, Development and validation of instruments to measure learning of expert-like thinking, Int. J. Sci. Educ. **33**, 1289 (2011).

[18] K. Gibson, Ph.D. thesis, Arizona State University, 2008.

[19] S. V. Chasteen, R. E. Pepper, M. D. Caballero, S. J. Pollock, and K. K. Perkins, Colorado Upper-Division Electrostatics diagnostic: A conceptual assessment for

the junior level, Phys. Rev. ST Phys. Educ. Res. **8**, 020108 (2012).

[20] Edwin F. Taylor and John Archibald Wheeler, *Spacetime Physics: Introduction to Special Relativity* (W. H. Freeman & Co., New York, 1982); Nathaniel David Mermin, *It's About Time* (Princeton University Press, Princeton, 2005); Sean Carroll, *Spacetime and Geometry* (Addison-Wesley, San Francisco, 2004); Wolfgang Rindler, *Relativity, Special, General and Cosmological* (Oxford University Press, Oxford, England, 2006).

[21] Expert input was solicited by Email from the members of the Australasian Society for General Relativity and Gravitation and from members of the Education Group of the Australian Institute of Physics. Responses were also obtained through a posting to the Matter and Interactions Yahoo group.

[22] Kirk Allen, Teri Reed-Rhoads, and Robert Terry, Work in progress: Assessing student confidence of introductory statistics concepts, Proc. Front. Educ. Conf., 13 (2006).

[23] Kirk Allen, Andrea Stone, Teri Reed-Rhoads, and Teri J. Murphy, The Statistics Concepts Inventory: Developing a valid and reliable instrument, *Proceedings of the American Society for Engineering Education Annual Conference and Exposition, 2004*, https://engineering.purdue.edu/ASEE/SCI/pubs/ASEE%202004%20SCI.pdf.

[24] Manjula Devi Sharma and James Bewes, Self-monitoring: Confidence, academic achievement and gender differences in physics, J. Learn. Des. **4**, 1 (2011) [https://www.jld.edu.au/article/view/76].

[25] The data collection for the analysis presented in this paper was approved by the ANU Human Research Ethics Committee: Human ethics protocol 2012/380.

[26] Real Time Relativity, http://realtimerelativity.org, accessed 2013.

[27] The ATAR score ranks graduating secondary school students. The Physics 2 median ATAR score of 95 indicates that the median student scored higher than 95% of other students that were in their cohort in year 7.

[28] Lei Bao and Edward F. Redish, Model analysis: Representing and assessing the dynamics of student learning, Phys. Rev. ST Phys. Educ. Res. **2**, 010103 (2006).

[29] Apisit Tongchai, Manjula Devi Sharma, Ian D. Johnston, Kwan Arayathanitkul, and Chernchok Soankwan, Consistency of students' conceptions of wave propagation: Findings from a conceptual survey in mechanical waves, Phys. Rev. ST Phys. Educ. Res. **7**, 020101 (2011).

[30] Nate Silver, *The Signal and the Noise* (Penguin Press, New York, 2012), Chaps. 5 and 8.

[31] William H. Press, Brian P. Flannery, Saul A. Teukolsky, and William T. Vetterling, *Numerical Recipes: The Art of Scientific Computing* (Cambridge University Press, New York, 1986), 1st ed.

[32] Dennis D. Wackerly, William Mendenhall, and Richard L. Scheaffer, *Mathematical Statistics with Applications* (Thomson, Belmont, 2008), 7th ed.

[33] Lin Ding, Ruth Chabay, Bruce Sherwood, and Robert Beichner, Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment, Phys. Rev. ST Phys. Educ. Res. **2**, 010105 (2006).

[34] L. Ding and R. Beichner, Approaches to data analysis of multiple-choice questions, Phys. Rev. ST Phys. Educ. Res. **5**, 020103 (2009).

[35] Richard R. Hake, Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses, Am. J. Phys. **66**, 64 (1998).

[36] We used the MultiNomialDistribution function of MATHEMATICA 8 to generate our random samples.

[37] Maja Planinic, Lana Ivanjek, and Ana Susac, Rasch model based analysis of the Force Concept Inventory, Phys. Rev. ST Phys. Educ. Res. **6**, 010103 (2010).

[38] We used MINISTEP to do the Rasch analysis: http://www.winsteps.com/ministep.htm, accessed 2013.

[39] Laura McCullough, Gender, context and physics assessment, J. Int. Wom. Stud. **5**, 20 (2004) [http://www.bridgew.edu/soas/jiws/May04_Special/Gender.pdf].

[40] R. D. Dietz, R. H. Pearson, M. R. Semak, and C. W. Willis, Gender bias in the Force Concept Inventory?, AIP Conf. Proc. **1413**, 171 (2011).

[41] Jennifer Docktor and Kenneth Heller, Gender differences in both Force Concept Inventory and Introductory Physics Performance, AIP Conf. Proc. **1064**, 15 (2008).

[42] Vincent P. Coletta, Jeffery Phillips, and Jeffery J. Steinert, FCI normalized gain, scientific reasoning ability, thinking in physics, and gender effects, AIP Conf. Proc. **1413**, 23 (2011).

[43] L. E. Kost-Smith, S. J. Pollock and N. D. Finkelstein, Gender disparities in second-semester college physics: The incremental effects of a smog of bias, Phys. Rev. ST Phys. Educ. Res. **6**, 020112 (2010).

[44] Mercedes Lorenzo, Catherine H. Crouch, and Eric Mazur, Reducing the gender gap in the physics classroom, Am. J. Phys. **74**, 118 (2006).

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

[46] Robert J. Beichner, Testing student interpretation of kinematics graphs, Am. J. Phys. **62**, 750 (1994).

[47] W. K. Adams, K. K. Perkins, N. S. Podolefsky, M. Dubson, N. D. Finkelstein, and C. E. Wieman, New instrument for measuring student beliefs about physics and learning physics: The Colorado Learning Attitudes about Science Survey, Phys. Rev. ST Phys. Educ. Res. **2**, 010101 (2006).

[48] J. McKendree and C. Smith, FAQ—I have heard that multiple-choice questions (MCQs) are biased in favour of males. What is the evidence for this?, http://www.heacademy.ac.uk/resources/detail/subjects/medev/FAQ-I-have-heard-that-multiple-choice-questions-are-biased-in-favour-of-males, accessed 2013.

[49] N. Cole, "The ETS gender study: How females and males perform in educational settings," Educational Testing Service Technical Report, 1997, http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED424337, accessed 2013.