Relativity Concept Inventory

John Stewart Aslanides

**A thesis submitted for the degree of
Bachelor of Science with Honours in Physics of
The Australian National University**

October, 2012

Declaration

This thesis is an account of research undertaken between February and October 2012 at the Department of Quantum Science and Physics Education Centre in the Research School of Physics and Engineering, The Australian National University, Canberra, ACT, Australia.

Except where acknowledged in the customary manner, the material presented in this thesis is, to the best of my knowledge, original and has not been submitted in whole or part for a degree in any university.

John Stewart Aslanides
October, 2012

Acknowledgements

This year has been amazing. Honours is an experience I will never forget, not least because it has profoundly changed who I am. This change didn't occur spontaneously, though - it was with the help and tireless effort of several people that this was possible, and it is to them that I owe deep and sincere thanks; you are all tremendously important to me.

Firstly, thank you to my supervisor Craig Savage, who has been a role model to me, both as a scientist and as a person. Your patience, generosity, wisdom, and humour have uplifted me and inspired me. My time working with you is the one thing I will cherish the most from this year - so thank you Craig, for being a great mentor and a great friend.

Thank you to my parents Timoshenko and Jenny, for your unconditional love and support, for your advice and forbearance when I was down, and for your praise and jubilance when I was up. I owe everything to you, and then some. I love you.

Lastly, to my beautiful girlfriend Lulu. Thank you for being who you are, for your constant affection and understanding, and for putting up with all my silly habits while I wrote my thesis. Getting to know you this past year has changed my life. Happy anniversary, baby! I love you so much.

In addition, I would also like to give thanks to:

- Anna Wilson, for her advice on how to do the student interviews, and for having them transcribed for me.
- Susan Scott, Joe Hope, Roy Kerr, and other participants in the expert survey, for their helpful comments and insights.
- Don Koks and Edwin Taylor, for the illuminating discussions about relativity, and for their detailed suggestions and remarks.
- Aaron Titus, for administering the RCI as a post-test in his class, and for his thoughtful input on the RCI itself.
- Emmanuel, Jarrod, Hannah, Paul, and the rest of the honours class for making honours fun - I'm going to miss you guys!

Abstract

Concept inventories are increasingly being used as formative assessments in science education, and they are proving to be powerful instruments for evaluating the effectiveness of instructional strategies. Until now, no concept inventory for special relativity has been published; in this thesis, we detail the development and use of the Relativity Concept Inventory (RCI). We explore the validity and reliability of this instrument with the standard statistical techniques of classical test analysis, factor analysis, and item response theory.

We also extend these techniques in a number of important ways; a critical part of the internal validation of a concept inventory is the analysis of correlations between items on the inventory. However, much of the existing concept inventory literature neglects the statistical significance of these correlations. We find that Monte Carlo simulations are a useful supplement to the standard methodology, which can be used to rigorously estimate the statistical significance of correlations.

We also explore the use of student self-assessed confidence data, an approach that is under-developed and under-discussed in the concept inventory literature.

We use data from administering the RCI, in conjunction with summative assessments to make inferences about our class. We report a significant gender difference in the inventory; males achieved higher normalised gains, significant at the 5% level. We also discuss some hitherto unstudied student misconceptions, related to the "absolute rest frame" misconception discovered previously, in a Galilean context.

Contents

Chapter 1

Introduction

"Professor Mazur, how should I answer these questions? According to what you taught us, or by the way I think about these things?"[1]

In the past half-century, the role of the introductory physics course at universities has changed. It is no longer the case, as it was half a century ago or more, that university physics classes consist mainly of highly gifted and motivated students intending to make their careers in physics [2]. The modern workplace requires people in many professions to have some knowledge of science, including basic physics. As a result, a typical introductory physics class nowadays consists of a diverse group of students, covering a broad spectrum of academic background, programme focus, and expectations of the course. It has been argued that the goals of the modern physics teacher should include adjusting to this change and ensuring that their teaching is as effective as possible in this new classroom context [3]. In Lasry, Mazur, et al.'s article "Are most people too dumb for physics?", they argue that *"The appealing—yet suspiciously conceited—notion that physics is only for smart or industrious people is in fact quite questionable"* [4].

In comparison with physics research, the science of physics teaching remains very undeveloped. The Nobel laureate Carl Wieman argues strongly for a new approach to teaching, emphasising the need for using the scientific method [3]. Wieman is is not alone: Reif and Redish, two prolific PER investigators in the 1990s, independently gave damning assessments of the unscientific approach many scientists take to teaching [5, 6]. Hammer [7] summarised their conclusions: "Both Reif and Redish noted a contrast between the careful thought physicists apply to physics and the 'seat of the pants' approach many take to teaching. Like the intuition physics-naive students have developed from their extensive, unstudied experience of the physical world, and in which they may be confident, these instructors' intuitions are inadequate and often incorrect."

In this context, it is important to develop the most effective teaching methods possible, particularly in relation to the most difficult areas of physics, such as relativity. This project aims to address the question: "How do we teach relativity better?" In particular, this project addresses the important subsidiary question: "How do we measure student understanding of special relativity better, and what does such measurement tell us?"

[1] Question from a Harvard first year physics student to Professor Eric Mazur, upon being given the Force Concept Inventory as an in-class quiz. Extract from [1].

The role of physics education research

In the early 1980s there was a growing awareness among physics educators that in the new classroom context, traditional teaching methods were not effective for the majority of students [8]. In response, an increasing number of academics and educators engaged in the line of enquiry now called physics education research, the purpose of which is to probe students' understanding (or lack thereof) and to modify teaching methods and materials on the basis of research findings. Physics education research has since yielded a rich literature of studies and experiments. The major proponents of physics education research argue that changes in teaching methodologies should be informed by results from cognitive science and from studies in physics classrooms themselves [3, 6]. Results from such studies are forming the basis for teaching techniques that are slowly being adopted worldwide.

Just as in physics itself, a key issue in physics education research is that of standards and measurement [9]. Measuring students' conceptual understanding of topics in physics is generally time consuming and difficult. Performance in standard summative assessments such as exams and homework problems are often an insufficient measure of students' understanding; it has been shown in numerous cases that it is possible for students with strong misconceptions to perform well in such assessments (see [2, 10], and references therein). Standard techniques for researchers to understand student thinking are the use of one-on-one student interviews and careful analysis of assessment questions; both being time consuming processes.

For this reason, it is desirable for physics education researchers to possess refined instruments with which to efficiently and accurately probe the conceptual understanding of a large cohort of students. A common name given to an instrument that probes students' conceptual understanding of a given topic is the *concept inventory.* A valid and well constructed concept inventory will help address the issues previously mentioned - to set a standard for instruction, and to give a reliable means of measuring student understanding of basic concepts in a given physics topic.

Concept Inventories

Physics is distinctive among the academic disciplines in its dependence on general principles and broad concepts; thinking like a physicist involves applying a small number of key concepts to make inferences about a broad range of phenomena. McDermott [8] summarises a lot of research with the following generalisation: "Facility in solving standard quantitative problems is not an adequate criterion for functional understanding. *Questions that require qualitative reasoning and verbal explanation are essential.*" She deems it insufficient for students to memorise specific formulae and problem-solving recipes; yet this is the approach that many students take, independent of the method of instruction. This is referred to in some of the literature as the "hidden curriculum" - the things the students perceive to be expected of them, independent of the nominal learning objectives of the course [2]. It was in the context of these findings that the first concept inventory was constructed.

The purpose of the concept inventory is to probe student understanding in a given topic, primarily as a formative assessment - to evaluate the effectiveness of instruction.

The process of improving physics pedagogy involves a lot of experimentation; formative assessments such as concept inventories serve to guide this process by giving a quantitative measurement of student understanding before and after instruction. Hake's study [11], using a large amount of compiled FCI data, shows the value of concept inventories in this respect, as an indicator of the relative effectiveness of two different teaching methodologies.

Concept inventories are generally required to be multiple choice, so as to facillitate administration, grading, and interpretation of results. In particular, care must be taken to ensure that as much as possible, inventory questions do not conflate conceptual understanding with other factors such as exam technique, problem solving skills, reading comprehension, or technical skill. This places constraints on the subject matter of the inventory, the level of difficulty of the questions, and the presentation of the concepts.

A procedure to develop and validate these instruments has been created and refined by Adams and Wieman [12]. The authors emphasise scientific rigour and objectivity in the development and validation process; when the goal is to produce a reliable assessment instrument for wide use, it is critical that the development stages be sound. The validation process ensures that a concept inventory measures what it purports to measure.

Why special relativity?

The significance of special relativity for contemporary physics is pivotal - all fundamental theories of nature are formulated relativistically. For those students choosing to major in physics, studying special relativity is then indispensable - moreover, it is introduced in most introductory university textbooks, which has lead to its being taught in first year curricula, and even in many high schools at the senior level [13].

Unlike quantum mechanics - the other pillar of contemporary physics - the formulation of special relativity is mathematically simple, and this makes it more accessible for students at the introductory level. Moreover, the deductive approach used in most first courses in relativity exposes students to the process of thinking like a scientist - a crucial part of their education if they are to succeed in science. From the student's perspective, special relativity is consistently the topic they most anticipate learning about in first year physics - this has been borne out in both formal and informal class surveys [14].

Learning special relativity requires the student to let go of their intuitions about space and time, which have been developed over years of everyday experience. One of the main difficulties of teaching relativity is that the consequences of the theory are so remote from this experience. Even though applications of relativity appear in some "real-world" applications (e.g. the Global Positioning System) and in "big-science" (e.g. the Large Hadron Collider), its more profound results are as removed from experience as they were in Einstein's time. Considerable effort has been spent to reduce this abstraction and help students to visualise relativistic effects, most notably in the Real Time Relativity collaboration between the Australian National University and the University of Queensland [14].

From a cognitive science perspective, the challenges to student learning of relativity are similar to those in other areas of physics, in particular mechanics. In the mechanics context, some students have built up an everyday intuition of kinematics and dynamics that is largely incompatible with Newtonian mechanics. Likewise with special relativity, which deals in the fundamentals of space and time. The challenge for students in both contexts is to confront their deep-seated and often implicit assumptions about the nature of reality, and replace them with scientific conceptions - in the literature, this process is normally referred to as conceptual change [15].

Replacing students' prior conceptions of spacetime is not only difficult, but it can be unclear whether or not the replacement has been entirely successful. Rachel Scherr, in an extended study at the University of Washington, found that many students attribute the relativity of simultaneity to the difference in signal travel time for different observers, and so "reconcile statements of the relativity of simultaneity with a belief in absolute simultaneity and fail to confront the startling ideas of special relativity" [16]. Situations like this are common in physics tuition; students nominally know the material and may perform well on assessments, but when their conceptual understanding is deeply probed, they often demonstrate that instruction has done little to displace their existing beliefs.

If we accept the proposition that it is worthwhile to teach special relativity at the introductory level, and we accept the arguments (made by Scherr and others) that there are deep subtleties in the process of conceptual change in relativity that need to be addressed, then we may conclude that appropriate conceptual tests should be developed. This is where the Relativity Concept Inventory comes in.

The Relativity Concept Inventory

The goal of this project is to develop and validate a Relativity Concept Inventory (RCI). One previous attempt has been made at constructing a RCI, as part of a doctoral thesis by Kevin Gibson at Arizona State University [17]. However, the scope of that inventory was narrow and not well-defined, and no attempt was made to validate the test items, either through expert review, or with student interviews. An exploratory factor analysis[2] of student responses was inconclusive, although some of the results were useful for this project. No publications arose from Gibson's thesis.

Our RCI is being developed in a similar format to the FCI: a multiple choice test of around 25 questions, that is designed to take around half an hour for students to complete. The list of concepts to be tested is to be defined by a survey of special relativity educators and experts, and by interviews of undergraduate students at ANU.

In their treatise on concept inventory development, Adams and Wieman emphasise the importance of iterations - the process of trial and error. There is no universal "recipe" to create a fully formed and useful concept inventory on the first attempt. A painstaking process of interviews (both of subject experts and of students), analysis of student responses, trials, statistical analysis and experimentation is required to make incremental improvements towards an end product. Combine this with the fact that it is only feasible to do large-scale trials when a physics class is studying the topic of interest, and the timescale for developing a concept inventory can easily spread to several years, going through many iterations and revisions [10, 12].

[2] Factor analysis is a commonly used tool in psychometrics and test analysis, which we will describe in section 5.7.

A clear and significant difficulty in developing an RCI is the subtlety of the concepts involved, and in presenting them in an way that is easily understandable across different institutions and student populations. Although the presentation of special relativity is usually not as technical as that of electrodynamics or quantum mechanics, it is highly abstracted from everyday experience. Relativistic thought experiments (the scenarios around which conceptual questions are constructed) are generally stranger and more complex than the analogous scenarios in the Newtonian context. A great advantage of the FCI is that its subject matter allows the questions to be phrased in terms of everyday scenarios, and with minimal use of technical language. In topics of greater complexity and abstraction, some sort of compromise between testing concepts and technical skill is sought [18]. Another difficulty peculiar to the development of an RCI is the relative rarity of experts in the field, in comparison to mechanics, electromagnetism, or even quantum mechanics. For this reason, a special online survey had to be constructed to collect expert opinion on the proposed list of concepts to be tested.

Research questions and scope of the project

The main research question for this project is: "How do we measure student understanding of special relativity better, and what does such measurement tell us?" The literature seems to suggest that concept inventories are a valuable tool for education research, and an effective way of investigating student understanding. As no concept inventory yet exists for special relativity, this suggests that the development of an RCI is a step towards answering the research question.

This project is focused on the initial development and validation of a concept inventory, using the first year second semester *Physics* 2 class at ANU as a trial. The student body consisted of $N = 99$ students. After reviewing the relevant literature and polling a selection of international relativity experts, we developed a draft version of the RCI. After some preliminary testing and student interviews, this draft was administered to the class, as a pre-test and, with some modifications, as a post-test, after three weeks of instruction on special relativity. The results were quantitatively analysed, with an emphasis on statistical rigour where it was appropriate. On the basis of this evidence, a working "beta" version of the RCI is now available for further testing and refining, and can be found in appendix A.

This thesis aims to add to the rich and expanding concept inventory literature, and to serve as a detailed account of the process of developing and validating a concept inventory. It will also fill a gap in the special relativity education literature, of which there is still very little published. The developments in physics education research over the last 20 years have furnished Newtonian mechanics and electromagnetism with a wealth of survey instruments and associated studies and experiments; many studies based on these instruments have proved highly fruitful. This project will attempt to illuminate the way for relativists to start along the same path.

Outline of the thesis

- Chapter 2 reviews the literature on physics education and concept inventories in particular. This chapter provides the background for the approach we take in designing

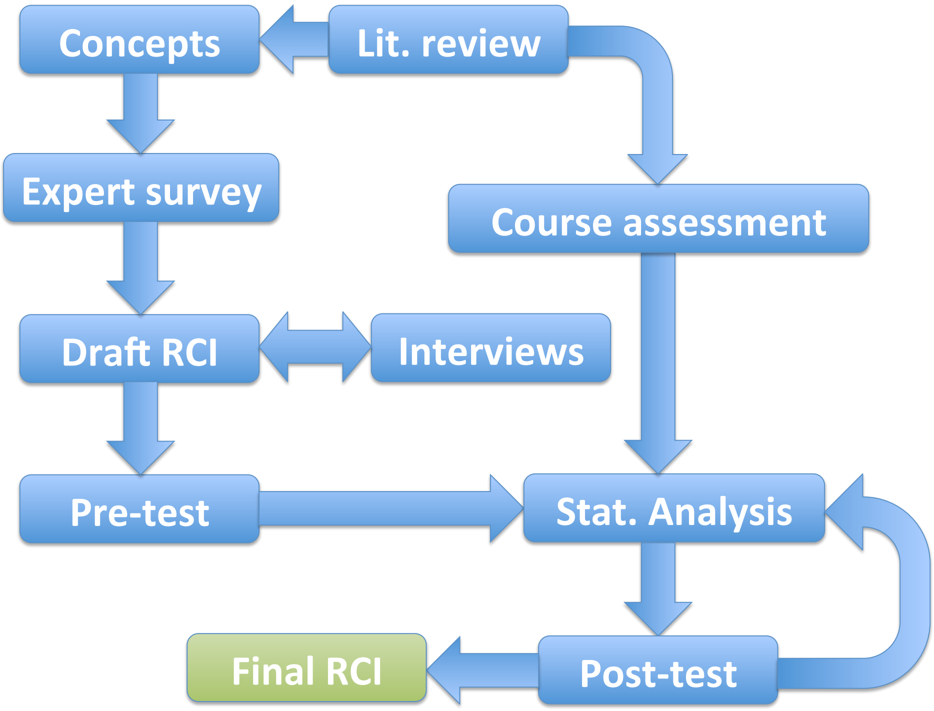


Figure 1.1: Flow chart outlining how different aspects of the project relate to each other. In particular, the arrows indicate how each step informs subsequent steps - this summarises the iterative development process.

and using the RCI.

- Chapter 3 describes conceptions and misconceptions in special relativity. This allowed us to focus our investigation, and provided a framework around which we constructed the RCI itself.

- Chapter 4 details the design stage. The centrepiece of this chapter is the expert survey, in which we take an overview of the opinions of 30 international relativity experts on concepts in relativity.

- Chapter 5 describes the statistical methods used in our analysis, and provides the details of the tools used, including their strengths and weaknesses. This chapter shows how simulation techniques can be used to make rigorous inferences from the data.

- Chapter 6 presents and interprets the results from the classroom study. In particular, it evaluates the validity RCI as an instrument, and also uses it to help draw conclusions about our students' learning.

- Chapter 7 has some concluding remarks, recommendations for teachers, and suggestions for further work.

- The Relativity Concept Inventory itself can be found in Appendix A.

Figure 1.1 shows the conceptual links between these chapters.

Summary of important results and methods

The primary new results and original work presented in this thesis are listed in the following dot points:

- Monte Carlo methods were used to simulate virtual classes, to give a robust way of determining the statistical significance of results. This is described in chapter 5.
- There was a significant gender difference on the RCI, which did not show up in the course assessment. This is discussed in chapter 6.
- Students assessed their own confidence in their responses as they took the RCI. We adapt some analysis methods that were suggested in the literature regarding student confidence in concept inventories, and show new ways that this data can be used to better understand and categorise student responses. This can be found in chapter 6.
- New student misconceptions in special relativity were discovered and explored. These can also be found in chapter 6.

Chapter 2

Background

Through a review of the relevant literature, this chapter will address the questions: What does physics education tell us about student thinking, and what is known about developing and using concept inventories? In chapter 3, we will address the question of what is known about student understanding of special relativity.

The main findings in the literature that will influence our work are the development methodology advocated by Adams and Wieman [12], the use of question pairings suggested by Singh [19], the study of student confidence used by Allen et al. [20] and others, and the analysis approach of Heller and Huffman, and others [21].

2.1 Review of previous research on teaching physics

2.1.1 Depth of learning

There are many studies in the physics education research (PER) literature detailing the ways in which physics students fall short of the expectations of their teachers. The work of McDermott and Redish [8, 22] provides us with a foothold in the existing literature, by drawing together existing physics education research to generate criteria for student learning objectives. I'll reproduce those that are relevant to us:

- Facility in solving standard quantitative problems is not an adequate criterion for functional understanding[1]. *Questions that require qualitative reasoning and verbal explanation are essential.*

- A coherent conceptual framework is not typically an outcome of traditional instruction. *Students need to participate in the process of constructing qualitative models that can help them understand relationships and differences among concepts.*

- Certain conceptual difficulties are not overcome by traditional instruction. *Persistent conceptual difficulties must be explicitly addressed by multiple challenges in different contexts.*

All of these conclusions were drawn from the results of research conducted in introductory calculus-based and algebra-based physics courses at North American universities in the

[1] We'll define *functional understanding* to mean skills and knowledge that can be used scientifically - that is, to make testable predictions about phenomena. For an introductory physics student, this may mean simply being able to reliably solve a problem that they haven't seen before. In contrast, *nominal understanding* comprises merely the ability to memorise a definition or formula, and apply it narrow, almost "pre-programmed situations" [6].

1980s and 1990s. We argue that these results are relevant to Australian universities in the 21st century; the issue of deep learning at universities has been discussed at length in Ramsden's book, which, though drawing on international research, is rooted in the context of Australian universities [23]. Ramsden's book is about learning and teaching in universities in general, and creates an important link between the PER focus of this work and the broader educational reform context. In particular, Ramsden emphasises that traditional teaching and assessment methods often instill a shallow approach to learning in the student. This conclusion is reflected in the PER results of Reif. et al [24], in which they argue that much of student knowledge obtained in physics courses is "nominal, not functional".

2.1.2 Interpretation of learning problems

A key goal of physics education research is to attempt to understand student thinking, and identify the areas in which it is functional and those in which it is not. Redish et al. [6] put forward a framework for making sense of student learning, which describes properties of "mental models" that students may use. This challenged the notion that students' conceptions in physics were organised into a coherent and self-consistent logical framework. This framework forms the basis of our treatment of student learning, so we'll summarise its principles briefly:

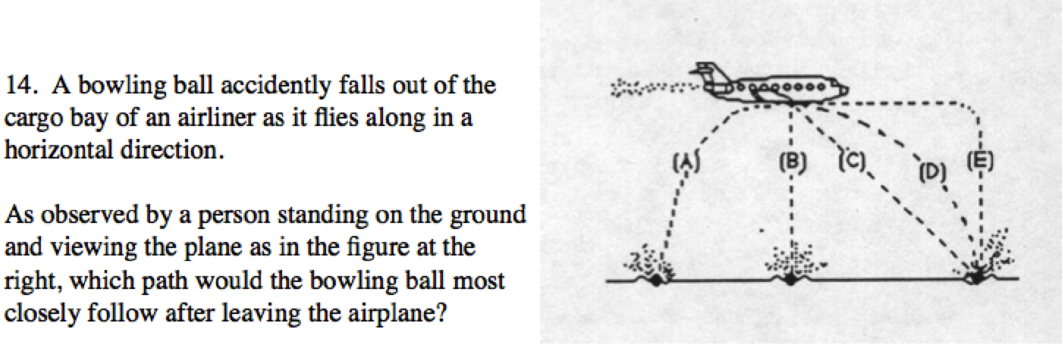
- Mental models typically contain images and procedures.
- Models may be incomplete and/or not self-consistent.
- Elements of mental models don't have firm boundaries. Similar elements may get confused.
- Mental models tend to minimise expenditure of mental energy.

An important corollary of these principles is the idea that most traditional testing fails to test the student's mental model, and this has been observed in several experiments [22].

2.1.3 Instructional strategies, and how to evaluate them

Having first delivered a critique of traditional teaching methods, PER results then motivated and informed the development and dissemination of research-based instructional strategies, particularly in the United States [25]. While many of these strategies are well-grounded in cognitive theory and in experiments, a calibrated measurement instrument is needed to objectively compare the effectiveness of different strategies.

This is the domain of the concept inventory. Since the introduction of the Force Concept Inventory (FCI) in 1992, "a number of these instruments have been developed to measure student learning of science at the university level in a systematic way, and these are having a growing impact on teaching and learning" [12]. Concept inventories have been developed and used in other areas of physics, including electromagnetism [18, 26, 27] and quantum mechanics [28, 29], and in other areas of science, including statistics [30], genetics [31], and molecular biology [32]. One has yet to be developed for special relativity.



14. A bowling ball accidently falls out of the cargo bay of an airliner as it flies along in a horizontal direction.

As observed by a person standing on the ground and viewing the plane as in the figure at the right, which path would the bowling ball most closely follow after leaving the airplane?

Figure 2.1: An example multiple choice question from the Force Concept Inventory.

2.2 Concept Inventories

The concept inventory is a specific type of formative assessment. Its purpose is to measure student learning as a result of instruction, and indicate areas in which instruction has been particularly effective or ineffective for the class as a whole. The main goal is formative assessment of teaching, not summative assessment of student ability. The original Force Concept Inventory [10] is the prototype concept inventory, on which most concept inventories are modeled. Some have made departures from its conceptual emphasis, testing technical skills and problem solving to greater extents [18]. All concept inventories follow the same rough format though - a multiple choice ("forced answer") test of around 20-30 multiple choice questions, which take the student around half an hour to complete.

The FCI consists of a set of 30 qualitative multiple choice questions, designed to probe student understanding of Newtonian mechanics. Answer alternatives were carefully researched so as to represent common student misconceptions, giving the FCI power as a diagnostic instrument. The questions themselves are, to an expert, straightforward; the assumption of many physics teachers was that their students would score highly on the test. The results were surprising, and uniformly worse than expected - even students that did well in course assessments performed poorly on the FCI. It was previously observed that many students held pre-Newtonian (Aristotlean) conceptions of mechanics before taking first year physics [33], and these FCI results showed that traditional instruction had done very little to change the beliefs held by those students [10].

This raised the general awareness in the teaching community of students' conceptual difficulties in studying mechanics, and the dissemination of the FCI effected a revolution in physics education research in the 1990s. It has since been used in hundreds of institutions on tens of thousands of students, including at the Australian National University (ANU). When administered as a pre-test (before instruction) and a post-test (after instruction) on the same student body, the effectiveness of the instruction can, in principle, be measured.

To this end, a large collection of FCI and Mechanics Diagnostic data was used in a comparison[2] of research-based courses and traditional courses (see figure 2.2). Concept

[2] Traditional courses and low-scoring interactive engagement courses are under-represented in the sample, an effect which the author attributes to a reluctance on the part of teachers to report scores whose gains are "embarrassingly minimal". The author argues that this does not significantly bias the results,

inventories in this mould (see figure 2.1) have since been constructed in other areas of physics such as electromagnetism and thermodynamics, and in other areas of science such as chemistry and biology, and there is a growing reliance on concept inventories and related instruments as education research tools. For a detailed bibliography of concept inventories and their impact on science education, see [34].

Development methodology

The process of developing concept inventories has been well documented [12, 30, 26]. Adams and Wieman recommend that in selecting which concepts to include on the inventory, the experimenter should choose those that are (1) widely taught, (2) considered to be valuable, and (3) have been shown to cause students difficulties. Adams and Wieman argue against the inclusion of two-tier questions such as "fact followed by reason" because "brevity and ease of interpretation are more important than detailed characterisation of student learning." They add that teachers should avoid the temptation to make questions more technical and precise than is necessary to probe student thinking.

Adams and Wieman emphasise the use of student interviews in concept inventory development and validation, and recommend that this process be spread over a number of years to make finding appropriate participants easier. Ideally, through the interviews and open-ended responses, the spectrum of common misconceptions will be determined, and this will provide ample material to serve as distractors (alternative incorrect multiple choice answers). The distractors serve to inform the instructor, too: "The primary challenge in creating good multiple-choice questions is to have incorrect options (distractors) that match student thinking. Typically three to five distracters are offered, although there are exceptions." In the case of the Relativity Concept Inventory, several of our questions fall into the "exceptions" category, as they amount to a true/false question, as no suitable distractors could be invented these questions.

We now turn to the issue of validation, and use previous analyses of the FCI as a model, as the FCI is by far the most used, and studied, concept inventory [34].

Validating a Concept Inventory

Does the Force Concept Inventory test what it purports to test? A number of techniques have been advocated to answer this question. Steinberg et al. [35] wrote exam questions that corresponded to selected FCI questions, and did a study on the correlations between student performance in the FCI (multiple choice) and the open-ended versions of FCI questions. Alarmingly, the connections between FCI questions and the corresponding exam questions, while significant, were not as strong as expected. For example, in relation to a question on Newton's First Law, from a sample of 28 students, 54% answered the FCI correctly, whereas 90% answered the corresponding exam question correctly. Some students that answered correctly on the FCI supplied incorrect reasoning in the exam, and some responses on the exam didn't have a corresponding FCI response. This result is an example of the sensitivity to context that was discussed by Redish et al. [6]; they

as there was no disincentive for traditional courses to report high gains, and there are none of these. In particular, it is telling that no traditional course obtained a $\%\langle\text{Gain}\rangle$ greater than 20, whilst several interactive engagement courses did.

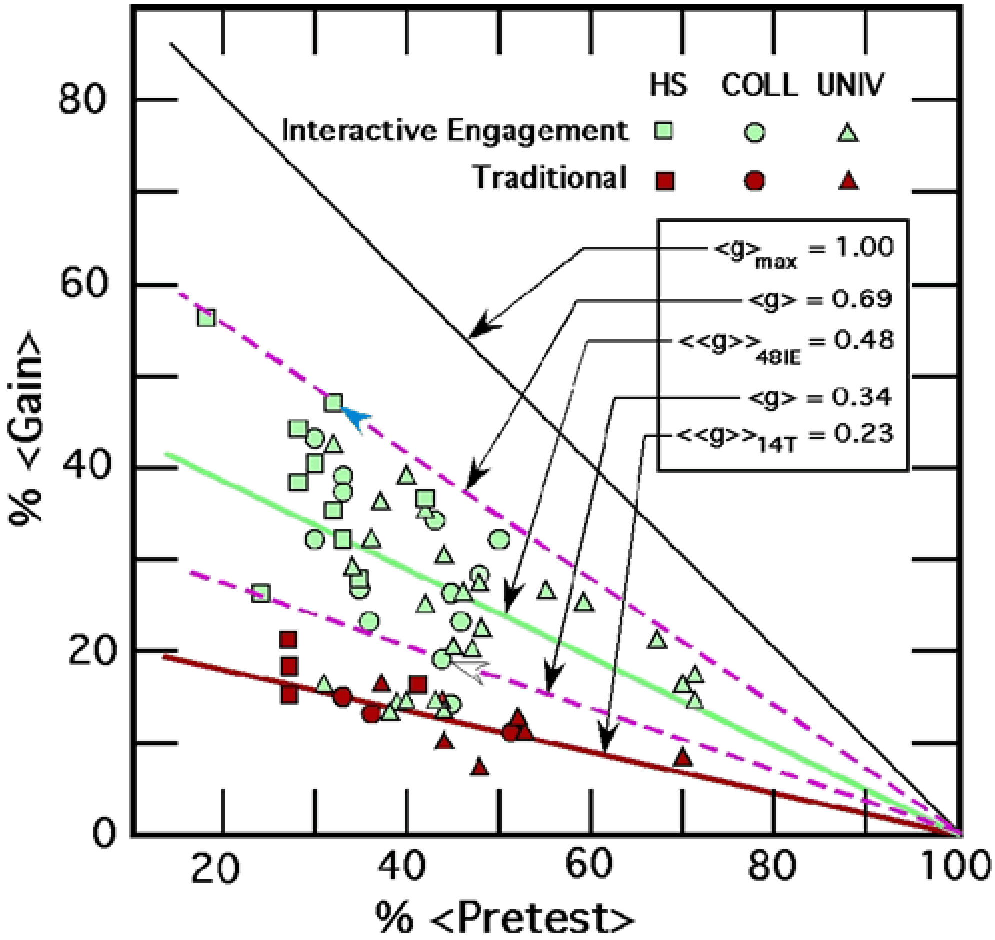


Figure 2.2: $\langle gain \rangle$ vs $\langle pretest \rangle$ scores on the conceptual Mechanics Diagnostic and FCI tests for 62 high school (HS) and first year university (COLL, UNIV) mechanics courses enrolling a total $N = 6542$ students , where $\langle gain \rangle = \langle posttest \rangle - \langle pretest \rangle$, averaged over each class. The diagonal lines are lines of constant normalised gain, $\langle g \rangle = \langle \frac{gain}{1-pretest} \rangle$, which is a better measure of learning than absolute gain (see Section 5.2). The red markers represent traditional courses, and the green markers represent research-based courses. The mean normalised gain for the research based courses is 0.48, more than twice the mean normalised gain for traditional courses (0.23). This result shows research-based courses consistently outperforming traditional courses with respect to student learning gains, and so was a milestone in the physics education literature [11].

suggest student thinking can consist of pieces of knowledge, different parts of which can be expressed at different times or in different situations.

Singh [19] ran a study in which she presented students with a pair of “isomorphic” problems in Newtonian mechanics - these are problems that, to a physicist, involve the same set of concepts, and on a deep level are effectively the same question, but are presented in different contexts - usually a physical system with different “surface features”. This and previous works showed that as a student’s expertise grew, they were more likely to identify the deep connection between the two questions, and answer them in a consistent way. The converse was also found to be the case, and this is consistent with Scherr’s “pieces model” of student understanding [36]. Stewart et al. [37] applied a restricted class of context transformations to a subset of the FCI questions in a study using 647 students at the University of Arkansas, to ascertain whether changing the superficial context of the physical concept had any effect on student performance. In contrast to the result from Steinberg et al. [35], they found the effect of context transformations to be negligible. In this project, we will attempt to make the most of both approaches: we usually address concepts in the RCI with “isomorphic” pairs of questions, and the mid-semester exam was designed to be a quantitative version of a handful of RCI questions.

An important idea put forward by Hestenes is that it should be possible to arrange the FCI questions into groups based on common concepts, and that this grouping would be reflected in student responses . Heller & Huffman [21] challenged this claim, and argued, based on a factor analysis of a large set of FCI data, that students do not have a coherent conceptual framework within which to approach the FCI questions, and so the conceptual coherence that Hestenes expected did not exist. A discussion in the literature ensued [38, 39], and Hestenes subsequently withdrew his suggestion regarding conceptual grouping [40]. This presents us with an important lesson about concept inventories and their interpretation: while it is desirable that our test questions have a one-to-one correspondence with students’ mental models, and thus give us a snapshot of their conceptions, we have to be prepared that this will often not be the case - this is a noisy signal!

The “conceptual coherence” debate also relates to Scherr’s 2007 paper, in which she discusses two distinct models for student thinking: the “pieces” model and the “misconceptions” model [36]. The “misconceptions” model that is favoured by many researchers (that students hold a rigid, coherent, context independent mental model of reality that can mirror historical misconceptions) and the “pieces” model, in which students hold networks of loosely connected, often inconsistent ideas, which may be malleable and context dependent. In this sense, Hestenes et al. were more aligned with this misconceptions model, while Huffman and Heller argued that something like a pieces model is more appropriate for interpreting FCI data. In our investigations, we will attempt to make use of both models in interpreting the data.

2.3 Student confidence

Recently, some concept inventory developers have begun using student self-assessed confidence to supplement the data from the inventory questions themselves, although the use and analysis of this data for validation purposes has not been investigated in detail.

Allen et al. [20] report the use of a Likert-style confidence rating scale in their Statistics

Concept Inventory, and briefly outline possible ways that student confidence data could be related to other properties of the test, such as item difficulty and discrimination. Recently, Siewiorek et al. reported using confidence scales in concept-inventory like instruments, and note a positive relationship between student confidence and test score. They also use confidence scores to differentiate between students with misconceptions (confident and incorrect) and what they term "misunderstandings" (unconfident and incorrect). More recently still (September 2012), Lawrie et al., at the University of Queensland, collected student confidence data for questions drawn from a number of chemical concept inventory tests, and found that students in the mid-range of academic ability were most likely to be over-confident in their responses [41]. A number of these studies also reported gender differences with respect to score, and confidence; these will be reported in section 6.8, where we discuss a possible gender bias in the RCI.

Based on the favourable use of student self-assessed confidence in this work, we included a confidence scale in both pre-test and post-test iterations of the RCI, and extend these previous methodologies relating to confidence and score.

2.4 Justification of my approach

The results reported in the literature affected my design and interpretation of the RCI in the following ways.

Objectives: The desired end-state for students is mastery of the material. This consists of a correct, coherent, and self-consistent conceptual framework, which can be applied to solve problems in different scenarios. The RCI attempts to measure this by avoiding re-use of the "staple" relativity problems, and presenting the student with unfamiliar scenarios.

Context: There was a disagreement in the literature on the importance of context in conceptual questions. Our position is that it is important to control for context when testing conceptual understanding, so we used multiple contexts to test each concept.

Conceptual coherence: This is another contentious issue. We adopt a compromise position; we primarily use the misconceptions model to interpret the data, while recognising that it will be necessary to supplement it in some cases with the pieces model.

Student confidence: Allen et al. collected student confidence data in their Statistics Concept Inventory [20], and some further work has been done in this area. We collect confidence data from our students, and explore the inferences that can be made with it.

Technicality: The approach we have taken in the design of the RCI is more aligned with the FCI of Hestenes et al. [10] than the CSEM of Maloney et al. [18] - that is, we have tried to avoid technical language and any requirement that the students use mathematical formalism to arrive at their answers. A type of test that should be avoided is one that assesses "naming things", rather than applying concepts in situations.

Methodology: As Reif points out, quantification is a major issue for PER [9]. With a large number of uncontrolled variables, a small sample size[3], and only one "shot" at the experiment on each iteration, the number of definitive conclusions that can be drawn from quantitative data will be small. Consequently, we were conservative in the conclusions that we drew from the data.

[3] Although this is a serious problem for this project, it is not always the case. Well-funded PER experiments often run studies over several years and across numerous institutions, and it is not uncommon for these studies to have a sample size well in excess of 1000 (see [42] or [43] for example).

2.5 Conclusion

The previous literature review highlights the importance of concept inventories for physics education research. While concept inventories have been created in a number of important areas of physics, none has been published for special relativity; Gibson's work was a step in that direction, but was incomplete and his results lacked rigour; a critique of his work can be found in section 4.2.2. By developing a Relativity Concept Inventory, we will address a significant deficiency in the tools available to physics educators.

Using a mixed methods approach[4], we build on the work of Scherr and others, with respect to special relativity, and also extend the quantitative techniques used in concept inventories to create a valuable measurement instrument. This is important for PER because, as McDermott and Redish point out, PER suffers by comparison with the more precise physics research epistemologies.

In the next chapter, we present and interpret the previous research on special relativity education, drawing in particular from the seminal work of Rachel Scherr et al. at the University of Washington [16, 44, 45].

[4] Consisting of both qualitative and quantitative techniques.

Chapter 3

Special relativity: conceptions and misconceptions

3.1 Overview

This chapter sets out the conceptual underpinnings of special relativity and known misconceptions found by prior research, and hypothesises the existence of other misconceptions. The chapter forms the framework around which we will design the concept inventory, and, in conjunction with the expert survey in chapter 4, will provide the justification for the content, and conceptual emphasis of the RCI. In particular, we go into detail about the following concepts:

- Inertial reference frame
- The postulates of special relativity
- Time dilation
- Length contraction
- Relativity of simultaneity
- Velocity addition
- Causality
- Consistency
- Mass-energy equivalence

3.2 Inertial reference frame

An inertial reference frame can be defined as a non-accelerating coordinate system with which we can assign time and space coordinates to events; it is a framework within which observers make measurements of physical phenomena. This is a system which allows all observers in a common frame of reference - that is, at rest with respect to each other - to assign the same space and time coordinates to distinct events. This is the so-called “rods and clocks” idea, popularised by Taylor and Wheeler in their book (see figure 3.1). If we allow our grid of imaginary measuring rods and clocks to be arbitrarily fine, then we can define the time of any event as the time we read on a clock directly adjacent to that event, at the same time as we see the event occur. This allows us to account for the delay for the signal to reach us[1], provided our clocks are all synchronised using Einstein's clock synchronisation method [46]. Accounting for light delay in this way is also referred as the "intelligent observer" by Scherr [45]; this simplification is crucial in most treatments of relativity, as it allows us to neglect optics, which have a significant effect in relativistic scenarios.

All inertial frames are equivalent; the only thing distinguishing two frames of reference is their relative velocity. Let's consider two inertial frames S and S', with S' moving at constant velocity v in the x direction with respect to S, and so that their coordinate origins coincide at $t = t' = 0$ (see figure 3.1). Let an event described in S have coordinates (t, x, y, z). Then the same event described in S' has coordinates (t', x', y', z'), given by the transformations in section 3.11.

Misconceptions

It has been shown by Scherr et al. [16, 45] that many student difficulties with special relativity stem from a lack of formal understanding of the reference frame formalism: "*Students at all levels have significant difficulties with the ideas that form the foundations of the concept of a reference frame. In particular, many students do not think of a reference frame as a system of observers that determine the same time for any given event.*" Scherr further proposed that this misunderstanding of reference frames has a severe impact on students' ability to understand the concept of the relativity of simultaneity.

Difficulties with the reference frame concept are not restricted to students studying special relativity; studies by Panse et al. [48] and Ramadas et al. [49] were conducted in courses focused on Galilean relativity. In student responses to questionnaires and in interviews, the authors encountered numerous misconceptions relating to reference frames. Those that are particularly pertinent to this investigation are:

1. The notion that reference frames have finite spatial extent, and that an object can "enter or leave" a reference frame; in this misconception, frames are seen as local and position-dependent, instead of non-local and velocity-dependent.

2. The idea that phenemona depend on how they are viewed, a view that the authors dub "pseudorelativism". This has its relativistic counterpart, in which relativistic effects may be attributed to optical effects, or perception.

3. The idea that some motion is real and some is only apparent.

We will refer to this third misconception as the "absolute reference frame" misconception, because it is consistent with the belief that there is some priveleged reference frame, with respect to which all "real" motion occurs. This interpretation is reinforced in studies by Villani and Pacca [50], in the context of special relativity, and by Saltiel and Malgrange [51], in the context of Newtonian mechanics. The idea that the motion of an object is an intrinsic and absolute, rather than extrinsic and relative, was discovered to be present in 11 year olds, first year undergraduates, and fourth year undergraduates. Villani and

[1] While it is possible to define the time of an event as the time that an observer *sees* the event, without compensating for signal delay, this would make time measurements position-dependent, which is undesirable. Einstein considered and rejected this definition, in favour of the one given above.

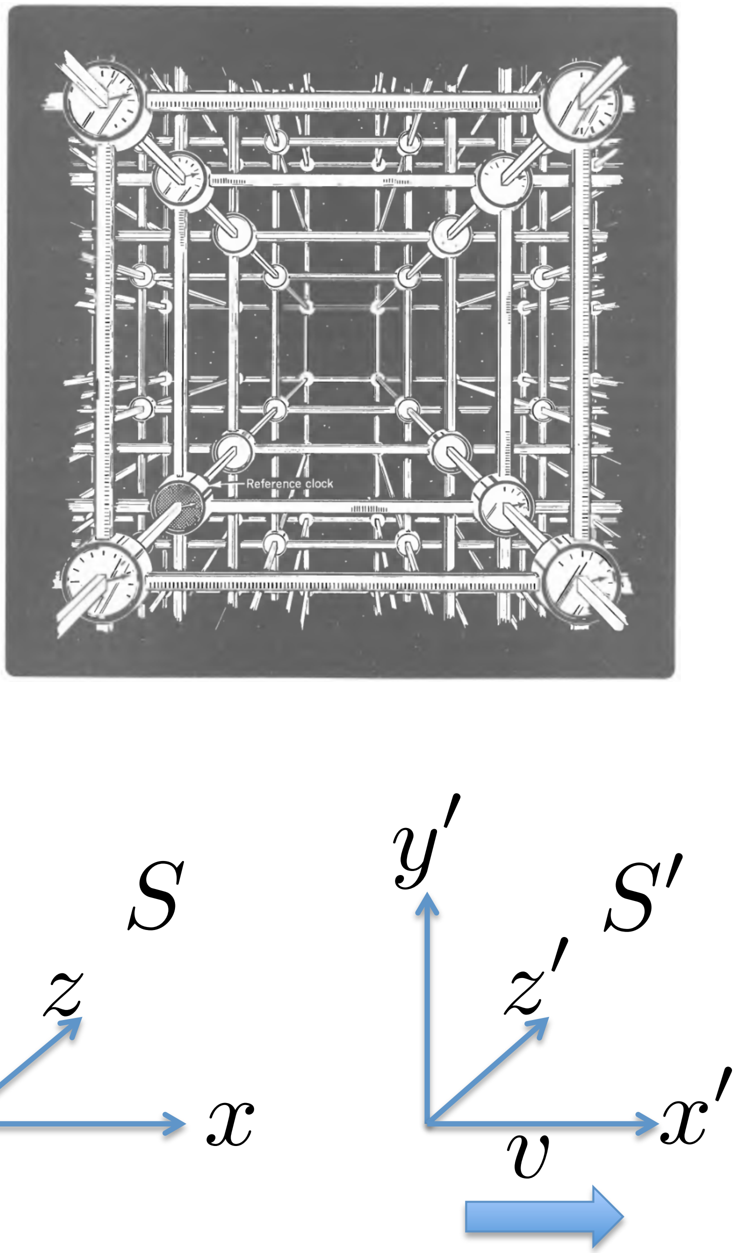


Figure 3.1: Top: The “rods and clocks” picture of a reference frame. Image taken from Taylor and Wheeler’s *Spacetime Physics* [47]. **Bottom:** Two inertial reference frames S and S' in uniform motion with respect to one another.

Pacca concluded from this evidence that "*Concerning the teaching of special relativity ... it is not realistic to assume that students have fully understood Galilean relativity*" and recommend that a course in special relativity should begin by giving students a firm grounding in Galilean relativity: "*... to build up in the student's mind a Galilean intuition which will at least in part liberate students from the idea of an absolute reference frame.*"

Based on the findings and recommendations of this research, we may conclude that an understanding of Galilean relativity, and reference frames in particular, is critical to a good conceptual understanding of special relativity. This view is particularly advocated in David Mermin's textbook *It's About Time* [52]. These recommendations were implemented in the PHYS1201 curriculum. Galilean relativity is a stepping stone to special relativity, and the inclusion of the Galilean transformations in the pedagogy is a necessary precursor to the Lorentz transformations. However, no explicitly Galilean (i.e. $c = \infty$) scenarios were included in the RCI, because this would diminish the overall coherence of the instrument; the instrument is designed to test conceptually where students end up, and not how they got there. Inertial reference frames are tested in two RCI questions, the results of which can be found in chapter 6.

3.3 The postulates of special relativity

In the first of his two 1905 papers on relativity, Einstein presents us with two postulates from which he derives the theory of special relativity [46]:

- The laws of physics are the same in all inertial reference frames.
- The speed of light in a vacuum is independent of the motion of the emitter or the receiver.

The first is Galileo's relativity principle but in stronger form - applying to all laws of physics, and in particular, to Maxwell's theory of electromagnetism. The second postulate (constancy of the speed of light) is derivable from the first postulate, and Einstein makes the subsidiary nature of the second postulate clear in his second 1905 paper [53]:

> *"The laws by which the states of physical systems alter are independent of the alternative, to which of two systems of coordinates, in uniform motion of parallel translation relatively to each other, these alterations of state are referred (principle of relativity)."*

and follows this with the footnote:

> *"The principle of the constancy of the velocity of light is of course contained in Maxwell's equations."*

A minority of the experts we surveyed argued for the inclusion of the second postulate under the first postulate, for this same reason (discussed in section 4.2.1). However, for the purposes of the first year course, we refer to the constancy of the speed of light as the second postulate, in accordance with Einstein's first paper, and treat it as an independent concept.

Misconceptions

Student misconceptions with respect to the first and second postulate have not yet been investigated in detail. Based on the outcomes of previous work, and in particular results from Gibson's inventory (see Section 4.2.2), suggest that the second postulate concept is the among the easiest for students to understand and apply, despite its apparent strangeness. The conclusions made by Posner et al. [54], based on their study, agree with this assessment: *"Constructing a coherent representation of the theory's two postulates individually is not particularly problematic. One can imagine a state of affairs in which each in turn is true, although the more one accepts Newtonian mechanics the harder it will be to imagine a world in which the postulate about constancy of the speed of light is true ... The intelligibility of the theory as a whole, however, is a different matter."* The first postulate, and to a lesser extent, the second postulate, were strongly supported by experts in the expert survey (section 4.2.1), so they are well represented in the RCI.

3.4 Time dilation

In our exposition of the three main kinematic relativistic effects (time dilation, length contraction, and the relativity of simultaneity), we used the light-clock thought experiment to derive and explain the details, following the deductive approach of Einstein. We use the relativity principle to argue that all results that apply to light clocks are general, and apply to all clocks and measuring rods.

The light clock is a box with an emitter and receiver at one end, and a mirror at the other. The light clock "ticks" by emitting a pulse of light at one end, which then is reflected off the mirror at the other end and is received at the detector, which for all intents and purposes is at the same position as the emitter, relative to the clock (see figure 3.2). The time of this round trip constitutes one tick of the clock.

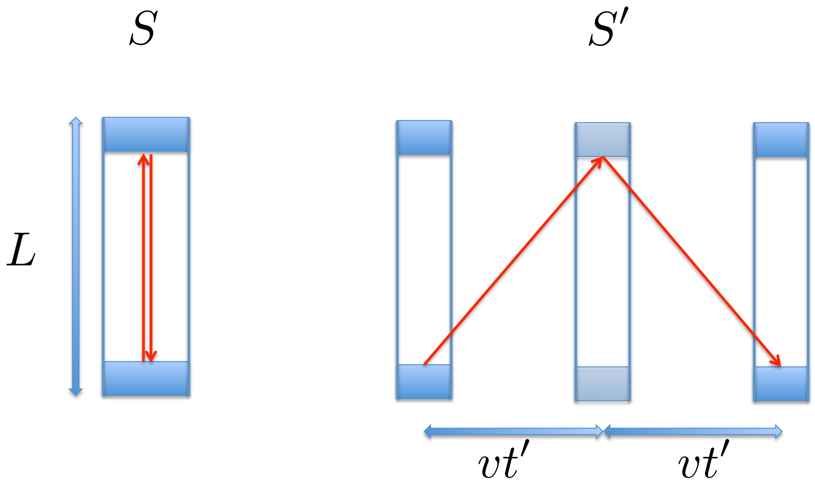


Figure 3.2: A light clock, in the S frame (left), and the S' frame (right).

Let S be the rest frame of the light clock and let the light clock have a rest length L.

The "half-tick" of the clock is given by the time it takes for the light pulse emitted at one end of the clock to hit the mirror at the other end of the clock. In the rest frame of the clock, one half tick takes $t = \frac{L}{c}$ seconds.

Consider a light clock of rest length L oriented along the y axis, and stationary in a reference frame S (the clock frame). Let S' be a reference frame (the lab frame) moving at velocity $-v$ in the x direction relative to S (Equivalently, the clock frame S is moving along the x axis of the lab frame S' at velocity v). In the lab frame S', the path of the light pulse is longer[2] than the path in the clock frame S, since the mirror and light pulse are moving to the right in this frame. Invoking the second postulate and analysing the geometry of the situation, the time for one half-tick t' satisfies:

$$L^2 + \left(vt'\right)^2 = \left(ct'\right)^2 \tag{3.1}$$

From the analysis in the rest frame of the clock, we have the relation $L = ct$. Substituting into equation 3.1 we have:

$$(ct)^2 + \left(vt'\right)^2 = \left(ct'\right)^2 \tag{3.2}$$

Rearranging:

$$t' = \frac{t}{\sqrt{1 - \frac{v^2}{c^2}}} \tag{3.3}$$

This is the familiar "time dilation" result.

Misconceptions

Issues with students indiscriminantly applying the time dilation formula have been investigated by Scherr [45] and Gibson [17]. Time dilation is a special case of the Lorentz transformation for time: the time dilation result is only valid when $\Delta x = 0$ in the rest frame of the object being observed, and length contraction is only valid when $\Delta t = 0$ in the rest frame of the person doing the observing. I hypothesised that there was a broader misconception relating to time dilation, and that this misconception was related in some way to the "absolute rest frame" misconception, discussed previously. I labelled this misconception the "asymmetric time dilation" misconception. A pair of questions were designed for the RCI to address this hypothesis, the results of which are reported in chapter 6.

3.5 Length contraction

Now consider a light clock of rest length L oriented along the x axis, and stationary in an inertial reference frame S (the clock frame). Let S' be a reference frame (the lab frame) moving at velocity $-v$ in the x direction relative to S. Consider now a full tick of the light clock. Let the time of a full tick in S be $\tau = \frac{L}{c}$. Then, by the time dilation result

[2] This presupposes that the length of the light clock in S' is the same as its rest length in S, which is *only* true if it is oriented perpendicular to the direction of motion in S'. We will prove this result below, in the section on length contraction.

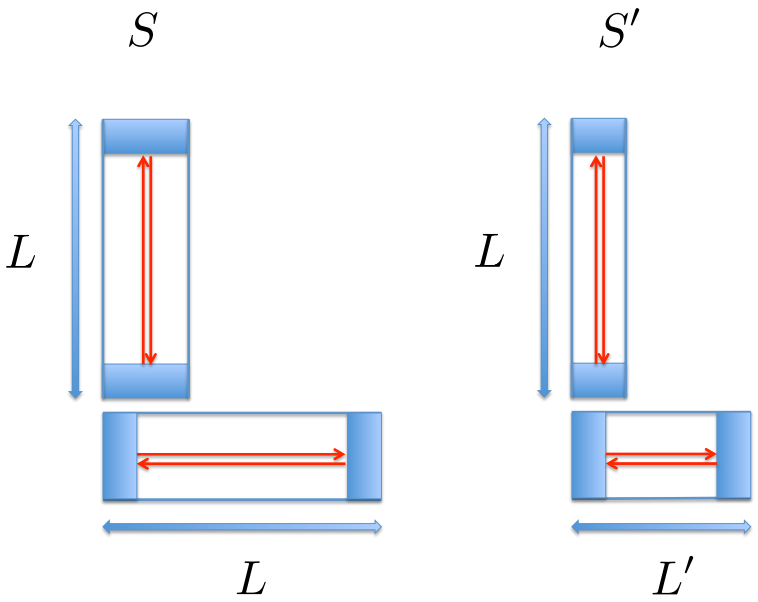


Figure 3.3: Two light clocks, in the S frame (left), and the S' frame (right), showing length contraction.

(equation 3.3), the time of one tick in S' is[3]:

$$\tau' = \frac{\tau}{\sqrt{1-\frac{v^2}{c^2}}} \tag{3.4}$$

Now, let us assume that the length of the clock in S' is given by L', which not necessarily equal to the rest length L. Then, by analysis from S', we have the time taken for the inward (t'_1) and outward (t'_2) half-ticks as:

$$t'_1 = \frac{L'}{c-v} \tag{3.5}$$

$$t'_2 = \frac{L'}{c+v} \tag{3.6}$$

In the rest frame of the light clock, $t_1 + t_2 = \tau$, so $t'_1 + t'_2 = \tau'$ must be the case, by the relativity principle, which gives:

$$L'\left(\frac{1}{c-v} + \frac{1}{c+v}\right) = \frac{\tau}{\sqrt{1-\frac{v^2}{c^2}}} \tag{3.7}$$

Rearranging:

$$L' = L\sqrt{1-\frac{v^2}{c^2}} \tag{3.8}$$

[3] The time dilation result was derived in the context of a vertical clock. This result holds for a horizontal clock, since the two clocks are interchangeable in their rest frame; they tick in synchrony, as long as they are at the same position.

which is the familiar "length contraction" result.

Misconceptions

Issues with students indiscriminantly applying the length contraction formulas have been investigated by Scherr [45] and Gibson [17]. Length contraction is a special case of the Lorentz transformation for space: the length contraction is only valid when $\Delta t = 0$ in the rest frame of the person doing the observing. This distinction, in particular in the case of length contraction, which involves measuring the distance between two simultaneous events, was shown by Scherr to be non-trivial and confusing for many students. Indeed, Roy Kerr commented[4], surprisingly, with respect to length contraction: *"This is far too complicated."* A study on a high-school physics class by diSessa and Levrini [55] "affirms a contention" from Scherr that procedures for measuring length and time intervals need to be constructed to supplement the reference frame formalism.

This led to our proposing for the expert survey a distinct concept, "The operational definition of length and time measurements". The idea was that time and lengths could be specified in terms of events, which connects them to the Lorentz transformations. However, many experts pointed out that this did not add any value to the existing time dilation and length contraction concepts, so it was decided to scrap this concept.

It has also been noted that there is a tendency for students to attribute length contraction and time dilation to optical effects, and to effectively dismiss them as illusions, or distortions of perception. A student interviewed in Posner's study states: *"I see them as changing their length, or changing their time ... I feel they haven't changed, but the way I'm looking at them has changed ... I'm not at all uncomfortable with the idea of foreshortening* [length contraction]. *I do feel it is a perception."*

I also hypothesised a misconception analagous to the asymmetric time dilation misconception, mentioned earlier, and created exam questions to probe this. A further hypothesis was the asymmetric time dilation and asymmetric length contraction misconceptions would be correlated. This investigation is reported on further in chapter 6.

3.6 Relativity of simultaneity

The relativity of simultaneity is the most counter-intuitive and conceptually difficult of the three kinematic relativistic effects, but also arguably the most important, with respect to its consequences. According to Edwin Taylor[5], it is *"... the ambush trap for many students. I have had retired doctors and engineers send me endless papers 'disproving' this one."* I'll continue to use the light clocks to elucidate this concept:

Consider two identical light clocks of rest length L oriented at right angles to each other, with one along the y axis of our coordinate system (Clock A), and one along the x axis (Clock B). Let the event E_A be the half-tick of Clock A, and let the event E_B be the

[4] Comment on the expert survey - see section 4.2.1

[5] Co-author (with John Wheeler) of Spacetime Physics [47]. Commentary on expert survey, 8/7/2012.

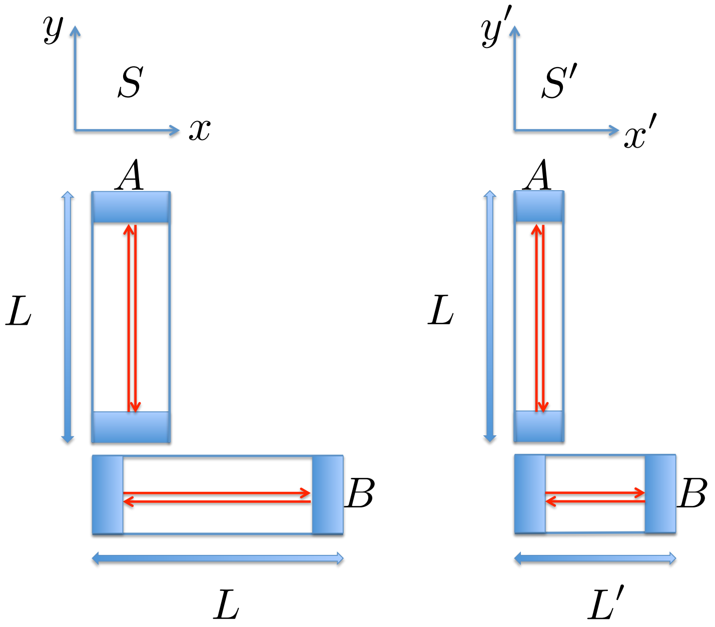


Figure 3.4: Relativity of simultaneity. The half-ticks E_A and E_B occur simultaneously in S, but not in S'.

half-tick of Clock B. In the clock rest frame, the half-tick times for each clock are given by:

$$t_A = \frac{L}{c} \tag{3.9}$$

$$t_B = \frac{L}{c} \tag{3.10}$$

In the clock rest frame, the time interval between the half-tick events E_A and E_B is zero, so the half-ticks are simultaneous in this frame. Let's now analyse the kinematics in a reference frame S' moving at a velocity $-v$ relative to the clocks in the x direction.

The path of the light pulse for a half-tick of Clock A is perpendicular to the direction of motion of the clocks, so the path length is the same in this scenario as in the time dilation scenario:

$$\Delta s' = \frac{L'}{\sqrt{1-\frac{v^2}{c^2}}} \tag{3.11}$$

The path of the light pulse for a half-tick of Clock B is parallel to the direction of motion of the clocks, so the path length in S' is contracted (From Equation):

$$\Delta x' = L\sqrt{1-\frac{v^2}{c^2}} \tag{3.12}$$

Hence for an observer in S', the first half-tick of Clock A takes time t'_A:

$$t'_A = \frac{L}{c\sqrt{1-\frac{v^2}{c^2}}} \tag{3.13}$$

And the first half-tick of Clock B takes time t'_B:

$$t'_B = \frac{L\sqrt{1-\frac{v^2}{c^2}}}{c-v} \tag{3.14}$$

In general, $t'_A \neq t'_B$. The difference between the half-ticks in this frame is:

$$\begin{aligned} \Delta t' &= t'_B - t'_A \\ &= \frac{-\frac{vL}{c^2}}{\sqrt{1-\frac{v^2}{c^2}}} \end{aligned} \tag{3.15}$$

Hence, in the S' frame, the vertical clock half-ticks occur *before* those of the horizontal clock. This is the relativity of simultaneity result: events that are simultaneous in S, and separated along the direction of motion with respect to S' will be non-simultaneous in S'. Moreover, if we change the sign of v, then ordering of E_A and E_B is reversed.

Misconceptions

Student difficulties with the relativity of simultaneity form a centrepiece of the study done by Rachel Scherr and colleagues at the University of Washington [44]. Scherr posed a problem to students involving the eruptions of two distinct volcanos (Mt. Rainier and Mt. Hood), and asks what order the eruptions will occur in for a person flying from one to the other in a high-speed spaceship[6]. Here are some example student responses:

> *"The spaceship is near Rainier, so he gets the signal about the same time Rainier erupts. So the spacecraft pilot would say Rainier erupts before Hood."*
>
> *"Mt. Rainier erupts first because the light from Mt. Hood takes time to reach the spaceship."*

A common issue here is that many students attribute the relativity of simultaneity to signal travel time, or light delay, and thus distort the concept in order to make it compatible with their prior belief in absolute simultaneity. Scherr summarises: "*They [students] often attribute the relativity of simultaneity to the difference in signal travel time for different observers. In this way, they reconcile statements of the relativity of simultaneity with a belief in absolute simultaneity and fail to confront the startling ideas of special relativity.*" In response to this, Scherr produced a modified version of the volcano question, in which it's made explicit that all observers are "intelligent", and compensate for light signal delay. Student responses to this modified question indicated a belief that once signal delay is accounted for, simultaneity will be absolute. This led Scherr to conclude that *"... students held three beliefs that fit together into a coherent, but incorrect, understanding of the nature of spacetime. The three beliefs were that events*

[6] Lecture question 22/8/12B is based on this question - see section D.5.

are simultaneous if an observer receives signals from the events at the same instant, simultaneity is absolute, and every observer constitutes a distinct reference frame."

From this result, it's clear why relativity of simultaneity is so infamously hard: in this case, several misconceptions converge to give a plausible but incorrect understanding. My intention in the RCI is to disentangle these different aspects of the relativity of simultaneity misconception. I supplemented the RCI questions on this with exam questions, to further elucidate the nature of student misconceptions in this area, and to provide a reliable way to identify these misconceptions when they exist. These results are reported in chapter 6.

3.7 Velocity addition

Let a particle have a velocity $u = \frac{dx}{dt}$ in frame S. Then in frame S' (moving at velocity v in the x direction) its velocity is given by:

$$\begin{aligned} u' &= \frac{dx'}{dt'} \\ &= \frac{\gamma\left(dx \pm vdt\right)}{\gamma\left(dt \pm \frac{vdx}{c^2}\right)}, \qquad \text{(from equations 3.22 and 3.23)} \\ &= \frac{dx \pm vdt}{dt \pm \frac{vdx}{c^2}} \end{aligned} \tag{3.16}$$

Dividing top and bottom by dt:

$$u' = \frac{u \pm v}{1 \pm \frac{uv}{c^2}} \tag{3.17}$$

This the velocity addition result. In particular, velocities add such that no object can be observed moving faster than the speed of light, and that adding any velocity u (provided $u < c$) to the speed of light c results in $u' = c$, which amounts to the second postulate.

Misconceptions

There are no documented misconceptions regarding relativistic velocity addition. While Gibson addressed velocity addition in the tutorials that he developed, he didn't probe this concept in his RCI. A common comment from experts was that this concept wasn't as fundamental or as important as the others on the proposed concept list (see section 4.2.1). Nevertheless, in the interests of completeness, the concept was still included on the inventory, although it was not investigated in detail in this project.

3.8 Causality

Special relativity places strong restrictions on whether or not events can be causally connected. To describe the causal structure of spacetime, it is useful to first define the spacetime interval, which is a useful concept, but too abstract to be included on our proposed concept list, for practical purposes.

Using the interval and its invariance as a starting point for elucidating relativity was first introduced by Taylor & Wheeler's *Spacetime Physics* [47]; this approach has since been popular with many special relativity educators[7]. The space-time interval describes

[7] See Section 4.2.1.

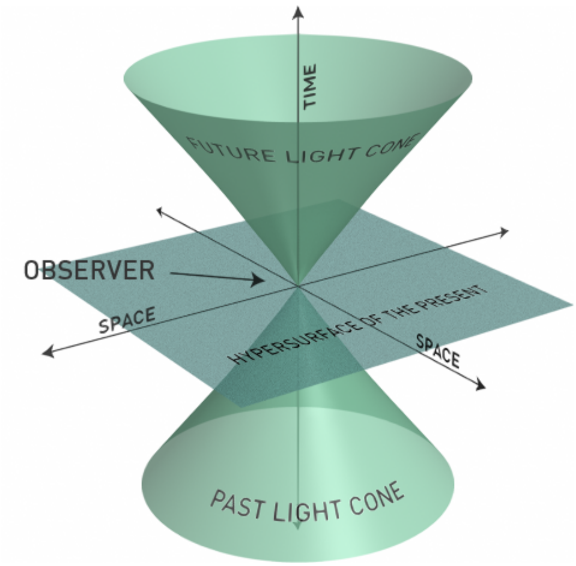


Figure 3.5: Representation of the future- and past-light cones for an event. One spatial dimension has been suppressed. Any events outside these light cones can have no causal connection to the event at the origin, and so their ordering is frame-dependent. Image sourced from Wikipedia.org (GNU license).

the geometry of Minkowski spacetime, and is also known as its metric[8]:

$$\Delta s^2 \quad = \quad \Delta t^2 - \Delta x^2 - \Delta y^2 - \Delta z^2$$

What makes the spacetime interval useful is the fact that it is Lorentz invariant:

$$\Delta s'^2 = \Delta s^2 \tag{3.18}$$

We define two events to be *space-like separated* if $\Delta s^2 < 0$, *time-like separated* if $\Delta s^2 > 0$, and *light-like separated* (or *null*) if $\Delta s^2 = 0$.

This highlights the causal structure of the Minkowski spacetime: two time-like or light-like separated events may have a causal link between them, because causal effects must propagate at a speed less than or equal to the speed of light. This means that their ordering is fixed for all inertial reference frames. Pairs of space-like events, on the other hand, can have no causal connection, and so their ordering is not invariant under Lorentz transformations. This concept relates closely to the relativity of simultaneity: two events A and B may only be simultaneous in a given reference frame if they are space-like separated, in which case their ordering is not fixed: there are frames in which A precedes B, and there are frames in which B precedes A. Time-like separated events, on the other hand, are “immune” to the relativity of simultaneity, since their order is preserved in all inertial frames.

[8] There is no consensus as to which sign convention to use. Many books use $\Delta s^2 = -\Delta t^2 + \Delta x^2 + \Delta y^2 + \Delta z^2$.

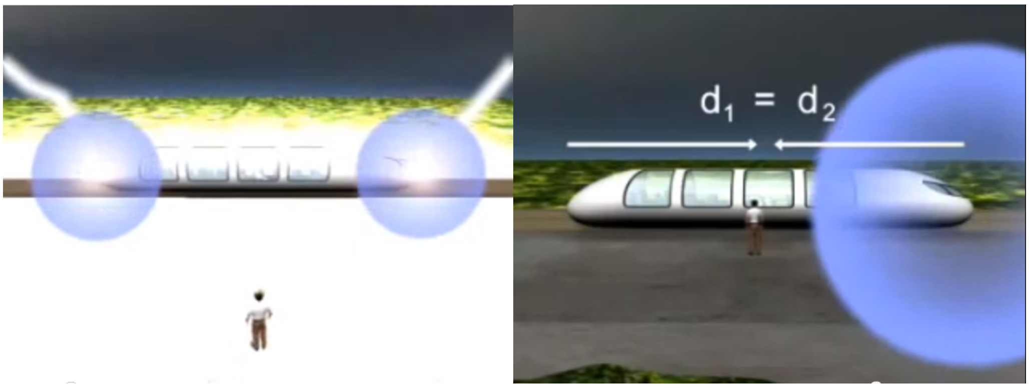


Figure 3.6: Einstein's train scenario, illustrating the relativity of simultaneity. The blue circles represents wave-fronts of light emitted from the two lightning strikes. **Left:** The lightning strikes occur simultaneously as viewed from the platform reference frame. **Right:** The lightning strikes are *not* simultaneous as viewed from the train reference frame. The issue of the order in which the wavefronts reach a passenger, located at the centre of the train, is problematic for many students. Both images are stills from a popular YouTube video that explains the scenario well, which can be accessed at: https://www.youtube.com/watch?v=wteiuxyqtoM.

Misconceptions

There is little data dealing with the issue of causality, although Rachel Scherr mentions it in the context of dealing with the relativity of simultaneity in Einstein's train scenario (see figure 3.6), in which two bolts of lightning hit the front and back of a speeding train, simultaneously in the reference frame of a person standing on a railway platform [45]. A common student misconception is that the platform observer will reason that the train observer receives the light from the wavefronts at different times, while the train observer receives the light simultaneously - a conclusion that violates causality:

> Many treatments of the train paradox devote little attention to the transition from the ground frame to the train frame. Our interactions with students ... indicate that this sequence of reasoning is highly nontrivial for students. Many students fail to recognize that events with a possible causal relationship in one frame must have a possible causal relationship in all frames. In particular, they fail to recognize that events that occur at the same location in one frame in a certain time order must occur in that same time order in all reference frames ... the majority of students are quite ready to ignore requirements of causality in order to retain their incorrect belief that simultaneity is absolute.

This example shows the close relationship between the causality concept and the relativity of simultaneity. For this reason, it was included in the proposed concept list. Expert commentary on this concept can be found in section 4.2.1, and RCI results in chapter 6.

3.9 "Consistency" or, the independence of events from frames of reference

In response to student difficulties with Einstein's train paradox mentioned above, Scherr developed an alternative version of the scenario, in which a tape player is set up at the centre of the train, such that if the two wavefronts originating from the lightning strikes hit it at different times, it will play a segment of a Beethoven symphony, and if the wavefronts hit it at the same time, it will not. If, in the thought experiment, the train is then stopped and the tape player is brought out so the two observers can compare their observations, consistency requires that if the tape has unwound "in the train's reference frame", then it has unwound in all refrence frames. This is the requirement of consistency. Some students were prepared to dispense with this requirement, in the face of all the weirdness: *"Although some students realise that if the music plays in the ground frame, it must do so in any frame, many claim that the music plays in the ground frame but not the train frame."* Scherr doesn't make the distinction between consistency and causality, instead collectively referring to them both as causality. I argue that they are distinct concepts: the requirement that events be independent of frames of reference is different from the requirement that the ordering of all time-like pairs of events is frame-invariant. For this reason, I proposed the concept on the expert survey, and it received favourable enough comments to make it into the RCI pre-test. However, the question wasn't successful, so it was dropped for the post-test. Further discussion of this concept can be found in section 4.2.1 and in chapter 6.

3.10 Mass-energy equivalence

It can be shown that energy has inertia [53]. This mass-energy equivalence can be expressed as:

$$E = \gamma m_0 c^2 \tag{3.19}$$

where m_0 is the rest mass of an object. This can be decomposed into kinetic energy:

$$E_K = (\gamma - 1)\, m_0 c^2 \tag{3.20}$$

and the rest energy:

$$E_0 = m_0 c^2 \tag{3.21}$$

Misconceptions

No misconceptions have yet been documented in the PER literature regarding mass-energy equivalence. This is surprising, given the popular fame of equation 3.21! This concept was included on the RCI, with interesting results (see chapter 6).

3.11 Other concepts

Below I describe other basic concepts that are part of many treatments of special relativity, but which are not included in the RCI, for various reasons.

Lorentz transformations

The Lorentz transformations are the fundamental building blocks of relativity; all of the previous relativistic results may be derived from the Lorentz transformations:

$$t' = \gamma\left(t - \frac{vx}{c^2}\right) \tag{3.22}$$

$$x' = \gamma(x - vt) \tag{3.23}$$

$$y' = y \tag{3.24}$$

$$z' = z \tag{3.25}$$

where in the above:

$$\gamma = \frac{1}{\sqrt{1-\frac{v^2}{c^2}}} \tag{3.26}$$

The Lorentz transformations are a property of the geometry of flat spacetime, and are derived from four assumptions: isotropy, homogenity, and the two postulates of special relativity [56]. They are indispensable for a course in special relativity, but they fall into the category of mathematical formalism, and do not lend themselves well to a qualitative concept inventory, so they are not included on the RCI.

Relativistic Doppler effect

Some of the misconceptions already mentioned have involved students being preoccupied with optics, and light signal delay in particular. In a sense, they are *right* to be so preoccupied with optics, because they are the dominant effect, as the size of these effects are of order $\mathcal{O}\left(\frac{v}{c}\right)$, while the kinematic relativistic effects are of order $\mathcal{O}\left(\frac{v^2}{c^2}\right)$:

$$f' = f\frac{\sqrt{1-v^2}}{1 - v\cos\theta'} \tag{3.27}$$

where f is the frequency of emissions (or frequency of light) emitted, and f' is the frequency received, and θ' is the angle between the source and the receiver, in the receiver's frame. Though this concept does form a part the Real Time Relativity laboratory exercises (see section 6.4), this concept is not covered in detail in the course. This won't be the case for all courses: Gibson's RCI has two questions on the relativistic Doppler effect. This is a possible area for the RCI to be extended, although one of our design principles was to minimise length - this is most easily achieved by excluding subsidiary concepts (see chapter 4).

Relativistic mass

One may choose to define a "relativistic mass" for an object:

$$m = \gamma m_0 \tag{3.28}$$

An argument for using this notation is that it provides intuition for why it's hard to accelerate objects to relativistic speeds, and impossible to exceed the speed of light: the object's inertia increases. But this is not true in the rest frame of the object itself, in which $\gamma = 1$; the first postulate dictates that even a spaceship travelling at $0.99c$ with respect to the Earth will not detect any changes to their inertial mass (this in fact forms the basis for RCI question #18, a first postulate question - see Appendix A).

It has been argued by some physicists that not only is the concept of relativistic mass dated and not useful, but that it is in fact “pedagogically suspect” [57], “a historical artifact” [58], and, regarding its continued use in many textbooks: “it is our duty ... to stop this process” [58]. More recently, it has been treated as a misconception in a PER study in Turkey [59]. The issue is not dealt with explicitly in the RCI, since many treatments - particularly at the introductory level - deal with relativistic mass; it is still a controversial topic among relativists.

3.12 Conclusion

Relativity is hard to learn and hard to teach. By presenting the key concepts of special relativity, and analysing misconceptions applying to each, this chapter forms the starting point for the development of the Relativity Concept Inventory. The next chapter brings to bear the insights of international special relativity experts and educators, which forms the next stage in the development process.

Chapter 4

Inventory development

Adams and Wieman present a robust concept inventory development procedure, but suggest that it requires several years and many iterations to create a final product [12]. They recommend that interviews of students doing the course be used to feed into the design of preliminary inventory questions, which are then administered to students in the next offering of the course (usually the following year). This forms the basis of their iterative process of development. This time was not available to us, and this placed constraints on our development procedure. In particular, we reduced the role of interviews, and instead used mainly exam questions and lecture questions to probe student thinking. The development procedure followed for this thesis is a modified version of the above version:

1. With a review of the relativity education literature, identify what is known about student difficulties in special relativity.
2. Critique the previous RCI attempt by Gibson [17], and draw from his results.
3. Establish topics that are important to teachers using an expert survey.
4. Create open-ended questions to probe aspects of student thinking in test form, for use in tandem with the concept inventory.
5. Create a concept inventory test that measures student thinking.
6. Test it on a small scale with selected student "think-aloud" interviews.
7. Administer to the class as a pre-test and run statistical tests on the results.
8. Use pre-test data to create another iteration of the RCI, and administer this as a post-test.
9. Use post-test data and open-ended question data to create final iteration of the RCI.

4.1 Desirable properties for our concept inventory

These are the principles with which the inventory was designed. We determined that it should:

- Be easy to administer and grade.
- Test concepts that are valued by educators worldwide, so that it is useful at different institutions.

- Have a clear correspondence between the test items and individual concepts.
- Be appropriate to administer as both pre-test and post-test. There are no quantitative or technical questions - the emphasis is on making predictions in simple scenarios, in the style of the Force Concept Inventory.

4.2 Pre-test iteration

4.2.1 Expert survey

To determine some consensus among experts about which concepts would be appropriate to include on the concept inventory, we sent out a survey[1] to content experts in special and general relativity, and educators of special relativity at universities, in Australia and overseas (see Appendix B for the full survey). A total of 31 responses were received, from a diverse group. The experts were asked to respond to our proposed list of concepts with one of three choices (Agree, Neutral, Disagree), and were encouraged to include comments and suggestions to each of their responses.

Results

Below is a brief summary of the relevant results, the conclusions that were drew from them, and how they affected the RCI. Some quotes from the expert comments are also included in appendix D.1.

- Experts were unanimous in the inclusion of the first postulate, and almost unanimous on the relativity of simultaneity (30 out of 31 experts). This is unsurprising, as these are the fundamental underpinning, and the most difficult consequence of the theory, respectively. These concepts are well-represented on the RCI, with four questions each.
- The inclusion of the second postulate was mostly agreed upon (28 out of 31). The only instances of disagreement were related to whether or not the second postulate should be treated as a consequence of the first postulate. We treat it as a separate concept, and gave it two questions on the RCI.
- The "staples" of relativity, time dilation (27 out of 31) and length contraction (25 out of 31), while popular, were not uniformly agreed upon. The lack of agreement was attributed to the "invariant-centric" approach to teaching that some educators prefer. Time dilation and length contraction are dealt with in most texts and at most institutions, so they are retained, with four and three questions each, respectively.
- There was no obvious consensus as to whether mass-energy equivalence, non-inertial frames, and invariance of the interval should be included. The more prominent experts (to whose opinion we gave more weight), tended to disagree with their inclusion in a concept inventory for use in introductory courses (see appendix D.1).

[1] The survey was set up online, using Survey Monkey (www.surveymonkey.net), and mailed (with permission) to the following three mailing lists: The Australasian Society for General Relativity and Gravitation, the Australian Institute of Physics (Education Group), and the Matter & Interactions *Yahoo!* group.

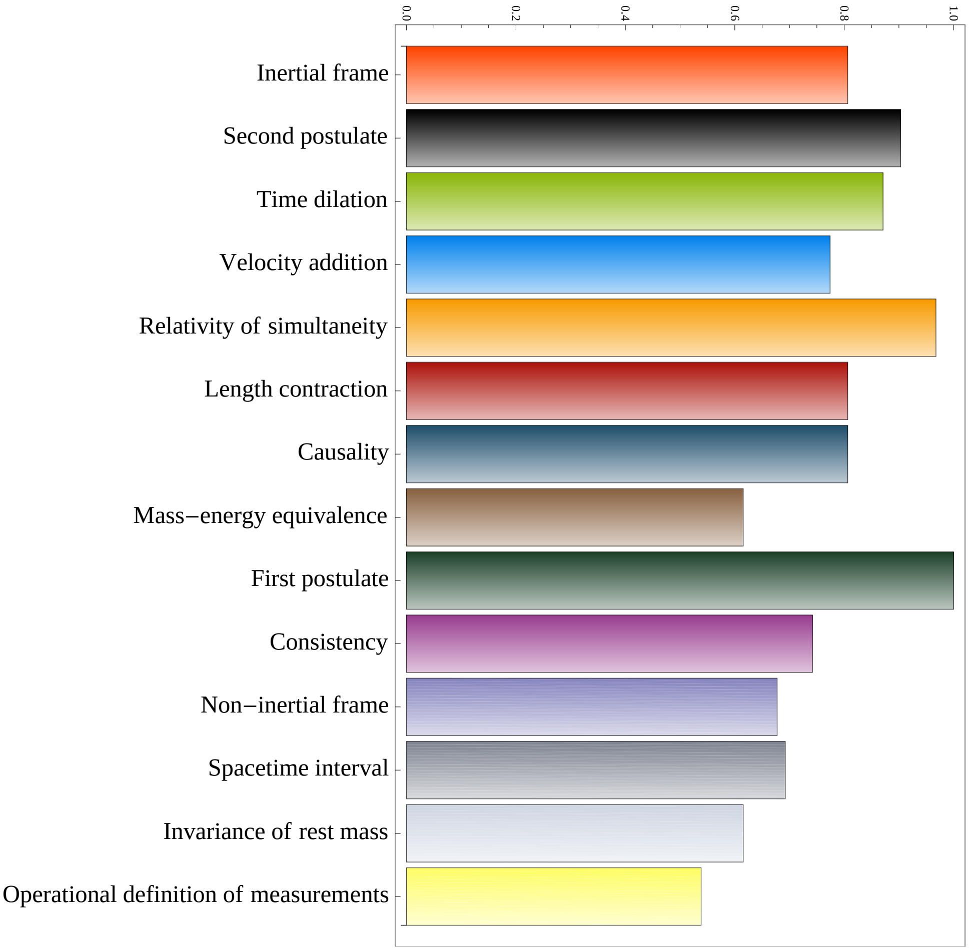


Figure 4.1: Concept survey results. The length of the bars indicate the proportion of experts that agreed with the inclusion of that concept. Concepts are colour-coded in all plots, for ease of interpretation.

Concept	Question(s)
Relativity of simultaneity	1, 2, 3, 4, 5
Second postulate	6, 8
Twin 'paradox'	9
Doppler effect	10, 11
Length contraction	12, 14, 16
Time dilation	13, 15

Table 4.1: Concepts covered in Gibson's RCI.

Mass-energy equivalence was included on pre- and post-test, with one question, as an experiment.

- Velocity addition, causality, and consistency were all marginal cases in which expert votes and commentary were mixed, but these were included to give the RCI a broad scope, with the option of narrowing it down later.

4.2.2 Gibson's RCI attempt

A review and critique of Gibson's previous attempt at creating an RCI provided some input into our development process. Gibson's results were inconclusive, and his development approach was not well defined - in particular, his particular choice of concepts was not justified. We summarise his work here for completeness. No publications ensued from Gibson's investigation, so we use the results included in his thesis [17].

Gibson's RCI has 16 questions, and covers six concepts (see table 4.1). An important omission is the first postulate, which our experts unanimously agreed should be covered in the RCI.

Below, we summarise the contents of Gibson's RCI attempt, and its relevance to our RCI. Many of his questions suffered from important design issues, and so much of his work is not suitable to be built on - see figure 4.2, for example.

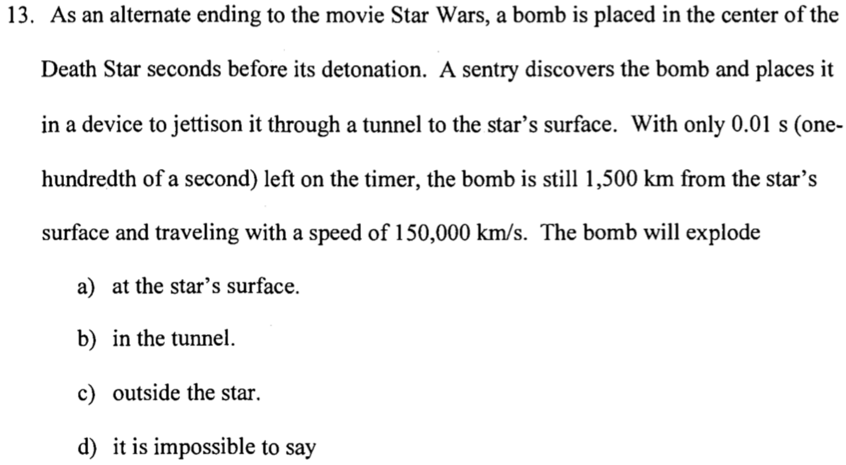
- Gibson's relativity of simultaneity questions are variations of Scherr's "volcanoes" scenario [44]; we used this scenario in a lecture question, for pedagogical purposes, so we avoided it in the RCI.

- The scenarios used in Gibson's second postulate questions are reasonably simple and effective; one of these was streamlined and adapted for our RCI.

- Gibson includes a question explicitly about the 'twin paradox'. Although we didn't poll experts on whether or not to include the 'twin paradox' as a concept, some mentioned that it was not fundamental, and that dealing with paradoxes on the assessment could be counter-productive.

- We decided not to include the Doppler effect, as it is not as basic as the other concepts, and it is not taught in detail in our course.

- Gibson's length contraction and time dilation questions are generally over-complicated and unclear, and were generally not useful for the RCI (for example, question 13 - see figure 4.2).



13. As an alternate ending to the movie Star Wars, a bomb is placed in the center of the Death Star seconds before its detonation. A sentry discovers the bomb and places it in a device to jettison it through a tunnel to the star's surface. With only 0.01 s (one-hundredth of a second) left on the timer, the bomb is still 1,500 km from the star's surface and traveling with a speed of 150,000 km/s. The bomb will explode

a) at the star's surface.

b) in the tunnel.

c) outside the star.

d) it is impossible to say

Figure 4.2: An example question from Gibson's RCI attempt. This question has numerous serious issues, and in my opinion, serves as a guide of what *not* to do when designing concept inventory questions. Aside from alienating students that are not familiar with the Star Wars films (!), the question is unclear as to which reference frame these time and length measurements are made in, which affects whether the correct answer is *(a)* or *(c)*. It also has an uninformative and superfluous distractor in *(d) "it is impossible to say"*. In addition, the inclusion of numerical information is an unnecessary and potentially confounding factor.

4.2.3 Student interviews

Time constraints made it unfeasible to interview PHYS1201 students, so the role of interviews in the development process was limited. I interviewed three senior undergraduates prior to administering the pre-test RCI to the first year class. The purpose of this interview was as a small "reality check" on the RCI questions - to ensure that they were being interpreted mostly as intended. I did two rounds of interviews on these three students: one "think-aloud" interview, and one "verbal probe" interview, separated by about a week; in both interviews, the students worked through a draft version of the RCI. The three students that participated in these interviews were a select group: they were all mentors in the Peer Assisted Learning (PAL) programme[2]. Both interview protocols are explained in detail by Adams and Wieman [12], but we will briefly summarise them:

Think-aloud protocol

The purpose of the think-aloud interview is to find out what a student would be thinking if they were taking the RCI under test conditions. The goal is to put the student in something like an authentic test-taking situation, and get them into the think-aloud "frame of mind" with minimal input from the interviewer. In particular, it is crucial to not press the student to clarify their thinking, or to explain their choices beyond what they are doing as they think aloud. In this way, we minimise back-action on the student and try not to influence their thinking or performance.

[2] The programme screens its mentors for students that are highly engaged and motivated and have above-average grades - these three are a biased sample of undergraduate physics students.

Verbal probe protocol

This is the more traditional interview style, in which the student is asked to explain their thoughts, and is prompted to clarify their ideas, if required. This is, however, not a conversation, and the interviewer should refrain from correcting the student, or asking leading questions. In our case, the verbal probe protocol was used to clarify student responses in the think-alouds.

The interviews served mainly to check whether there were any questions that completely failed to communicate their point, or that were being misinterpreted in unexpected ways. No questions had this problem; at most, the wording of some questions had to be tweaked as a result of suggestions from the interviewees. We will consider one important example of how the interviews fed, belatedly, into the development process, in the case of question 7, which remained unchanged from pre-test to post-test.

Question 7 was problematic, in that it showed a strong anti-correlation between student performance and student confidence, which is an indicator of either a trick question or strong student misconceptions - this is discussed further in section 6.6. Other results showed that there was something peculiar about this question, and indicated strongly that it was a candidate for being removed from the final iteration of the RCI. Student interviews suggested that the question was in fact doing its job properly, and so provided some justification for its not being removed.

This question involved applying time dilation in an unfamiliar context:

7. It is known that our galaxy is around 100, 000 light-years in diameter. True or false: "Travelling at a constant speed that is less than, but close to, the speed of light, in principle it is possible for a person to cross the galaxy within their lifetime."

   (a) True
   (b) False

Figure 4.3: RCI question 7. Our correct answer is *True.*

Two of the three students interviewed did not consider applying time dilation:

> *"It's 100,000 light years, that's got to be false."*
>
> *"No. Look, purely logical, no, because you can't live for 100,000 years, but for some reason I think you're trying to trick me. We'll just go false on that one."*

The third student took a long time with this question, and eventually got the correct answer, although by analysis from the ship frame, in which the relevant effect is length contraction:

> *"If you're travelling with a certain velocity compared to the galaxy, actually, you get length contraction and so the distance isn't as far. So that means that, assuming you can get as close to the speed of light as you like, then that actually is possible for their lifetime, if they're on the spaceship. That's true."*

Incidentally, this last student response with respect to question 7 suggested that perhaps the question ought to be included in the length contraction grouping as well, although for simplicity's sake, we classified it as a time dilation question. Further interviews would elucidate whether this category agrees with the way students interpret the question.

4.3 Post-test concept list

Having done a literature review, and polled the relevant experts, we have the concept-list used for the post-test RCI. The pre-test RCI concept list was the same, except with one additional concept: consistency. This was removed, based on informal comments from a handful of experts that reviewed the RCI itself. Since most of the analysis in chapter 6 pertains to the post-test, we present the post-test concept list only:

- First postulate: The laws of physics are the same in all inertial reference frames.

- Second postulate: The speed of light in a vacuum is the same in all inertial reference frames.

- Time dilation: The time interval between two 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.

- Length contraction: The length of an object (defined as the space interval between two simultaneous events at either end of the object) is longest in the frame in which the ends of the object are at rest, and is shorter in all other frames.

- Relativity of simultaneity: If two events A and B are space-like separated, then there exist inertial frames in which A precedes B, and there exist frames where B precedes A.

- Inertial reference frame: A coordinate system in which a free particle will move at constant velocity - in particular, the concept that all inertial frames are equivalent.

- Velocity addition: Velocities transform between frames such that no object can be observed travelling faster than the speed of light in a vacuum.

- Events are independent of reference frame: If X happens in one reference frame, then X happens in all reference frames (distinct from the first postulate).

- Causality: If two events are time-like separated, then the ordering of the events is fixed for all inertial reference frames.

- Mass-energy equivalence: Energy has inertia.

In table 4.2, we present the list of concepts we included in the post-test and the questions that we intend them to correspond to. Our hypothesis is that questions that we have grouped together under the same concept will be correlated with one another, and a large part of the analysis in chapter 6 is devoted to determining whether or not this is actually the case.

Concept	**Question(s)**
First postulate	16, 18, 19, 20
Second postulate	3, 4
Time dilation	5, 6, 7, 8
Length contraction	13, 14, 17
Relativity of simultaneity	11, 12, 15, 21
Inertial reference frames	1, 2
Velocity addition	9, 10
Causality	22, 23
Mass-energy equivalence	24

Table 4.2: Concept list, and corresponding questions. This table represents our expectations of the RCI: our intention is that the correlations in the student data recreate these groupings.

Chapter 5

Statistical methodology

In this chapter, we comprehensively describe the statistical techniques used in our analysis. Some of these are standardly used in the concept inventory literature (e.g. point-biserial coefficient, KR-20, and factor analysis), and some are not (the Kolmogorov-Smirnov test, and our own Monte Carlo technique for estimating the statistical significance of item-item correlations).

5.1 Correlation and statistical dependence

There are two kinds of quantitative data that will be analysed in this study:

- Dichotomous data. These are data that are binary: either 1 (correct) or 0 (incorrect). In our study these are individual student responses to RCI questions, exam questions, or the student's gender. For RCI data with N students and M questions, there will be $M \times N$ data points of this type.

- Approximately continuous data. These are data where many values are permissible. In our study these can be: RCI total scores, homework total scores, exam total scores, or Universities Admissions Index. These data are not truly continuous, since they are aggregated dichotomous data; all scores are binned in multiples of $\frac{1}{S}$, where S is the highest possible score, as no partial marks are given. Since S in our case is generally large (≥ 20), we will assume the data are sampled from a continuous distribution, for the purposes of correlation [60].

At many points in our investigation, we will want to measure the degree of association between two different data sets obtained from the study. Since there are two different types of data, there are three possible pairs of data sets, and we will need to be able to calculate the association (correlation) for all three combinations:

1. Two continuous variables: Pearson's product-moment coefficient.

2. A continuous variable and a dichotomous variable: Point-biserial coefficient.

3. Two dichotomous variables: ϕ coefficient.

The point-biserial and ϕ coefficients are both special cases of Pearson's product-moment correlation coefficient[1]. We will define the correlations used for each of the three cases:

[1] The derivation of the point-biserial coefficient and ϕ coefficient from Pearson's coefficient can be found in Chiang's book [60].

1. Pearson's coefficient is defined as:

$$r_{xy} = \frac{Cov\,(X,Y)}{\sqrt{Var\,(X)\,Var\,(Y)}} \tag{5.1}$$

where $Cov\,(X)$ is the covariance:

$$Cov\,(X,Y) = \langle (X - \langle X \rangle)\,(Y - \langle Y \rangle) \rangle \tag{5.2}$$

and $Var\,(X)$ is the variance:

$$Var\,(X) \equiv Cov\,(X,X) = \langle (X - \langle X \rangle)\,(X - \langle X \rangle) \rangle \tag{5.3}$$

Pearson's coefficient takes values in the range $[-1,1]$, and is a measure of the degree to which two random variables X and Y are related. We will use this to calculate the correlation between, for example, RCI scores and exam scores. None of the correlations in our data are large, and we will generally restrict discussions to relationships that are moderately correlated, with $0.3 \leq |r_{XY}| \leq 0.6$.

2. The point-biserial coefficient is defined as (in the context of the correlation between test and item scores):

$$r_{pbc} = \frac{\bar{X}_1 - \bar{X}_0}{\sigma_X} \sqrt{d(1-d)} \tag{5.4}$$

where $\bar{X}_1$ is the mean total score for those who correctly answer an item, $\bar{X}_0$ is the mean total scores for those who incorrectly answer an item, σ_X is the standard deviation of total scores, and d is the item difficulty defined in equation 5.6.

The point-biserial coefficient takes values in the range $[-1,1]$, and is a measure of the degree to which a dichotomous variable (such as gender, or whether or not a test question was answered correctly) is related to a continuous variable (such as total test score). The point biserial coefficient is mostly used as a measure of the the internal coherence of a test: if the point-biserial of a question is high, then it means it correlates well with the rest of the test, and is testing something (a concept or attribute) that is a good predictor of overall test performance. If the point-biserial coefficient of a question is low, this indicates that it is possibly testing something that is not closely related to the content of the rest of the test. One interpretation of a low point-biserial coefficient may indicate that the question is not measuring "relativistic thinking", but an alternative interpretation is that there are merely not enough questions on the test dealing with that particular concept [12].

3. Let X and Y be two distinct questions, whose results are characterised by dichotomous data (1 or 0). For an individual student, four outcomes for the question pair (X,Y) are possible: $(1,1)$, $(1,0)$, $(0,1)$, and $(0,0)$. For a set of N students attempting X and Y, we can then construct a 2×2 "contigency table", in which N_{11}, N_{10}, N_{01}, and N_{00} are the frequencies of each of the four outcomes, and $N_{11} + N_{10} + N_{01} + N_{00} = N$:

	X correct	**Y incorrect**
X correct	$\boldsymbol{N_{11}}$	$\boldsymbol{N_{10}}$
Y incorrect	$\boldsymbol{N_{01}}$	$\boldsymbol{N_{00}}$

The ϕ coefficient is then defined as:

$$\phi_{XY} = \frac{N_{11}N_{00} - N_{10}N_{01}}{\sqrt{(N_{11} + N_{10})(N_{11} + N_{01})(N_{00} + N_{10})(N_{00} + N_{01})}} \tag{5.5}$$

The ϕ correlation has the same range of values, and the same interpretation as the point-biserial and Pearson's r product-moment coefficients. It will be used to calculate the correlations between individual questions on the RCI, and so it is the most useful of the three correlations in our investigation.

5.2 Classical test analysis

The basic purpose of classical test analysis is to estimate certain attributes of questions on a test, and of the test as a whole. These include how well it discriminates between students of different ability, how well items performance correlates with test performance, and how consistent the material covered in the test is. Although there are generally accepted ranges of values for many of these statistics, there are several ways to interpret values given, and it is argued that they should be taken as guidelines only [12].

Properties of test items

Item difficulty d is defined as the proportion of students in the sample that get the question right:

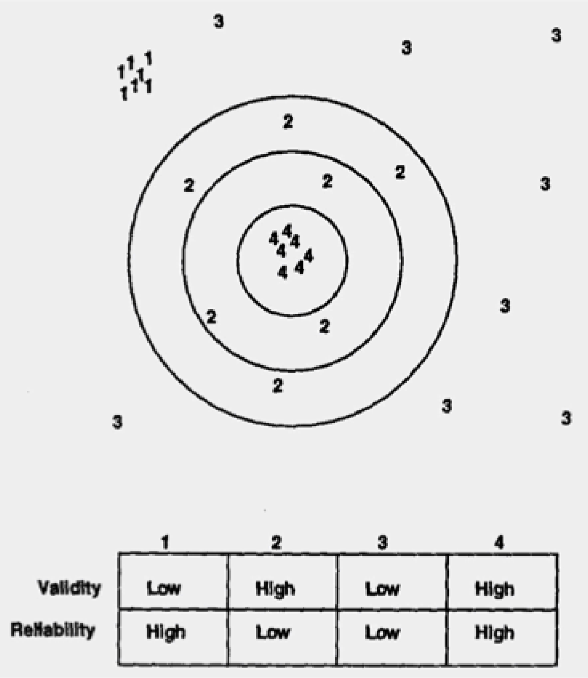
$$d = \frac{n_c}{n} \tag{5.6}$$

where n_c and n are the number of correct responses and the total number of responses for a given question, respectively. The difficulty d is actually a measure of *easiness* rather than difficulty, but we will continue with the standard nomenclature. It has the range $[0, 1]$, and it is desirable that question difficulties lie in the range $[0.3, 0.9]$ according to Ding & Beichner [61]. These bounds are generous, and need to be put into context: an average pre-test score of 0.9 on a concept inventory would be unacceptable (although desirable for a bright class in post-test), while a mean pre-test score of 0.3 would not be out of the ordinary (although in post-test, could be seen as a serious indictment of the standard of instruction).

The item discrimination index is a measure of how well each item differentiates between high-achieving and low-achieving students. It is defined as the difference in the proportion of correct answers between the top quartile and bottom quartile of students:

$$D_i = \frac{4\left(N_{H,i} - N_{L,i}\right)}{N_i} \tag{5.7}$$

where N_H and N_L are the number of students in the top and bottom quartile of the class that answered the question correctly, respectively, and N is the total number of responses to the question. The norm is to define “high” and “low” achieving internally - using total scores on the test only. If this is the measure used, then, according to Adams and Wieman,



	1	2	3	4
Validity	Low	High	Low	High
Reliability	High	Low	Low	High

Figure 5.1: An illustration distinguishing between test validity and external reliability; four different hypothetical tests are shown - the center of the bullseye is the attribute we wish to measure accurately, and repeatably. The center of the target is what each of the tests are attempting to measure. Image sourced from Beichner's paper [62].

not all questions with low item discrimination need be removed or revised, as is normally advised [61, 12] - an example is a question in which nearly every student performed poorly in pre-test, and in which every student performed well in post-test: this test item would have a very low discrimination index, but would nevertheless be an indicator of particularly ineffective teaching if the large gains were not observed.

The distinction between validity and reliability

Test validity is a measure of how well a test measures what it purports to measure, and ensuring the RCI is a valid instrument will form a large part of the analysis in chapter 6. Test reliability comes in two types: internal and external. External reliability is a measure of how accurately a test will reproduce its results, either by testing the same population, or over different populations. Internal reliability is essentially a measure of the internal consistency of a test, based on the relationships between items in a single administration of the test [12]. Figure 5.1 illustrates the difference between validty and reliability.

The internal reliability of a test is usually calculated with the Kuder-Richardson Formula 20 (KR-20). KR-20 is the analogue of Cronbach's α for dichotomous data, and is a widely used measure of internal consistency in psychometric tests [61]:

$$r_{test} = \frac{K}{K-1}\left(1 - \frac{\sum_{i=1}^{K} d_i\,(1-d_i)}{\sigma_x^2}\right) \tag{5.8}$$

A higher mean correlation between test questions will result in a higher KR-20. Low values of the KR-20 would imply that the test items are not well connected, and are testing a set of disparate concepts or skills.

Ferguson's delta is a measure of how broadly distributed the total test scores are, and so is a measure of the discriminatory power of the test as a whole:

$$\delta = \frac{N^2 - \sum_{i=1}^{S} f_i^2}{N^2 - \frac{N^2}{K+1}} \tag{5.9}$$

where N is the total number of students taking the test, K is the number of test items, and f_i is the number of students whose total score is i.

In both the case of Ferguson's δ and the KR-20, higher values are desirable, as they indicate a test with a high consistency (the questions are all testing the same sorts of things), and discrimination (it produces a broad range of scores).

Normalised gain

It is generally accepted in the concept inventory literature that the best measure of student learning is not absolute score, or even absolute gain, but the normalised gain [11]. Normalising the gain corrects for pre-test scores, and provides a good measure of the value-added by instruction. It is defined as:

$$g = \frac{d_{post} - d_{pre}}{1 - d_{pre}} \tag{5.10}$$

where d is the item difficulty defined previously. Normalised gain takes values in the range $[-1, 1]$. If we ever discover a question with a negative normalised gain, from pre-test to post-test, we must either strongly query its validity, or, if it is taken seriously, revise the teaching method relating to the concept that question tests.

5.3 Quantifying uncertainty

All of the measurements used in this study are counts (e.g. how many people answered a question correctly), and since there is no systematic uncertainty in this counting process, the uncertainty in the raw data is zero, and error bars are not appropriate. When we make inferences about the *population*, given an effect in our data, it is often appropriate to use the standard error in the mean to quantify this uncertainty [63]:

$$E_x = \frac{\sigma_x}{\sqrt{N}} \tag{5.11}$$

Since we will rarely make inferences about populations from our data, we will refrain from using error bars in most of our plots and graphs. On the other hand, we will frequently want to know the *statistical significance* of a result, and this is usually quantified with a p-value. In the case of all of our hypothesis tests, the p-value is the probability of our observing, by chance, an effect at least as big as the one observed. We will calculate our p-values with Monte Carlo simulations, and with well-known statistical tests.

5.4 Statistical tests

Pearson's χ^2 test

In comparing two data sets of binned data (score frequencies, in our case), we may calculate the following statistic [63]:

$$X^2 = \sum_i \frac{(R_i - S_i)^2}{R_i + S_i} \tag{5.12}$$

where R_i are the number of events in bin i for the first data set, and S_i are the number of events in the same bin i for the second data set. When the R_i and S_i are large, X^2 has an approximate χ^2 probability distribution. The corresponding p-value is calculated from the chi-square cumulative distribution function:

$$P(a,x) = \frac{\gamma(a,x)}{\Gamma(a)} \qquad (a > 0) \tag{5.13}$$

where $\Gamma(a)$ is the usual Gamma function, that interpolates between the integer factorials:

$$\Gamma(a) = \int_0^\infty e^{-t} t^{a-1} dt \tag{5.14}$$

and $\gamma(a,x)$ is the incomplete Gamma function:

$$\gamma(a,x) = \int_0^x e^{-t} t^{a-1} dt \tag{5.15}$$

For the purposes of our analysis, x is the value of χ^2 calculated with the formula above, and a is the number of of degrees of freedom, in our case the number of bins, which is equal to the number of total scores possible, which is the number of questions plus one. Hence:

$$p_{\chi^2} = 1 - P\left(\frac{n_q + 1}{2}, \frac{X^2}{2}\right) \tag{5.16}$$

A disadvantage of the chi-square test is the requirement for frequencies in every bin to be large, for the chi-square approximation to be valid [64]. For this reason, we will generally use the Kolmogorov-Smirnov test instead.

Kolmogorov-Smirnov test

The Kolmogorov-Smirnov test is a robust, non-parametric test for the equality of two one-dimensional probability distributions. It can be used as a goodness-of-fit test for whether a given data set conforms to a given continuous distribution (one-sample test), or to test the hypothesis that two different data sets were sampled from the same distribution (two-sample test).

The Kolmogorov-Smirnov statistic for the one-sample test is given by [63]:

$$D_N = \max |S_N(x) - P(x)| \tag{5.17}$$

where $P(x)$ is the continuous cumulative distribution function we want to test our data against, and $S_N(x)$ is the empirical cumulative distribution function constructed from the

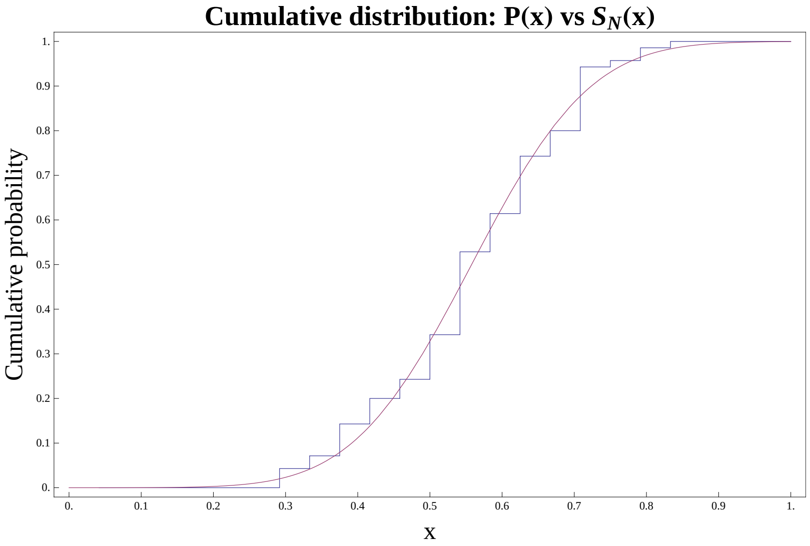


Figure 5.2: Example of an empirical cumulative distribution (blue) plotted against a normal distribution (red) with the same mean and variance.

data (assumed to be independent and identically distributed):

$$S_N(x) = \frac{1}{N}\sum_{i=1}^{N} I_{X_i \leq x} \tag{5.18}$$

and $I_{X_i \leq x}$ is the indicator function:

$$I_{X_i \leq x} = \begin{cases} 1 & X_i \leq x \\ 0 & \text{otherwise} \end{cases} \tag{5.19}$$

The distribution of the Kolmogorov-Smirnov statistic is given by the series:

$$Q_{KS}(\lambda) = 2\sum_{i=1}^{\infty} (-1)^{i-1} e^{-2i^2\lambda^2} \tag{5.20}$$

and the statistical significance is given by:

$$p_D = Q_{KS}\left(D_N\sqrt{N}\right) \tag{5.21}$$

If we wish to make comparisons between two data sets with n and n' elements each, without making any assumptions about the distributions that they are sampled from (as we would have to for the t-test, for example), we use the two-sample Kolmogorov-Smirnov test..

In this case, the Kolmogorov-Smirnov statistic is given by [63]:

$$D = \max |S_{N_1}(x) - S_{N_2}(x)| \tag{5.22}$$

where $S_{1,n}$ and $S_{2,n}$ are the empirical distribution functions for the first and second sample, respectively. The p-value is then:

$$p_D = Q_{KS}\left(D\sqrt{\frac{N_1 N_2}{N_1 + N_2}}\right) \tag{5.23}$$

The Kolmogorov-Smirnov test is quite general, and we will use it for all of our statistical tests, unless otherwise stated.

5.5 Monte Carlo simulations

Correlation plays a big role in our investigation; in particular, since we care about how the RCI questions relate to each other, we will study item-item correlations in the test data. When looking at data from N RCI questions, there are $\frac{N(N-1)}{2}$ distinct item-item correlations, and the possibility that some of these may be large by random chance is significant. The statistical significance of correlations is an issue that is not often addressed in physics education research, and we argue that this is a major oversight, particularly when these correlations are then used to inform factor analyses of the data, from which further inferences are often made.

We devised a Monte Carlo simulation to estimate the statistical significance of the item-item correlations, so as to have more confidence in our inferences about the internal structure of the RCI. This method is robust, in that it involves minimal assumptions about the data, or how it is distributed. In addition to this, Monte Carlo allows us to perform a “reality check” on our other statistical methods and tools - in particular, the factor analysis, which is used very uncritically in some PER, and in particular in Gibson’s RCI [17]. We describe the algorithm below.

Consider two questions, X and Y, for which we observe a correlation in the data of r_{XY}. We construct the familiar 2 × 2 contingency table for the frequencies of each of the four possible outcomes in our population:

	X correct	**Y incorrect**
X correct	N_{11}	N_{10}
Y incorrect	N_{01}	N_{00}

Table 5.1: 2 × 2 contingency table, giving the frequencies of each of the four possible outcomes, given a trial of $N_{11} + N_{10} + N_{01} + N_{00} = N$ students.

We generate a large population of students, and assume that the proportion of students that answered each question correctly in this population is the same as in our data (i.e., we estimate the population question means with our sample question means). This is the only strong assumption that goes into the simulation, and gives us the constraints:

$$N_{11} + N_{10} = N_X \tag{5.24}$$

$$N_{11} + N_{01} = N_Y \tag{5.25}$$

where N_X and N_Y are the number of students that answered questions X and Y correctly, respectively. We also require that all the frequencies sum to N, the number of students in our virtual population:

$$N_{11} + N_{10} + N_{01} + N_{00} = N, \tag{5.26}$$

We then specify the average correlation between the two questions in our virtual population:

$$\rho_{XY} = \frac{N_{11}N_{00} - N_{10}N_{01}}{\sqrt{(N_{11} + N_{10})(N_{11} + N_{01})(N_{00} + N_{10})(N_{00} + N_{01})}} \tag{5.27}$$

These four constraints allow us to uniquely specify the N_{11}, N_{10}, N_{01}, and N_{00} for our virtual population. In the case where we want to calculate the statistical significance of large correlations, we will set the population correlation ρ_{XY} to zero, but we may also set it to a high number, in the case where we want to check the statistical significance of *low* correlations. We then sample from this population, and count how many correlations are at least as large (or as small) as the correlation we observed. We model this situation with the multinomial distribution, which is a generalisation of the binomial distribution from 2 to k possible outcomes. The multinomial distribution tells us that over N independent cases (each student taking the test is independent, assuming there are no instances of cheating), the probability of measuring a result $\vec{x} = (x_1, \ldots, x_k)$ is given by the probability mass function [64]:

$$Pr\left(X_1 = x_1, \ldots, X_k = x_k\right) = \begin{cases} \frac{N!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k} & \text{when } \sum_{i=1}^{k} x_i = N \\ 0 & \text{otherwise} \end{cases} \tag{5.28}$$

where the x_i are the different frequencies for each of the k results. In our case, $k = 4$, and each of the x_i corresponds to one of the N_{11}, N_{10}, N_{01}, and N_{00}. It is then straightforward to calculate the correlation for this sample, using equation 5.27.

5.6 Item response theory: one parameter Rasch model

Item response theory assumes that there is one parameter that explains the performance of every student (their "ability"), and at least one parameter for the questions. In general, an Item Response theory model is formulated in terms of the probability of a student's probability of answering a given question correctly. When student i encounters question j, the probability of their answering it correctly is assumed to be given by:

$$P_{ij} = c_j + (1 - c_j)\frac{e^{a_j(\theta_i - b_j)}}{1 + e^{a_j(\theta_i - b_j)}} \tag{5.29}$$

where c_j is a constant for each question, which can be thought of as the base-line "guessing" probability. For each question, it would make sense to define $c_j = \frac{1}{n_j}$, where n is the number of answer alternatives. a_j is the item discrimination (distinct from discrimination in the classical test sense), and b_j is the item difficulty (again, distinct from the classical test theory sense).

The one parameter Rasch model makes the simplifying assumption that the $a_j = 1$ and

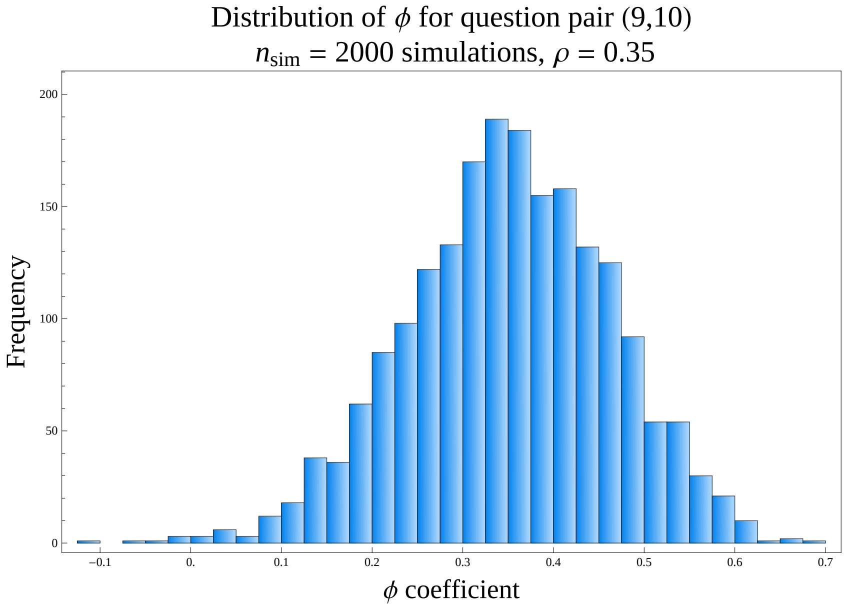


Figure 5.3: Monte Carlo simulation of correlations for the question pair (9,10). This is a population of 2000 trials $\times 63$ students = 126000 virtual students, where the mean population correlation is assumed to be $\rho = 0.35$, and the population means for questions 9 and 10 are given by our sample mean.

$c_j = 0$ for all j, so the questions are characterised only by their difficulties, b_j:

$$P_{ij} = \frac{e^{(\theta_i - b_j)}}{1 + e^{(\theta_i - b_j)}} \tag{5.30}$$

Although this is less general than 5.29, it has the advantage that it is a simple and transparent model, and will be adequate for our purposes. The goal of the Rasch model is to estimate all of the θ_i and b_j, such that equation 5.30 fits the data best. We'll use an iterative algorithm to do this.

The maximum likelihood estimator algorithm

I adapted this algorithm from Mark Moulton's demonstrative Excel Spreadsheet [65], and present it here.

We start with the matrix of raw student responses (1 or 0) to each question, which we will call $\boldsymbol{M}$. The component M_{ij} is student i's response to question j. Our first estimates of the student abilities and item difficulties (θ and b) are simply the logit[2] of the student raw scores and the question mean scores:

$$\begin{aligned} s_i &= \frac{1}{N_q} \sum_{j=1}^{N_q} M_{ij} \\ q_j &= \frac{1}{N_s} \sum_{i=1}^{N_s} M_{ij} \end{aligned}$$

[2] We use logarithmic units or "logits", so that difficulties and abilities can be composed additively, and reproduce probabilities when put through equation 5.30.

Hence (using the superscript to denote the number of the iteration):

$$\begin{aligned}
\theta_i^{(1)} &= \log\left(\frac{s_i}{1-s_i}\right) \\
b_j^{(1)} &= \log\left(\frac{1-q_i}{q_i}\right)
\end{aligned}$$

where s_i is the score of student i, and q_j is the mean score on question j, and N_s and N_q are the number of students and number of questions, respectively. The sign of b is reversed with respect to θ so that b represents the difficulty, rather than easiness of the questions. After adjusting the $\theta_i^{(1)}$ and $b_j^{(1)}$ so they each have a mean of 0, we plug them into equation 5.30, to produce a matrix of probabilities, which we will call $\boldsymbol{P^{(1)}}$. We compute the residual:

$$R_{ij}^{(1)} = M_{ij} - P_{ij}^{(1)} \tag{5.31}$$

and we estimate the variances of each of these probability estimates:

$$V\left(P_{ij}^{(1)}\right) = P_{ij}^{(1)}\left(1 - P_{ij}^{(1)}\right) \tag{5.32}$$

We then calculate our second estimate of the abilities and difficulties by adjusting our first estimate by the sum of the residuals divided by the sum of the variances:

$$\begin{aligned}
\theta_i^{(2)} &= \theta_i^{(1)} - \frac{\sum_{j=1}^{N_q} R_{ij}^{(1)}}{\sum_{j=1}^{N_q} V\left(P_{ij}^{(1)}\right)} \\
b_j^{(2)} &= b_j^{(1)} - \frac{\sum_{i=1}^{N_s} R_{ij}^{(1)}}{\sum_{i=1}^{N_s} V\left(P_{ij}^{(1)}\right)}
\end{aligned}$$

We re-adjust so that the means are again zero, and substitute the $\theta_i^{(2)}$ and $b_j^{(2)}$ into equation 5.30 to produce $\boldsymbol{P^{(2)}}$, and repeat the process. We keep iterating until the squared sum of the residuals:

$$E = \sum_i \sum_j \left(R_{ij}\right)^2 \tag{5.33}$$

converges to zero. Assuming the algorithm converges after n iterations, then the final result is the two lists:

$$\begin{aligned}
&\left\{\theta_i^{(n)}\right\}, \qquad i \in \{1, \ldots, N_s\} \\
&\left\{b_j^{(n)}\right\}, \qquad j \in \{1, \ldots, N_q\}
\end{aligned}$$

from which we can calculate $\boldsymbol{P}$, our matrix of probabilities with equation 5.30. This matrix is then used to correct the raw data for student's overall ability and question difficulties, to make inferences about the relationships between the test questions (see Section 6.7.4).

5.7 Factor analysis

Factor analysis is a dimension-reduction technique that attempts to explain the variance in an n-dimensional sample of data by extracting unobserved "factors" from the data, and was first proposed by Karl Pearson in 1901. It is frequently used in psychometrics and test design and analysis. Given N individuals with n data points each (in our case, n test questions), we ask whether there are $m < n$ unobservable "latent traits" (factors) that explain the n responses given for every student. This is represented with the linear model [66]:

$$s_{ij} = \sum_{p=1}^{m} a_{jp} F_{ip} + u_j Y_{ij}, \qquad i \in \{1, \ldots, N\}\,,\ j \in \{1, \ldots, n\} \tag{5.34}$$

where s_{ij} is the score of student i on question j, the coefficients a_{jp} are the so-called "factor loadings", and $u_j Y_{ij}$ is the residual. F_{ip} are the factor variables, usually taken to be normally distributed, and Y_{ij}, the error, is another random variable. The factor loadings a_{jp} are what we care about: they show how a given question is related to each of the factors. This is particularly relevant to concept inventories, because each linearly independent factor is expected to correspond to a different trait, or concept.

There are several different approaches to obtaining the model in equation 5.34. We won't elaborate on the details of the iterative algorithm here - they can be found in Harman's book [66]. Conceptually, one can think of the algorithm as searching the n-dimensional "question space" for m vectors whose projection onto the observed vectors is maximal. While factor analysis is a commonly used technique in concept inventory analysis [61], there are some issues about the sample size required for it to give reliable results. We will discuss this issue, and the interpretation of factor analysis results in section 6.7.2.

5.8 Conclusion

We are now armed with everything we need to collect, analyse, and interpret the data for our experiment. In the next chapter, we describe the experiment in detail, relate the key results, and draw conclusions from the analysis, using the tests and techniques from the previous chapter.

Chapter 6

Experimental results and analysis

This chapter will do two things: Interpret what the student responses on the RCI tell us about the concept inventory we've created, and find out what the RCI tells us about the students. A major question we want to ask of the RCI is whether or not the internal structure of the test matches our intended concept groupings from section 4.2 - this takes a large part of the analysis. We supplement the RCI question data with student confidence data, and data from other course assessments. Specific questions from the RCI are referred to periodically; the most frequently referenced ones are reproduced in this chapter, and the rest can be found in appendix A.

6.1 Curriculum

PHYS1201 (Advanced Physics 2) is the first year class which we will be studying. The curriculum for PHYS1201 consists of approximately three weeks on electromagnetism, three weeks on special relativity, three weeks on waves and optics, and three weeks on thermodynamics. The course prior to PHYS1201 is PHYS1101, in which students study mechanics and electromagnetism. The 1101 and 1201 courses are designed to give students a foundation in the main areas of physics (excluding quantum mechanics), so that they can focus in second year with 4 distinct courses: electromagnetism, waves and optics, statistical mechanics, and quantum mechanics.

The teaching package for the relativity section comprised 11 lectures, three tutorials, and one laboratory. The assessment consisted of two homework assignments, a lab log-book, and half of a two-hour mid-semester exam (the other half is on electromagnetism). The scope of the topic covered in lectures is: Galilei transformations, time dilation, length contraction, Lorentz transformations, relativity of simultaneity, space-time, four vectors, and mass-energy equivalence. The laboratory uses the Real Time Relativity software developed in a collaboration between ANU and UQ [67], with a focus on relativistic optics and getting students to construct a simulated experiment to verify time dilation.

Elements of the teaching package (lectures, homework, tutorials, exam) were all either adapted from pre-existing materials, or designed in tandem with the RCI. Though tutorials have played a large role in previous studies (most notably Scherr's extensive one at the University of Washington), they did not play a big role in this study.

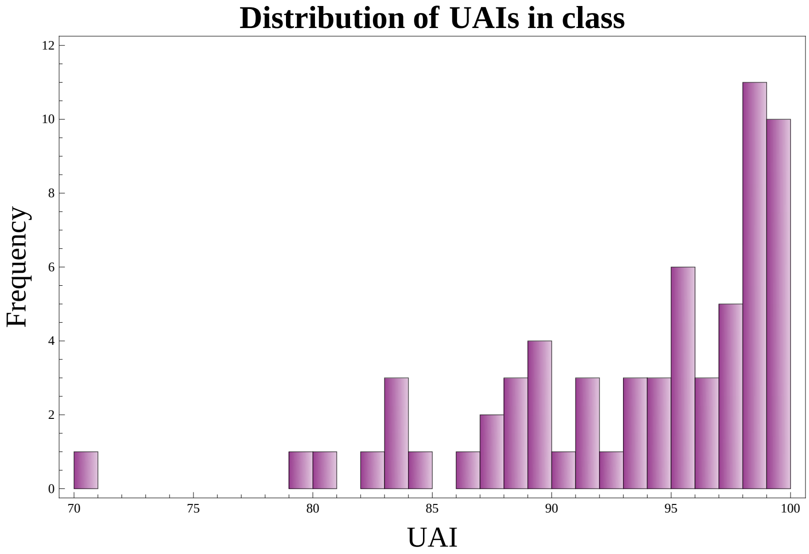


Figure 6.1: Distribution of UAIs for PHYS1201.

6.2 Student population

The number of students enrolled in the class was 99. The prior information we haev about our students is their degree concentration, gender, and Universities Admissions Index (UAI) score. The UAI is a percentile-based ranking system for prospective university students, based on their performance in their final two years of high school. We will make use of it as a good measure of prior achievement and general academic ability. The UAI data for the class was incomplete, though - UAI data was missing for many interstate and international students, so UAI data only existed for 64 of the 99 students. The median UAI for these 64 students was 95.3. Assuming the non-local students are of a similar calibre to the locals, this was an accomplished class; this has some bearing on our later analysis.

6.3 Ethics

This project was cleared by the ANU Human Research Ethics Committee, Human Ethics Protocol 2012/380. As part of the ethics requirements, an information statement on the study was made available to all PHYS1201 students (See Appendix E). This covered issues of confidentiality, and the storage and use of student data.

6.4 Set-up of the experiment

Adams and Wieman recommend that the pre-test be given on the first day of class, emphasising that it is not assessed for course grade, but that doing the test will benefit the students by giving the lecturer an idea of the strengths and weaknesses of the class. They also recommend giving the post-test on the second-last day of class - this advice is backed up by data from a study on the effects of the context of administering an electromagnetism concept inventory [68]. Both of these recommendations were adhered to.

Both pre-test and post-test were administered electronically[1]. Advantages of the electronic format include the ease of collection of data, and ease of administration to students that missed the lectures; students who missed the pre-test lecture were asked to take the test online, and unsupervised. The rationale for this was maximise the opportunity for students to participate in the study, so as to obtain the largest possible sample size. As the test was non-assessable, there was no incentive for students to cheat. Moreover, the data collection software monitors how long the students take to do the test, so spurious responses could be screened. In the end, the population of in-lecture and out-of-lecture responses were indistinguishable, both in terms of score, and time spent doing the test.

Pre-test

70 students took the pre-test (58 in lecture, 12 online). This test was administered in-class, in the first lecture on relativity, and online in the intervening two days between the first lecture and the second lecture. There was no indication that students didn't take the test seriously. Student responses were analysed for repeating patterns (e.g. answering "a" to every question), and none were found.

Post-test

63 students took the post-test (38 in lecture, 25 online). This test was administered in-class, in the second to last lecture on relativity, and online in the intervening two days between the second last lecture and the last lecture (a review session). The last lecture was on the same day as the mid-semester exam, and it was suggested to the students that doing the RCI post-test would help them prepare for the exam. To reward the students for taking the post-test, and to assist them in their exam preparations, we released the test solutions to the students after all the post-test responses had been collected. This practice is normally advised against in the concept inventory literature, as there is a concern that students will circulate the solutions and that these will find their way into the hands of the next group of students to be studied. We accept that there is a risk in future studies involving the RCI, but, given that the post-test was administered in a heavy exam week, we made our decision in the interest of fairness for the students.

Lecture questions

Lecture questions were designed to obtain additional data from the students in an informal setting. The results did not directly inform our conclusions, but we include them in appendix D.5, for archival purposes, should they be useful for future studies.

Homework

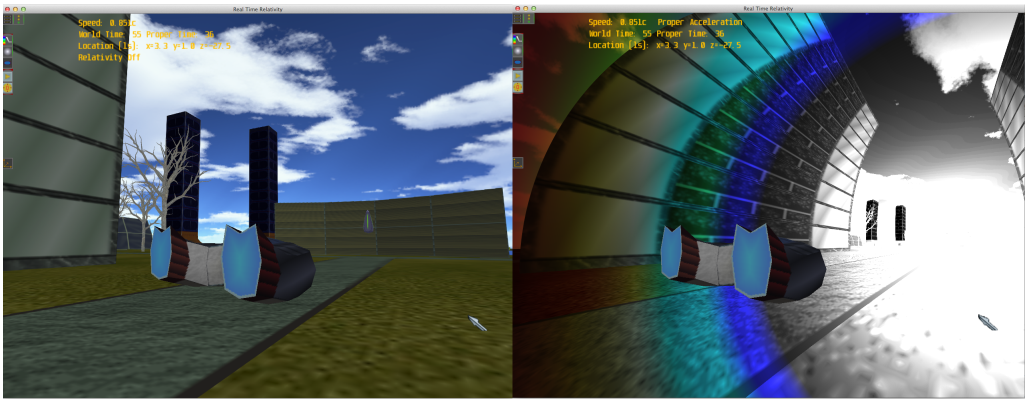
Students wrote open-ended answers to a qualitative time dilation problem. Their responses were qualitatively analysed, and the results are discussed in section 6.9.

[1] Most of the class had wi-fi enabled devices (laptops, iPads, or smartphones) with which to take the test. iPads were provided for students without such devices. We chose Survey Monkey to deliver both tests and collect results (www.surveymonkey.net).

Real time Relativity

Real Time Relativity (RTR) is a software package developed at the ANU to help teach special relativity [67]. Students explore relativistic scenarios simulated in a game-like environment, and do experiments in the simulation in a 3-hour lab format.

The point of the lab is to explore a realistic simulation, and to design an experiment to measure time dilation within the simulation. The student controls a spaceship in a game-like environment, and can fly around different scenarios at relativistic speeds. This requires the experimenter to find a way to minimise optical effects like aberration and Doppler effect, which can affect their measurements. The literature on special relativity education (**cite**) reports many instances of students interpreting relativistic effects (time dilation, length contraction, relativity of simultaneity) as "illusions", or optical effects only. The RTR simulation incorporates all relativistic efects, including optics, by using a raytracing algorithm to reproduce aberration, light travel delay, Doppler effect, and the headlight effect. Part of the idea is that, by exposing students to the messy reality of relativistic measurements, we can get them to differentiate between optical and non-optical relativistic effects. This proved to be a useful environment in which to informally probe student thinking about optical and relativistic effects, and the asymmetry misconception (see section 6.9).



Real Time Relativity

Speed: 0.851c
World Time: 55 Proper Time: 36
Location [ls]: x=3.3 y=1.0 z=-27.5
Relativity Off

Real Time Relativity

Speed: 0.851c Proper Acceleration
World Time: 55 Proper Time: 36
Location [ls]: x=3.3 y=1.0 z=-27.5

Figure 6.2: Screenshots from the Real Time Relativity Simulator. **Left:** Cityscape scenario with all relativistic effects switched off ($c = \infty$). **Right:** Same situation, but with relativistic effects enabled ($c = 1$). Length contraction, aberration, doppler shift, and the headlight effect can all be seen.

Mid-semester exam

A mid-semester exam worth 10% of the course grade was given to students after the post-test was administered, in order to externally validate aspects of the RCI. We structured the exam to address a situation similar to the one dealt with in questions 11, 12, 13 and 14 on the RCI. There was a mixture of calculation and explanation, with a particular emphasis on the physical origins of the relativity of simultaneity. There were several long-answer questions, designed to serve as written "think-alouds". The exam can be found in appendix C.

	$\bar{x}$	σ_x	$\bar{c}$	σ_c	$\bar{D}$	$\bar{r}_{pbc}$	KR-20	Ferguson's δ
Pre-test	0.56	0.13	0.50	0.12	0.22	0.27	0.48	0.93
Post-test	0.71	0.16	0.68	0.11	0.24	0.36	0.74	0.96
Desired values	$[0.30, 0.90]$	-	-	-	≥ 0.30	≥ 0.20	≥ 0.70	≥ 0.90

Figure 6.3: Descriptive statistics for pre-test and post-test, along with the desired values. Glossary: $\bar{x}$ is the mean total score, $\bar{c}$ is the mean confidence, $\bar{D}$ is the mean item discrimination, and r_{pbc} is the mean point-biserial coefficient, defined in section 5.2.

6.5 RCI Results and preliminary analysis

All analysis of the results were carried out using *Mathematica 8*, except for the factor analysis (*SPSS 20*), and the Rasch analysis (*Winsteps*).

Descriptive statistics

The desired values for these descriptive statistics are values that are widely agreed upon in the concept inventory community [61]. The RCI post-test results are within the acceptable range for all of these basic statistics, except one: the mean item discrimination. This means that the average RCI test item is not discriminating strongly between students that perform well on the test and students that perform poorly. This low mean discrimination is mostly due to low discrimination in questions 12, 13, and 14, which are three of the easiest on the test (see figure 6.6). In keeping with Adams and Wieman's recommendations (see section 5.2), this does not necessarily mean these items need to be revised or removed, as these are questions that are merely demonstrating strong performance over the whole class.

The pre-test and post-test participants are different subsets of the class (see figure 6.4). Before analysing the test in detail, we will check whether these subsets are biased samples of the class or not.

Testing for sample bias in the pre∩post group

We want to test for whether the pre∩post group (those that did both the pre-test and post-test) is a biased sample of the class, since we are using their results to calculate the normalised gain for the RCI, and it is possible that these students are more engaged than the norm for the class. We test this hypothesis with the Kolmogorov-Smirnov test, and comparing the total pre-test scores of students that did the pre-test only, and students that did both the pre-test and post-test, and likewise for the post-test:

	Pre-test only	Pre-test and post-test	Post-test only	p-value
$\bar{x}_{pre}$	0.55	0.56	-	0.86
$\bar{x}_{post}$	-	0.71	0.74	0.84

Table 6.1: Mean total scores for students that did pre-test only, post-test only, and both pre-test and post-test.

Recall that a high p-value indicates that it is likely that the two samples are from the same distribution. From the p-values in the table, we can see that there is no statistically

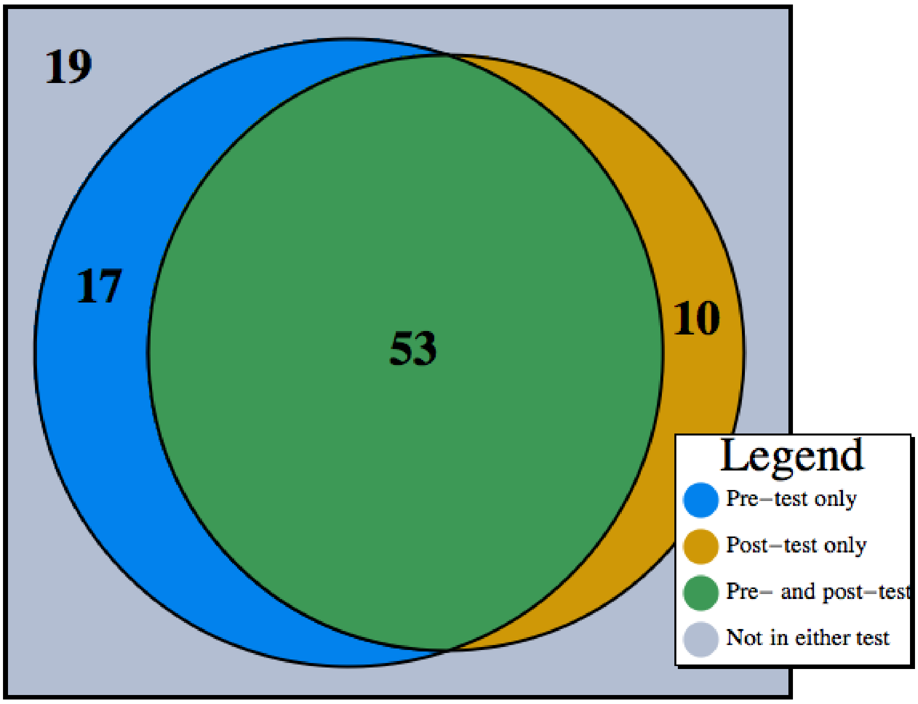


Figure 6.4: Venn diagram illustrating the class participation in the pre-test and the post-test.

significant difference between the different subsets, so we conclude that it is unlikely that the pre∩post group represents a biased sample of the class.

Testing for sample bias in the pre∪post group

We will test whether the pre∪post group (those that did at least one of the RCI tests) is a biased sample of the class. Although they are a majority (80 of the 99 students), the 19 students we missed may be a different population that would have affected the results if they had participated. We will use the relativity section of the mid-semester exam to benchmark these two groups. 93 of the 99 students enrolled took the exam, and this discrepancy is evenly shared between the RCI and no RCI groups:

	RCI	No RCI	p-value
N	77	16	-
$\bar{x}_{exam}$	0.66	0.46	0.008

Table 6.2: Mean mid-semester exam scores for students that took part in the RCI, and those that did not.

This low p-value indicates a strong sample bias in favour of the students that participated in at least one RCI test. The most plausible explanation for this is that this is a self-selection effect, as the students that did the RCI were those that were most engaged with the course, and their exam marks reflected this.

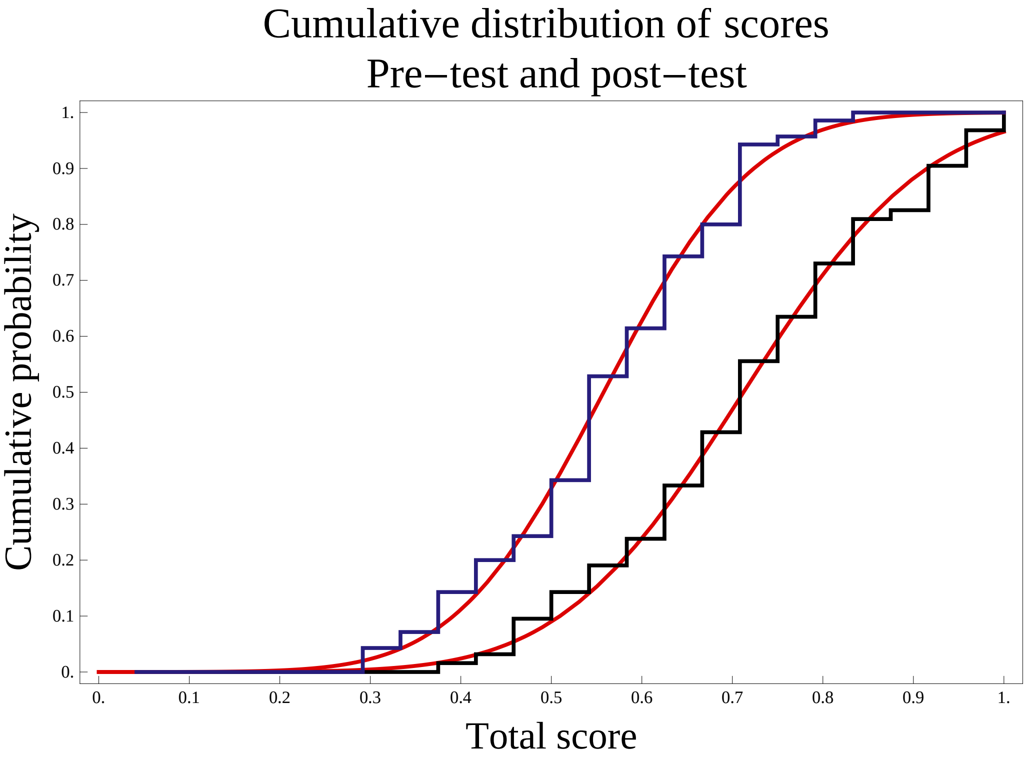


Figure 6.5: Pre-test (blue) and post-test (black) cumulative probability distributions with normal distributions of the same mean and variance superimposed (red).

Testing the effect of instruction

Are the pre-test and post-test scores significantly different? The Kolmogorov-Smirnov test gives a $p = 4.2 \times 10^{-6}$, and Pearson's χ^2 gives $p = 1.2 \times 10^{-3}$. This indicates that there is a very low probability that the observed gains could have occurred by chance, and so the effect of instruction has been to change the class's performance.

Are the total scores on the pre-test and post-test normally distributed? The Kolmogorov-Smirnov test gives p-values of 0.39 and 0.86 for pre and post-test, respectively. So, the null hypothesis that the test scores are normally distributed can be neither accepted nor rejected for the pre-test or the post-test. This is important, because it affects our statistical analysis, and which tests we choose to use. In the absence of a normal distribution, we need to be wary of using parametric statistics such as the χ^2 and student's t test. In general, we will use the non-parametric Kolmogorov-Smirnov test (defined in section 5) to calculate statistical significance, unless stated otherwise.

Connection with prior instruction

All the students enrolled were emailed, and asked whether or not they had any formal instruction in special relativity prior to taking PHYS1201. Of the 99 enrolled in the class, 58 replied to the email with a yes or no. Many also gave additional information about informal self-instruction in relativity, e.g. reading popular science books or magazines.

	$N_{replies}$	N_{yes}	N_{no}	$\bar{x}_{yes}$	$\bar{x}_{no}$	$\bar{x}_{class}$	p-value
Pre-test	45	20	25	0.56	0.54	0.56	0.69
Post-test	44	19	25	0.71	0.72	0.71	0.74
Exam	58	25	33	0.62	0.66	0.62	0.22

Table 6.3: Mean pre-test, post-test, and exam scores for students with prior instruction (x_{yes}), and without prior instruction (x_{no}).

From this data we conclude that there is no statistically significant connection between prior instruction and pre-test, post-test, or exam results. This suggests that at the conceptual level, high school physics teaching may not be particularly effective.

The mean pre-test score was 0.56. As there was no significant correlation with prior formal instruction, this suggests either that the test was too easy, or that many students had acquired some knowledge of special relativity in informal ways. All questions in which pre-test scores were high were reviewed (see figure 6.6), and one (question #21) was removed, on the grounds that it was too easy. The others were retained, because they were necessary for the comprehensiveness of the test.

Normalised gain

Mean normalised gain for class, averaged over the static[2] questions was $\bar{g} = 0.40$, which is an indicator of moderately strong learning gains for the class as a whole [43]. Length contraction showed the largest normalised gain, and mass-energy showed a negative normalised gain (see figure 6.7). Mass-energy equivalence was not a focus of the course, but this is on its own is not a reason for the gain to be negative. A more likely explanation is that question 24 (mass-energy) was not sufficiently well specified, and students' greater knowledge in the post-test caused them to over-interpret it. Variations in normalised gain within concept groups generally reflect different degrees of difficulty: items 7, 10, 15, and 17 all involve applying the concepts in unfamiliar and counter-intuitive scenarios, so their lower normalised gains are to be expected. The large normalised gain in item 6 with respect to item 5 can be attributed to the high teaching intensity on this aspect of time dilation. The "perfect" normalised gain of 1, for question 13, is an indicator that this is a straightforward question in a concept that the class uniformly understood as a result of instruction.

6.6 Student confidence

We implemented a confidence scale on all both the RCI pre-test and post-test, for students to self-assess their confidence in their answers to each question. In principle, this extra data allows us to distinguish between students that are guessing, and students that are confident; this provides us with more data which will help to interpret their responses. Confidence scales have been used before by Allen et al. in the Statistics Concept Inventory [20] and others (discussed in section 2.3). We will extend this technique in our investigation. Without a method of implicitly assessing student

[2] All questions except 18, 19, and 21 were held constant from pre-test to post-test.

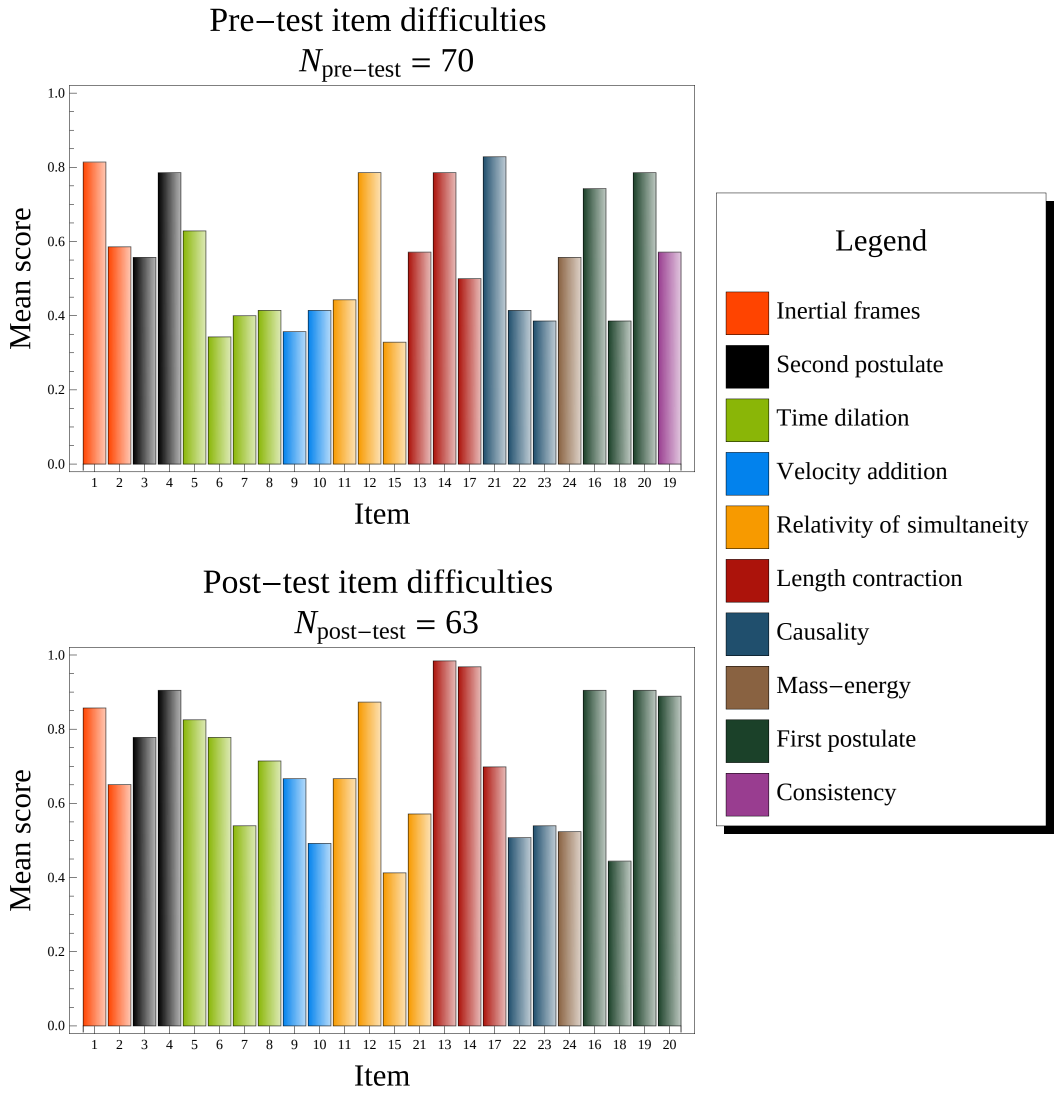


Figure 6.6: Mean item scores, pre-test and post-test. Note that questions 18, 19, and 21 were changed from pre-test to post-test, and the "consistency" concept (purple) was dropped for the post-test.

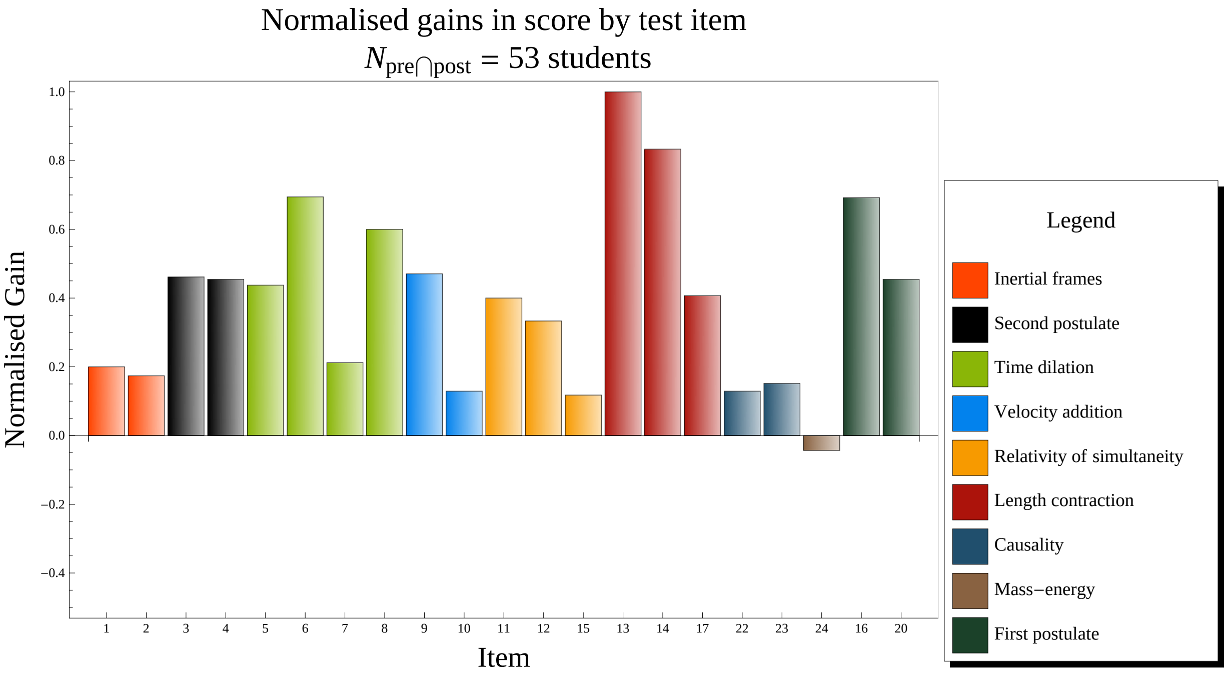


Figure 6.7: Normalised gain by item, grouped according to concept. Note that items 18, 19, and 21 are omitted, since they were changed from pre-test to post-test.

confidence, we are forced to make the strong simplifying assumption that students' self-assessment of their confidence is honest and accurate, which has been suggested to not be true in general [69] - this is a weakness of the confidence data, but one we will live with.

In the RCI, students are presented with a Likert-type scale underneath each question, which is rated on a numeric scale (1-5):

Rate how confident you are in your answer:

○·········○·········○·········○·········○

guessing unconfident neutral confident certain

We present the confidence data in appendix D.3, to avoid cluttering the discussion. As expected, confidence increased from pre-test to post-test in all static questions.

In general, confidence was moderately correlated with performance on the post-test, with an average correlation of $r = 0.189$, by question. I didn't calculate the statistical significance of these correlations, since there are far fewer of them than there are item-item correlations, so the chance of "false positives" is much lower, by approximately a factor of 10 (the number of correlations here is 24, whereas there are 276 item-item correlations). Correlation between confidence and performance is an indicator of mastery; questions with positive correlation are demonstrating student expertise in this respect. There are several with low, or negative correlations, and these merit investigation in detail, as this could be an indicator of problems with the questions. There are two ways that low or anti-correlation can manifest itself, with very different implications: a large number of (1) students guessing the question right, or (2) students confidently answering the question wrong. Let's find out:

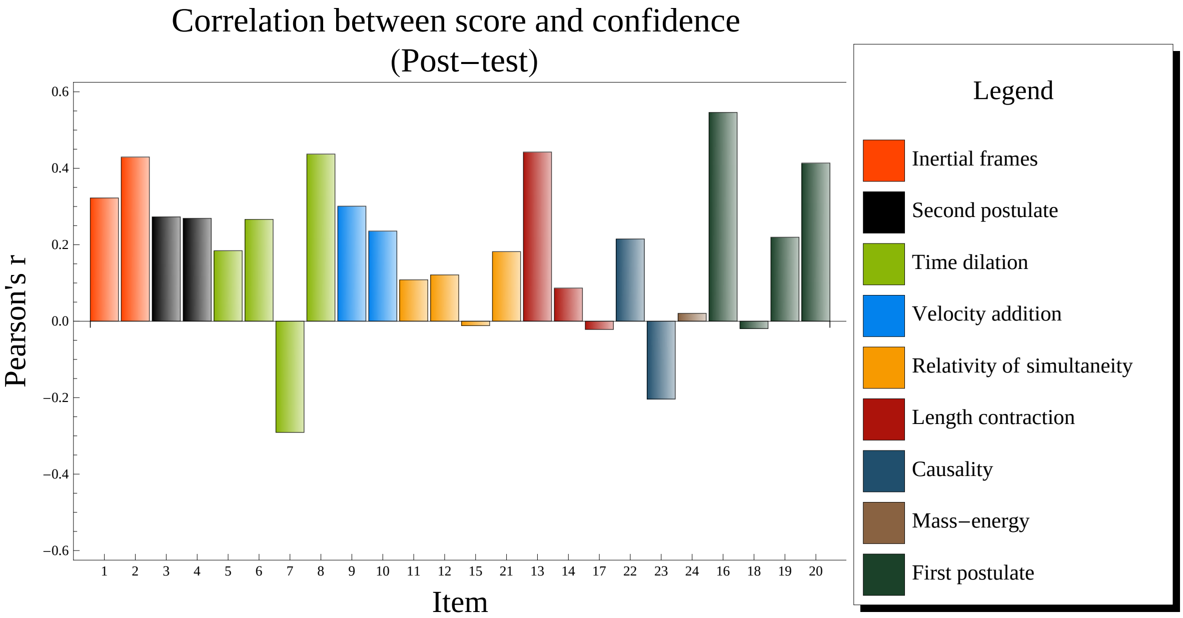


Figure 6.8: Correlation between score and confidence, post-test.

Terms on the diagonal of the contingency matrix contribute to the positive correlation component of the signal. What interests us is where most of the contribution in the cross terms is coming from: whether it is mainly from "confident and incorrect" students , or "unconfident and correct" students. Let $\boldsymbol{M}$ represent one of our contingency matrices in figure 6.9:

$$\boldsymbol{M} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \tag{6.1}$$

I'll now define k as the difference $b - c$ normalised by the total. I'll call this k the *confidence-score characteristic*:

$$k = \frac{b - c}{a + b + c + d} \tag{6.2}$$

Q 7	Confident	Unconfident	
Correct	22	4	
Incorrect	23	1	
			50

Q 15	Confident	Unconfident	
Correct	18	4	
Incorrect	25	5	
			52

Q 17	Confident	Unconfident	
Correct	26	6	
Incorrect	12	2	
			46

Q 23	Confident	Unconfident	
Correct	5	17	
Incorrect	10	9	
			41

Q 24	Confident	Unconfident	
Correct	14	7	
Incorrect	12	7	
			40

Q 18	Confident	Unconfident	
Correct	10	11	
Incorrect	12	9	
			42

Figure 6.9: 2×2 contingency matrices for confidence-score correlations for questions with low or negative confidence-score correlation. Note that "certain" and "confident" were binned together into "confident", and "unconfident" and "guessing" were binned together into "unconfident". Students that rated their confidence as "neutral" were excluded.

A large, negative value of k, as defined here, would suggest a question in which the bulk of the anti-correlation component of the signal is coming from the "confident and incorrect" students. This would indicate a question in which students have a strong misconception, or possibly a "trick" question.

On the other hand, a large, positive value of k would suggest a question in which the bulk of the anticorrelation component of the signal is coming from the "unconfident and correct" students. This would indicate a question in which students are not confident of their understanding, but are able to guess the right answer, possibly due to a low number of answer alternatives N_a. Let's investigate further:

Question	**Concept**	k	N_a
7	Time dilation	-0.38	2
15	Relativity of simultaneity	-0.40	5
17	Length contraction	-0.13	3
18	First postulate	-0.02	3
23	Causality	0.17	4
24	Mass-energy	-0.125	3

Table 6.4: Confidence-score characteristics for questions with low or negative confidence-score correlations.

Interesting! The only question with a positive k-value is question 23, and it has a high number of answer alternatives, which would seem to suggest that random guessing is not the source of the "guessing and correct" signal. A possible explanation is that the distractors used in this question are not very plausible, so students are able to guess the answer without any real understanding of causality. However, question 23 is one of a pair of questions on causality, and its pair, question 22, has the same set of distractors, yet exhibits a positive correlation between score and confidence. This is puzzling.

Questions 17, 18, and 24 seem to have "guessing and correct" and "confident and incorrect" responses in roughly equal measure, so it's difficult to draw any conclusions about these. Incidentally, these three questions had almost zero correlation between performance and confidence. Questions 7 and 15 have large, negative k-values, so we must conclude that either they are trick questions, or the class has strong misconceptions about these concepts. Given that student performance in the other time dilation questions (5,6, and 8) on the post-test was generally good, and correlation with confidence is positive for these questions, we must conclude that question 7 is a "trick" question for many students, even though that is not the intention (see figure 6.10.

7. It is known that our galaxy is of the order of 100,000 light-years in diameter. True or false: "Travelling at a constant speed that is less than, but close to, the speed of light, in principle it is possible for a person to cross the galaxy within their lifetime."

   (a) True

   (b) False

Figure 6.10: RCI question 7. Our correct answer is *(a)*.

This analysis indicates some questions to keep an eye on, as they may be flagging

15. Two separate light bulbs emit flashes of light, distant from an observer. This observer receives the light from both flashes at the same time. From this alone it is possible to conclude that:

    (a) The flashes occurred at the same time for all observers

    (b) The flashes occurred at the same time for the observer at that location

    (c) The flashes occurred at the same time if the observer is not moving relative to the light bulbs

    (d) It is not possible to make any of the above conclusions

Figure 6.11: RCI question 15. Our correct answer is *(d)*.

strong misconceptions. It is well known that students have difficulty with the relativity of simultaneity, and question 15 is a slightly unorthodox question addressing this problem (see figure . Question 7 seems to be counter-intuitive for many students, but possibly not betraying a strong misconception, given the relatively good performance in other time dilation questions in the post-test. Other items to note from 6.4 are 17 (applying length contraction in an unfamiliar context), and 18 (first postulate in the context of inertial mass), which we also suggest are of the "counter-intuitive" type, rather than of the "strong misconception" type, due to the relatively good performance in other questions in those concept groups. These hypotheses could be checked with student interviews.

6.7 Validating the RCI

In this section we will ascertain the validity of the RCI as a instrument for measuring student understanding. We will examine its relationships with other course assessments, which is a measure of its predictive power, and we will investigate the internal correlations between questions, to check its conceptual coherence.

6.7.1 External validity

Normalised gains on the RCI are reasonably well correlated with UAI ($r_{g,UAI} = 0.44$) and with exam scores ($r_{g,exam} = 0.39$). In the literature, a variety of correlations are reported; McKagan et al. [29] report no significant correlation between their concept inventory and exam results, while Smith et al. [31] report a large correlation between normalised gain and exam of $r = 0.65$. Our result is somewhere in the middle. This moderate level of correlation is plausible, given the broad range of concepts on the RCI, and the relatively narrow focus of the exam (see C).

Validity with respect to a key concept: the relativity of simultaneity

In open ended responses on the mid-semester exam, some students expressed the "relativity of simultaneity as a light delay phenomenon" misconception, thus reproducing Scherr's result, discussed in chapter 3. Students are required to calculate the time interval between the impacts of two balls dropped out of a speeding train (as measured by an observer on a platform), and are also required to give the physical reasoning for their answer (the mid-semester exam can be found in appendix C). 20% of students gave responses in which they attribute the relativity of simultaneity to light delay, for example:

> *"The ball that hit the ground closer... The light coming from this ball reaches him* [the platform observer] *quicker than the other ball, giving the illusion that one hits before the other."*

We also observed some (rare) instances of students calculating the correct answer with the Lorentz transformations, but then dismissing it, presumably because of a strong belief in absolute simultaneity:

> *"Despite the different answers given by Lorentz laws:* $t_{fall,L} = \gamma t'_{fall}$ *and* $t_{fall,R} = \gamma \left(t'_{fall} + \frac{vD}{c^2} \right)$*, if both balls hit the ground at the same time for Amanda, they must for Bryan."*

The exam mostly focused on student understanding of relativity of simultaneity; this is because: (1) the experts emphasised its importance, and (2) Scherr's work suggested that the relativity of simultaneity was a "hub" concept, and that relativity of simultaneity misconceptions were symptomatic of problems in other areas of relativity. Scherr's result suggested that students that believe that the relativity of simultaneity can be attributed to light delay were also having difficulties with the reference frame concept, and our hypothesis is that this will translate to low performance in the RCI overall. Question *(i)* on the exam addresses the question of light delay directly (see figure6.12).

(i) (2 marks) *Can the relativity of simultaneity be described simply as a light delay effect, i.e. can the lack of agreement on the simultaneity of separate events for different observers be account for just by signal delays? Explain briefly.*

Figure 6.12: "Light delay" question on the mid-semester exam.

Of 93 students that took the exam, 61 answered no, 17 answered yes, and 15 gave no response. From the subset of these that took the post-test, 43 answered no, and 8 answered yes. This gives us a small sample to work with, and so the statistical significance of these results is not high, but nevertheless worthy of investigation:

	$\bar{x}_{correct}$	$\bar{x}_{incorrect}$	**p-value**
RCI Post-test (N=51)	0.74	0.64	0.52
Exam (N=93)	0.73	0.42	0

Table 6.5: Mean scores on the RCI post-test and exam, for students that answered question *(i)* correctly ($x_{correct}$) and incorrectly ($x_{incorrect}$).

From table 6.5, we see that students that answered exam question *(i)* wrongly are not a statistically different group to those that answered it correctly, with respect to RCI responses . Although the sample size is low, this would suggest that our earlier hypothesis about students with the "light delay" misconception is incorrect. On the other hand, the two groups are strongly differentiated by the exam, with a difference in mean between the two groups of over 30%. This implies that the "light delay" misconception severely affects student reasoning ability in both quantitative and qualitative questions, with respect to the relativity of simultaneity. The RCI is a good predictor of exam score for those that got *(i)* correct, but overestimates the exam scores of those that didn't by over 20 points, and so its predictive power with respect to the "light delay" group is lower than for the "correct" group.

Although there is no statistically significant difference between the mean scores of the two groups ("light delay" and "correct"), the relative performance of the two groups across all RCI questions shows a suggestive signal (see figure 6.7.1). In particular, the "light delay" group perform worse on all of the relativity of simultaneity questions (in yellow), which is encouraging, although not conclusive. The fact that the light delay group performed *better* on questions 8, 10, 17, and 24 is puzzling, although again, this is not a statistically significant result. This is a possible avenue for further work - in particular, student follow-up interviews would further elucidate the thinking of this "light delay" group. Overall, this is encouraging for the validity of a key part of the RCI.

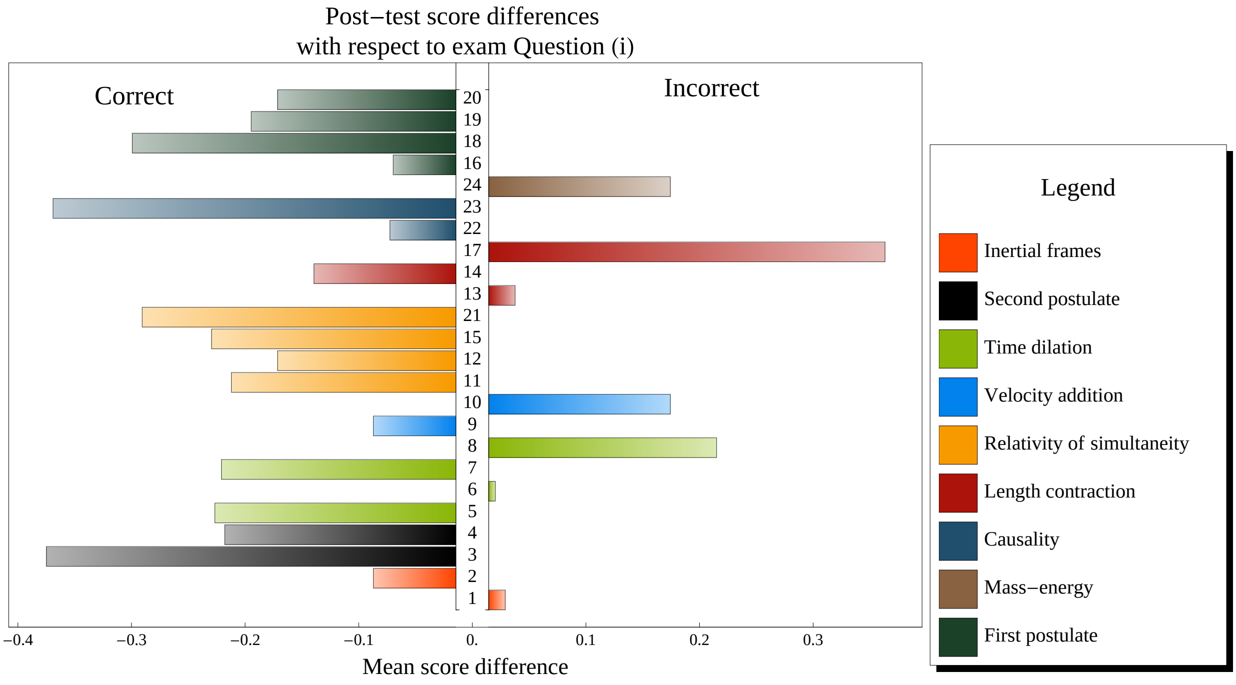


Figure 6.13: Post-test score difference between light delay and correct groups, by question.

6.7.2 Coherence in the internal structure of the RCI

We now look at the internal structure of the RCI - in particular, the relationships between questions, and whether these relationships align with our intentions: the questions have been grouped into concepts, and we want to see if the student data reproduces this grouping. Factor analysis is frequently used in the concept inventory literature [21, 70, 42, 17] for this purpose.

Factor analysis

Despite the ubiquity of factor analysis, there is some controversy in the literature as to what sample size is required for a factor analysis to be valid. A general rule of thumb adhered to in the PER community [12, 42], is that a meaningful factor analysis requires at least 10 times as many responses as there are questions. Our study falls significantly short of this requirement; for the post-test we have 63 responses and 24 questions, giving a responses-to-questions ratio just shy of 3:1. However, MaacCallum et al. argue, based on Monte Carlo simulations of continuous data, that a sample size of 60 is sufficient to reliably extract a small number of factors [71]. These contradictory results suggest, at minimum, caution in interpreting the results from the factor analysis. I performed an exploratory factor analysis on the pre-test and post-test data, as well as a set of uniformly distributed random responses from Monte Carlo, using the *SPSS 20* statistics software (see Figure 6.14).

The "Scree" plot is a way of presenting the eigenvalues for each factor that is extracted from the data. The eigenvalue is a measure of how much variance is accounted for by each factor: the higher the eigenvalue, the more of the raw data it accounts for. Particularly noteworthy, in Figure 6.14, is the fact that there is very little difference between the pre-test eigenvalues and the Monte Carlo eigenvalues, despite the fact that the pre-test responses have no resemblance to any set of randomly-chosen responses. This suggests that the factor analysis algorithm is capable of finding factors, or patterns in the data, even when there are none, simply by virtue of its optimisation technique. From this we argue that there are no clear factors in the pre-test.

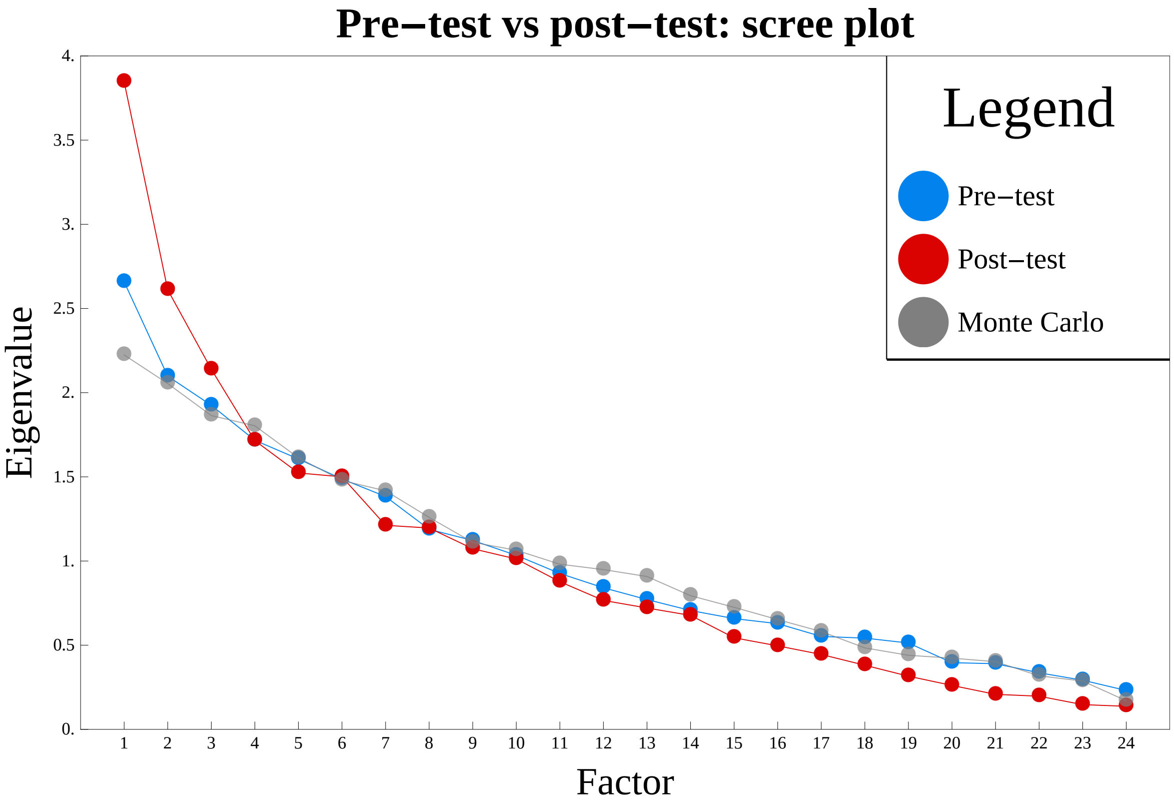


Figure 6.14: Scree plots for the RCI, including pre-test, post-test, and Monte Carlo.

The post-test scree plot (in red) differs significantly from the Monte Carlo in two or three factors. The factor loadings showed no clear patterns, although questions 1 and 2 (inertial frames) both strongly correlated with factor 1, questions 5 and 6 (time dilation) with factor 2, and questions 19 and 20 (first postulate) with factor 3, which is suggestive of at least some coherence in the data. The largest factor (factor 1) is only strongly

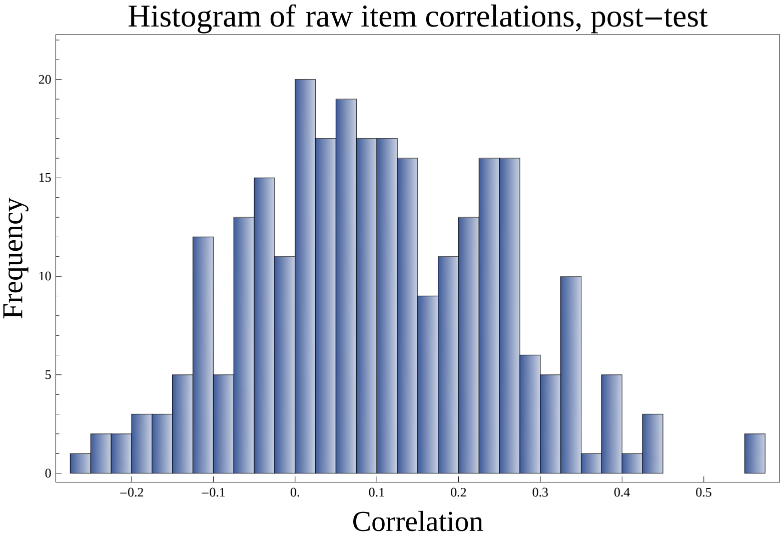


Figure 6.15: Distribution of the raw correlations, post-test.

correlated with 6 questions, so we can't interpret this as a "relativistic thinking" factor, or as any other indicator of general ability. It turns out that the post-test factor loadings don't tell us anything that the raw correlations don't. In any case, we are sceptical of the factor analysis results because of the issue with sample size, so we will move on to our own analysis of the item-item correlations to get a clearer idea of the structure of the RCI.

Raw correlations between questions

The number n_C of distinct correlations is clearly related to the number n_Q of questions by:

$$\begin{aligned} n_C &= \frac{n_Q\,(n_Q - 1)}{2} \\ &= 276 \end{aligned} \tag{6.3}$$

Note in figure 6.15 that the correlations aren't centered around zero: the mean correlation was $\mu_C = 0.098$, with standard deviation $\sigma_C = 0.15$. There is a lot of positive correlation here - most of this can be attributed to student ability, as we will see. In table 6.6, we present the 12 statistically significant correlations between questions in the data. The statistical significance is calculated with Monte Carlo simulations, detailed in section 5.5.

Question	Question	Pearson's r	p-value (one-tailed)
2	1	0.557	0
6	5	0.559	0
7	2	0.392	0
9	1	0.385	4×10^{-3}
9	3	0.432	0
12	11	0.438	5×10^{-4}
16	13	0.391	0.019
19	13	0.393	0.014
20	13	0.359	0.028
20	19	0.401	4×10^{-3}
22	9	0.382	5×10^{-4}
22	15	0.438	5×10^{-4}

Table 6.6: Statistically significant item-item correlations, post-test.

6.7.3 Lack of correlation?

Below is a table of post-test questions that we expected to be correlated, but which are not significantly correlated at the 95% level. We calculate the significance of their *lack* of correlation, assuming the students' responses are sampled from a population in which the questions are strongly correlated ($\rho = 0.35$).

Question pair	Concept	Observed correlation	p-value (one-tailed)
$(3, 4)$	Second postulate	0.22	0.2
$(9, 10)$	Velocity addition	0.090	0.016
$(16, 19)$	First postulate	0.26	0.33
$(6, 8)$	Time dilation	0.34	0.48
$(15, 21)$	Relativity of simultaneity	0.27	0.23
$(22, 23)$	Causality	0.30	0.34

Table 6.7: Question pairs that were not significantly correlated, with corresponding p-values, assuming the population *is* correlated.

From these p-values, we can only rule out any relationship between questions 9 and 10 - both intended to test velocity addition. This result means it is clear that one or both of these questions is not doing its job correctly, and we are inclined to think that it might be question 10, given that question 9 has many desirable item characteristics, including high item discrimination and point-biserial coefficients (see appendix D.4). Question 10 is then a candidate to be removed.

6.7.4 Correcting for student ability with the Rasch model

Even though all the correlations in Figure 6.16 are statistically significant, that doesn't necessarily mean we can attribute all of the correlation signal to conceptual links between the questions. One expects that the dominant effect driving the correlations is student ability: i.e. given a question pair (X, Y), good students will tend to get both right, and bad students will tend to get both wrong - this strengthens the overal correlations. We will use a one-parameter Item Response Theory (IRT) model, the Rasch model, to iteratively calculate the student abilities, so as to correct for them, and examine the residual correlations for conceptual coherence. This method was suggested to me by Paul Francis, at ANU.

I used *Ministeps*[3] to do the Rasch analysis. The Rasch algorithm is described in Section 5.6 - it gives us a set of abilities θ_i, one for each student, and a set of difficulties b_j, one for each question. We can then use these to create matrix of probabilities:

$$P_{ij} = \frac{e^{(\theta_i - b_j)}}{1 + e^{(\theta_i - b_j)}} \tag{6.4}$$

subtract this from our raw data M_{ij} (binary data containing every students response to every question), to give a residual:

$$R_{ij} = M_{ij} - P_{ij} \tag{6.5}$$

and then calculate the cross-correlations between questions, by taking the product between the residuals for each question pair, and average over the students:

$$C_{jk} = \frac{1}{N}\sum_{i=1}^{N} R_{ij} R_{ik} \tag{6.6}$$

where N is the total number of students. The number of residual cross-correlations C_{jk} is again 276, given by 6.3. Naturally, the residuals are small, and they average to zero (by construction of the model), but by looking at the outlying negative or positive ones, we can find residual correlations between questions once student abilities have been subtracted off. In order to extract the significant ones, we make the simplifying assumption that they are normally distributed[4], and look for correlations that are more than 3 σ away from the mean, corresponding to p-values of $< 3 \times 10^{-3}$.

Question pair	**Concept**	**Residual cross-correlation C_{jk}**	σ
$(1, 2)$	Inertial frames	0.066	3.4
$(5, 6)$	Time dilation	0.080	4.0
$(11, 12)$	Relativity of simultaneity	0.066	3.4
$(7, 8)$	Time dilation	-0.083	3.6
$(23, 24)$	Causality, mass-energy	-0.086	3.7

Table 6.8: Large residual item-item correlations, after student ability has been corrected for.

This result is encouraging. Notably, all correlations involving the "hub" questions 13 and 9 (see 6.16 have vanished, indicating that their strong correlations were indeed due to the overwhelming "ability" signal. The strongest correlations that existed in the raw

[3] http://www.winsteps.com/ministep.htm, retrieved 29 September 2012.

[4] This is a major simplification, but not grossly invalid - One-sample Kolmogorov-Smirnov test gives p-value for the correlations being normally distributed of 0.69.

signal remain after correcting for student ability: $(1,2)$, $(5,6)$, and $(11,12)$. This result strongly suggests that these questions pairs are well connected, which is evidence that they are testing the concepts that they purport to test. As for the two anti-correlated question pairs $(7,8)$ and $(23,24)$:

- $(7,8)$: We already know that question 7 is an anomalous question, because it has a strong anti-correlation between performance and student confidence. Of interest is their commonality: they both involve using time dilation on the "galactic" scale. It is possible that this anti-correlation relates to the asymmetric time dilation misconception, and interviews would be an obvious way to confirm or deny this.

- $(23,24)$: This pair is interesting, given that 23 and 24 don't actually make up a complementary question pair, or even share the same concept. However, 24 is anomalous, as it was the only question to have a *negative* normalised gain (see Figure 6.7).

6.7.5 Discussion

We have established the general validity of the RCI, and identified some questions that are candidates for being removed. With some rigorous statistical analysis, we have shown that some of the item-item correlations match with our intended concept groups, although not many. As with much PER, there is a lot of noise in the signal, and it is hard to draw conclusions about the other item-item correlations (dashed red lines in figure 6.16). In particular, it is puzzling that the first postulate questions are not strongly correlated, and that the causality ones are not either - although there is a chance that these low signals are statistical flukes (see table 6.7). Another possibility is that our grouping of the concepts is wrong, and that more appropriate groupings are possible. A third possibility is analogous to Huffman and Heller's interpretation of the FCI (presented in the literature review), that there is no conceptual coherence in the RCI - that students simply do not see the concepts "grouped" the way teachers do. It is likely that it is a combination of these two cases, given the robustness of some of the conceptual pairings we have found.

6.8 A gender bias?

There is a statistically significant gender difference in our RCI results; males outperform females. This gender difference isn't reproduced anywhere else in the course assessment, or in the prior achievement of the class (measured with UAI). Concept inventories have some history with gender differences; prior investigations **[72, 73]** have identified that there exist gender differences in the FCI, biased towards males. Gender differences in concept inventory performance have also been observed in other areas of physics [62] and in chemistry [41]. Some attempt has been made to account for these gender differences, but no plausible explanation for their origin has yet been found. Over the following pages, we will carefully elucidate this gender difference, starting with the class demographics. As is typical of most physics classes [73], females are under-represented:

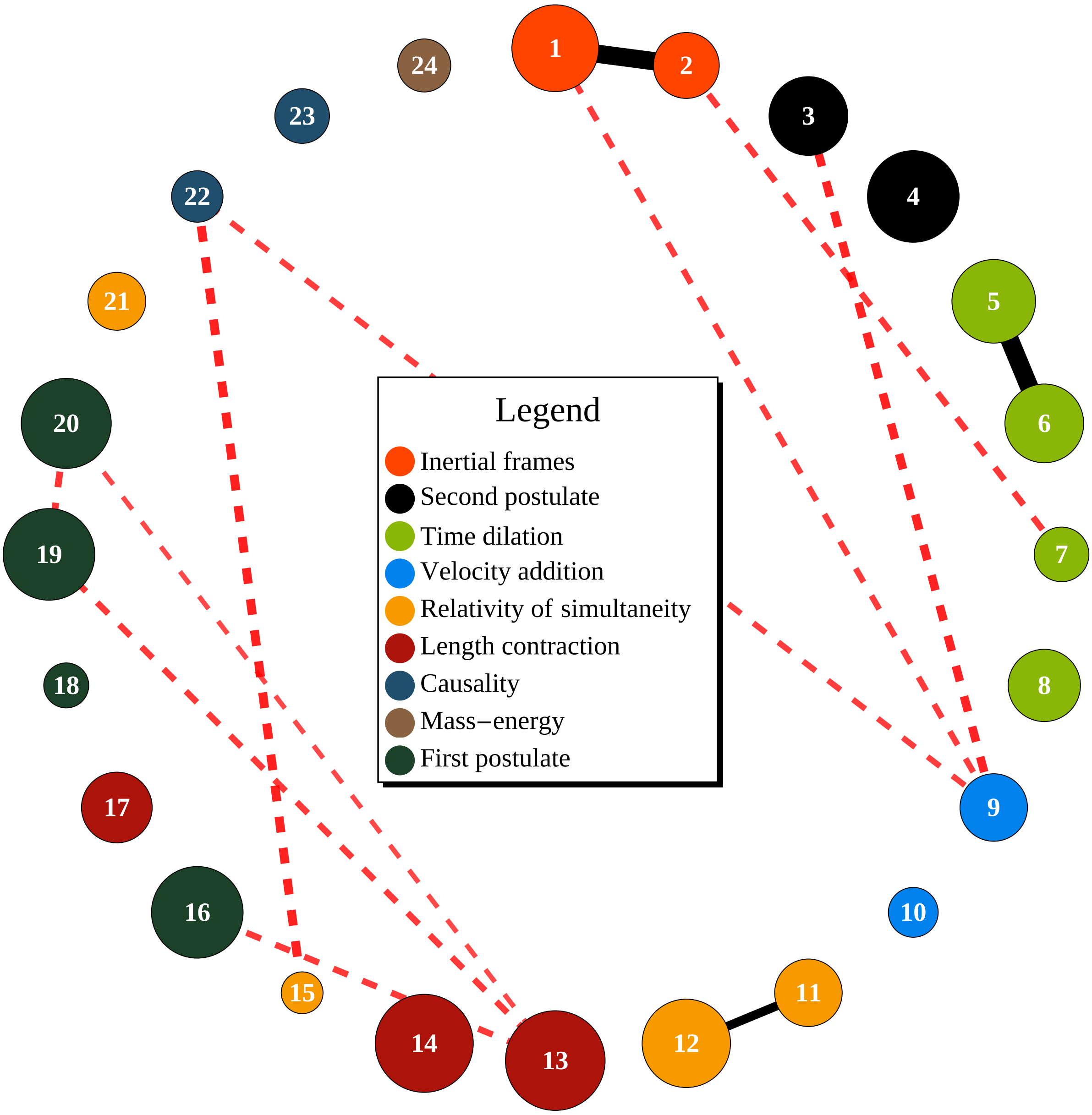


Figure 6.16: Graph connecting questions whose answers are significantly correlated. The red, dashed connections are the statistically significant "raw" correlations, and the black, solid connections are the (positive) correlations that remain after correcting for student ability with the Rasch model. The thickness of the connections is a function of the strength of correlation, but is non-linear, to exaggerate the differences. The radius of each question node is proportional to the number of students that got it correct; smaller nodes represent harder questions. The colours of each node represent which concept group they pertain to.

	Females	Males	Total
N_{pre}	19	51	70
N_{post}	18	45	63
$N_{pre \cap post}$	15	38	53

Table 6.9: Gender proportions in pre-test and post-test. The class consisted of 28 females and 71 males in total.

Following every item on the RCI, students were asked to self-assess their confidence in their response on a Likert scale. Males tended to rate their confidence higher on average, in both pre-test and post-test. This result is consistent with that reported by Sharma et al. [69], but inconsistent with the result of Lawrie et al., in which women were found to be more likely to be overconfident than men [41]. Recall that the p-values represent the probability that there is no gender effect and this result was produced by chance:

	Females	Males	p-value
$\bar{c}_{pre}$	0.41	0.53	0.02
$\bar{c}_{post}$	0.64	0.70	0.039

Table 6.10: Comparison of average confidence $\bar{c}$ for males and females, in pre-test and post-test. The difference in confidence between males and females diminished from pre-test to post-test, but remained statistically significant.

In addition to this, a statistically significant difference in the test scores of males and females was observed, in both pre-test and post-test.

	Females	Males	Class mean	p-value
$\bar{x}_{pre}$	0.50	0.58	0.56	0.022
$\bar{x}_{post}$	0.63	0.75	0.72	0.0030

Table 6.11: Comparison of average RCI score for males and females, in pre-test and post-test.

However, as has been noted, the primary measure of learning is not raw score, but normalised gain between pre- and post-tests, defined as $g = \frac{x_{post} - x_{pre}}{1 - x_{pre}}$:

	Females	Males	p-value
$\bar{g}$	0.23	0.38	0.047

Table 6.12: Comparison of mean normalised gain for males and females.

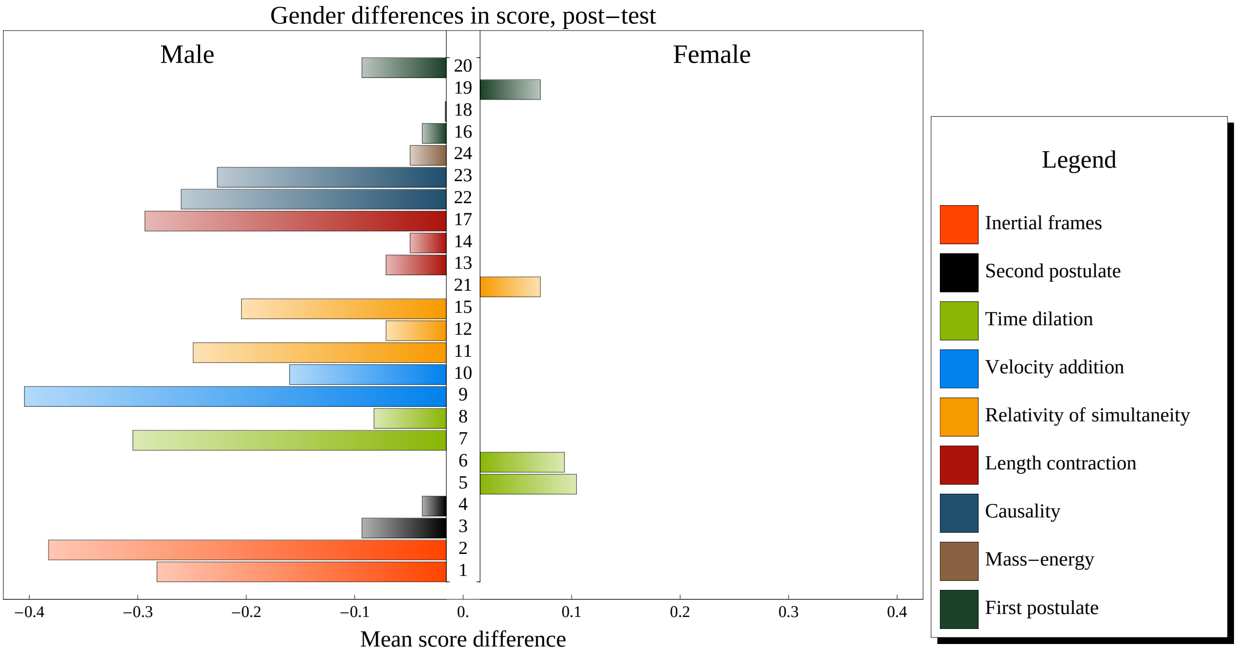


Figure 6.17: Post-test score differences, post-test. The length of the bars are equal to $|\bar{x}_{male} - \bar{x}_{female}|$, for each question.

This is a large disparity! A natural question to ask is whether there are a small subset of the RCI questions that are producing this effect, or whether the apparent gender bias is uniform across the RCI. With a small number of exceptions, the disparity in performance is spread out over all the RCI questions (see figure 6.17). Of note is that females out-perform males on the time dilation pair, questions 5 and 6, although this effect is not statistically significant. Using the Kolmogorov-Smirnov test, the only questions with statistically significant ($p \leq 0.05$) gender differences are questions 1, 2, 9, and 17 (see table 6.13).

Question	**Concept**	$\boldsymbol{\Delta x}$	**p-value**
1	Inertial frames	0.27	0.013
2	Inertial frames	0.37	0.0088
9	Velocity addition	0.39	0.0065
17	Length contraction	0.28	0.038

Table 6.13: RCI questions with statistically significant gender differences in the post-test.

Questions1 & 2 involve dropping a bowling ball out of a train, which is a similar context to item 14 on the Force Concept Inventory, a question which has been noted to exhibit gender differences [73]. We have no hypothesis for why questions 9 and 17 exhibit a gender difference.

This gender difference is not reproduced anywhere else in the course: there is no statistically significant gender difference in any other scores, so we can eliminate them as confounding factors. We outline this analysis below.

Prior academic achievement

A good measure of a student's academic ability going into university is their University Admissions Index (UAI), a percentile ranking that is based on their academic performance in the last two years of high school. There is no significant difference between the UAIs of the male and female students in the class:

	Females	**Males**	**p-value**
$\overline{\text{UAI}}$	94.2	93.5	0.96

Table 6.14: Mean UAI, males and females.

This lack of difference remains the same, regardless of how one partitions the class (students that did pre-test only, post-test only,or both pre-test and post).

Course assessment

Gender biases are conspicuously absent in all other special relativity assessments in the course. In the special relativity component of the mid-semester exam, the scores are not significantly different:

	Females	**Males**	**p-value**
$\boldsymbol{\bar{x}_{pre,exam}}$	0.65	0.64	0.97
$\boldsymbol{\bar{x}_{post,exam}}$	0.66	0.67	0.95

Table 6.15: Mean exam scores for students present at pre-test, and post-test.

Likewise, the relativity homework has no gender bias:

	Females	**Males**	**p-value**
$\boldsymbol{\bar{x}}$	0.75	0.75	0.998

Table 6.16: Mean relativity homework scores, males and females.

Discussion

A significant gender bias exists in the RCI, which is not apparent in other course assessments. The fact that no gender bias was discovered in the relativity homework or exam cannot rule out that there is a gender difference in the understanding of special relativity, though it suggests that it is unlikely. The alternative hypothesis, that it is a property of the concept inventory itself, and not of special relativity, that is producing the effect, is plausible, given that gender biases have been noted in other concept inventories, most notably the FCI [73], and the TUG-K [62]. This aspect of the RCI is puzzling, and an avenue for further research with larger samples.

6.9 New misconceptions

Here we outline previously undiscussed student misconceptions that were discovered in the course of our investigation:

Asymmetric time dilation

There was a disparity in performance between question 5 and question 6 on the pre-test (see figure 6.19 below), significant at the 5% level. Student performance in these two questions was anticorrelated ($r_{56,pre} = -0.25$, significant at the 2% level). Our interpretation of this result is that there existed a belief that time dilation effects are “asymmetric” or “reciprocal” - if A measures B’s clocks running slow, then B must measure $A's$ clocks running *fast*. This misconception would seem to be compatible with the “absolute reference frame” misconception explored by Panse et al., since it would seem to imply the belief that clocks that are “absolutely moving” will run slow, while clocks that are “absolutely stationary” will run “normally”, and hence *fast* in comparison. This “asymmetric time dilation” misconception was observed to exist (from student’s words, written and oral) in several different contexts:

1. Real Time Relativity lab. A representative encounter with a student in the process of exploring the RTR simulation (see Figure 6.2). The student’s ship is flying towards a clock at constant velocity, and so the readout on the clock is ticking faster than the proper time on the ship, due to the relativistic Doppler effect - a consequence of optics. The student is asked to explain the readout on the clock. They respond:

   > *“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...”*

   The student’s comment concisely summarises the connection between the “absolute rest frame” misconception and the “asymmetric time dilation” misconception.

2. Homework problem. Students are asked to write an open-answer response, explaining the situation in figure 6.18 to a sceptical friend who believes that there is a contradiction in the assertion that the two spaceships each observe the other’s clocks to be ticking at half the rate of their own clock. One student admitted that they couldn’t resolve the apparent contradiction and wrote in their submission:

   > *“It would seem to me that if both ships were travelling at about 86% of the speed of light then time for each would be equally slowed by* [a factor of] *two, and that for each ship one tick relates to two ticks of proper time and so their perception of each other would be the same as if they were travelling behind each other with the same separation. I cannot see how changing the direction of travel would change this.”*

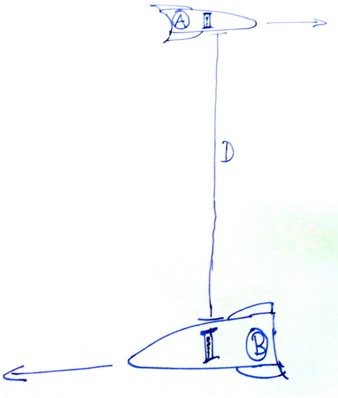


Figure 6.18: Symmetric time dilation homework problem. Two spaceships are travelling at uniform velocity relative to each other, and are observing each other's clocks as in the picture shown. The distance D between them is large, so that for the duration of the experiment, it is approximately constant. The result is that the astronauts on each ship measure the clocks on the other ship to be running slowly compared to their own clocks, due to the symmetry of their relative motion.

My interpretation of this response is that the student is analysing the situation from some "objective" frame, in which both ships are moving at equal speed, and hence their clocks should be ticking at the same rate. This would seem to be consistent with the "absolute rest frame" misconception.

In the following two questions, Abbey is in a spaceship moving at high speed relative to Brendan, who is standing on an asteroid (a very small piece of rock floating in space). She flies past him so that at $t = 0$, she is momentarily adjacent to Brendan.

5. At the instant that Abbey's ship passes Brendan, she sends two light pulses to him from her ship. If the light pulses are emitted a nanosecond (10^{-9} seconds) apart according to Abbey's clock, what will be the time interval between the pulses according to Brendan?

   (a) Greater than one nanosecond

   (b) Equal to one nanosecond

   (c) Less than one nanosecond

6. Also while Abbey's ship passes Brendan, Brendan sends two light pulses to Abbey. If Brendan sends the light pulses a nanosecond (10^{-9} seconds) apart according to his clock, what will be the time interval between the pulses according to Abbey?

   (a) Greater than one nanosecond

   (b) Equal to one nanosecond

   (c) Less than one nanosecond

Figure 6.19: Questions 5 and 6 from the post-test RCI.

Pre-test	**5 Correct**	**5 Incorrect**	**Post-test**	**5 Correct**	**5 Incorrect**
6 Correct	11	13	**6 Correct**	46	3
6 Incorrect	33	13	**6 Incorrect**	6	8

Table 6.17: 2x2 contingency table for questions 5 and 6, pre-test (left), and post-test (right)

Importantly, the RCI time dilation pair went from being anti-correlated in pre-test, to the *most* correlated pair of questions in the post-test, with $r_{56,post} = 0.559$, statistically significant at the 100% level. This means that even though overall performance on question 6 did not improve as much as the length contraction questions (13 and 14), it became tightly linked with question 5, which is the intention - a true relativist will give the same answer for 5 as they do in 6, because of the symmetry of the situation. A question to ask is, given that the correlation gain is so strong, why is the performance gain not as strong? Some students are answering symmetrically but incorrectly (see table 6.17).

We don't draw any hard conclusions from this result, as the number of students in each "bin" is low - all of the following results are suggestive only. The larger of the two cross-terms in both pre-test and post-test is "5 correct, 6 incorrect", and there are a moderate number getting both 5 and 6 wrong, explaining the high correlation but relatively low final score. Let's probe further: what are the actual answer choices of the incorrect students?

- 6 of the 8 students that got both wrong answered *symmetrically*: "Less than 1 ns, Less than 1 ns"

- All 6 students that got 5 right and 6 wrong answered *asymmetrically*: "Greater than 1 ns, Less than 1 ns"

- All 3 students that got 6 right and 5 wrong also answered *asymmetrically*: "Less than 1 ns, Greater than 1 ns"

There are a few suggestive things to note, and which may be followed up with future work:

- Alternative b "Equal to 1 ns" was only chosen twice, which suggests that the Galilean idea of absolute time has been displaced by instruction. It remains to see whether the ideas that have replaced it are correct or not!

- It seems reasonable to interpret the asymmetric answers (cross-terms) as exhibiting the asymmetric time dilation misconception, although interviews would have been needed to confirm or deny this. If this is indeed the case, then it is interesting that "Greater than 1 ns, Less than 1 ns" is the more common of the two. This is suggestive of the "absolute rest frame" misconception - since Brendan is "standing on an asteroid" and Abbey is "in a spaceship moving at high speed", these students use the "reciprocal time dilation" idea outlined earlier.

- The 6 students that answered wrongly but symmetrically ("Less than 1 ns, Less than 1 ns") present a wonderful question: given that they answered symmetrically, our assumption is that this indicates that they are thinking relativistically. Are they misinterpreting the question, or are they confused about whether time *dilates* or *contracts*?

Asymmetric length contraction

A reasonable corollary of the hypothesis that asymmetric time dilation is connected with the "absolute rest frame" misconception would be that students with the asymmetric time dilation misconception will exhibit an analogous misconception in the case of length contraction. This misconception exists, but its link with the equivalent time dilation misconception is unclear. From a student's response in the mid-semester exam (the question asks what length the length of a platform is in the reference frame of a speeding train - see appendix C):

> *"The length of the platform will appear longer in Amanda's frame. This is because it has the opposite effect as when Bryan observes Amanda's arms [...] because Amanda is travelling at a higher velocity."*

On the same exam, another question involves Bryan measuring the distance between the impacts of two balls dropped with a separation D and simultaneously in the rest frame of Amanda, who is on a speeding train. The students are encouraged to use the Lorentz Transformations, and almost all students did this. A counter-intuitive result is that although Amanda's armspan is reduced to $\frac{D}{\gamma}$ in Bryan's reference frame, the relativity of simultaneity means that the separation between the positions where Amanda *releases* the balls is in fact γD. Some students, after correctly applying the Lorentz transformations and getting the correct answer, attributed the result to length contraction:

> *"The distance measured by Bryan is greater ... because of the inverse length contraction."*

> *"The reciprocal length contraction means that Bryan will say that the balls impacted further apart than Amanda."*

This is a tricky question. Many students didn't consider the relativity of simultaneity at all:

> *"It would seem logical that Bryan would see the distance between the balls when they touch the ground to be the same as when Amanda was holding them."*

There was no length contraction question pair on the RCI, so these misconceptions could not be probed with the instrument in its post-test state. This is a possible area of improvement - adding a "symmetric length contraction" question - although we refrained from adding unvalidated questions in the final iteration.

Chapter 7

Conclusions

Special relativity is on the fringes of most people's conceptual understanding; this work aimed to bring it within the reach of far more people than in the past. I have drawn together the sparse special relativity education literature, and focused it into a comprehensive and validated conceptual survey, the Relativity Concept Inventory. This concept inventory was based on a careful review of existing knowledge of student misconceptions, and with significant input from international experts on special relativity and relativity education. A poll of these experts revealed the concepts that are most highly valued by teachers, and these concepts are well represented on the RCI. I have used this concept inventory, in conjunction with other assessments, to identify hitherto undiscussed student learning difficulties in special relativity thus significantly extending the prior research on student thinking in special relativity.

This project has also extended the methodological toolkit for physics education research. In particular, we argue that collection and analysis of students' self-assessed confidence in their responses can provide important information about their thinking. We also find that Monte Carlo techniques can be used to add rigour to the statistical analysis, in the case of the item-item correlations, where it was particularly appropriate to concern ourselves with statistical significance. Compared with much of the previous physics education research, this project is rigorous in terms of the strength of evidence that we required before making firm conclusions.

The findings of this thesis are to be presented by the author at the 20th Australian Institute of Physics Congress in December 2012.

7.1 Summary of main results

We draw the preceding results and analysis together to form the following conclusions:

Our rigorous statistical analysis of the item-item correlations reveals that there are three item-item correlations that are "robust", i.e. that remain even after student ability is accounted for. These question pairs are: $(1, 2)$ (inertial frames), $(5, 6)$ (time dilation), $(11, 12)$ (relativity of simultaneity). This is strong evidence that these questions are in fact testing the same concept, and a reasonable conclusion is that they are conforming with the intended concept grouping with which the test was designed. Strangely, after correcting for student ability with the Rasch model, two statistically significant anti-correlations emerge, that were previously masked in the raw correlation data. These were the question pairs $(7, 8)$ and $(23, 24)$; e have no plausible explanation for these anti-correlations. However, other aspects of the analysis indicate that questions 7 and 24 are anomalous in other respects. As a result, we put questions 7 and 24 up for review, and we make our justifications for their inclusion and removal, respectively, in section 7.2.

The rest of the validity analysis gave generally positive results. The RCI's relativity of simultaneity questions demonstrated a promising correlation with performance on the exam, although a larger sample size would be needed to make this result rigorous. The classical test analysis yielded results that were mostly within the normally accepted ranges for a concept inventory, with two important exceptions: the mean total pre-test score, and the mean post-test item discrimination. We put forward possible reasons for why these two measures are outside normal bounds, and why this not a matter of great concern at the moment:

- Taking part in the RCI study was not compulsory for students. We have established that the subset of the class that participated were significantly better students (as measured by the mid-semester exam) than those that didn't, so the RCI results represent a biased sample of the class. If the whole class had participated in the RCI study, it is likely that the pre-test mean score would have been lower.

- The class we studied was quite accomplished on average, with a median UAI of 95.3 for those students whose UAIs we had on record. Although pre-test results were not significantly correlated with prior formal instruction in special relativity, it was informally observed that many students that had not had prior instruction had nevertheless informed themselves about special relativity to an extent, out of their own interest. This may also account for the high pre-test scores.

- As for the low mean item discrimination, we know that the main source of low item discriminations are the questions 12, 13, 14, and 24. It is recommended that question 24 be removed from future versions of the RCI. We argue that the low discriminations in the other three questions is due to their being some of the easiest questions on the test, and thus represent concepts in which student learning was more or less uniform across ability ranges.

Further analysis of the RCI results revealed a statistically significant gender difference between males and females, with males outperforming females on both the pre-test and post-test. Normalised gain, the standard measure of student learning during instruction, also favoured males. A statistically significant gender difference in student confidence was also recorded. Gender differences in concept inventories have previously been documented, although no rigorous conclusions have yet been drawn from those investigations [73, 72]. This result remains a puzzle.

We also used student self-assessed confidence to differentiate between questions that exhibited a positive confidence-performance correlation, and those that exhibited a negative confidence-performance correlation. Those with negative correlations were indicative of either strong misconceptions, or students guessing the right answer - the specific cases were determined on closer inspection. This is an important extension of a nascent analysis methodology, which is increasingly being used in concept inventory analysis [20, 41].

Finally, we used the RCI, in conjunction with other assessment, to discover a previously undiscussed misconception, which we label "asymmetric time dilation". An analogous misconception for the case of length contraction was also discovered, although without involving the RCI directly.

7.2 Final iteration of the RCI

The last stage of our iterative process is a procedure of attrition - we ask if there are any items that can be justifiably removed. The inventory was developed with a broader scope than is necessary, so as to experiment with some concepts that had not been previously tested in any previous special relativity research - some of these experimental questions turned out to not work as intended. Based on our extensive analysis of the post-test results, we can put forward several items that are candidates to be dropped:

- Question 10 showed no correlation with question 9, and this lack of correlation was highly significant. This suggests that the question isn't testing the velocity addition concept, but something else instead. Because it performed well in other areas (good discrimination and point-biserial coefficient - see appendix D.4), there aren't compelling reasons to get rid of it.

- Question 7 showed a strong, negative correlation between student confidence and performance, and, after correcting for student ability, was anti-correlated with question 8 - a perplexing result, given both questions are nominally in the "time dilation" group. However, it probes an important aspect of time dilation in a new context, and for that reason, we consider it to be a valuable question, and left it in the final version.

- Questions 13 and 14 showed very low discrimination, but this is a due to the fact that in the post-test, the proportion of the class that got them correct was almost unity. These are certainly easy questions, but that is not an argument against their inclusion; if student performance on these questions was low, it would be an indicator of very ineffective teaching.

- We removed question 24 (mass-energy equivalence), which was a problematic question in several ways. It has a discrimination of zero (see appendix D.4), it was the only question to exhibit a negative normalised gain, and, after correcting for student ability, it turned out that there was a statistically significant anti-correlation between question 24 and question 23. Given that there was some doubt among the experts we surveyed as to whether the concept should be included or not (see appendix D.1), its inclusion was always tentative and experimental.

7.3 Limitations of the research and suggestions for further work

According to best practice in the field, this type of project should ideally takes a number of years, which would allow extensive use of student interviews, and a lengthier iteration process. This project, by contrast, was compressed into the space of a few months, and so the scope of the student interviews was limited. Nevertheless, our development process has reached an important milestone by synthesising a working concept inventory and accompanying analysis methods. Administering the RCI as a pre-test and post-test, as we have done, and then following up with student interviews would further validate and refine the RCI.

The next logical step would be to disseminate this version to other universities for further testing, with larger sample sizes. Aside from Aaron Titus, who administered the RCI as a post-test to his class at High Point University, a number of experts who responded to the expert survey expressed interest in using the RCI to evaluate their own teaching.

Some interesting results that are possible precursors to further work:

- It is interesting that there was an apparent disparity in the performance of students with respect to length contraction and time dilation, given that these are both the "staple" concepts of relativity. An issue with questions relating to time dilation and relativity of simultaneity, is to not conflate these with "light delay". All of the relativity of simultaneity questions explicitly state that the observers compensate for this, and some effort was gone to in questions 5 and 6 to set up the problem so that the Doppler effect wasn't an issue. It is possible that these efforts are complicating the questions for some students; perhaps we have been over-careful with the light-delay/Doppler effect issue, and a simpler question would get at the issue just as well.

- The gender effect in the RCI is puzzling, and fits into a history of unexplained gender effects in concept inventories. A first step would be to see if the result is reproduced in future uses of the RCI. A possible next step could be to attempt to rule out certain factors that might be discriminating based on gender, such as:

  - Whether or not it is assessed for course grade. This could be done by using subsets of the RCI questions in quizzes, exams, or other assessments, and checking if the gender effect is reproduced.
  - Whether or not it is multiple choice. This could be done by creating open-ended analogues of the RCI questions, and giving them in quizzes, tutorials, or exams, depending on whether or not they should also be assessed for course grade.
  - Whether the question contexts (trains, spaceships, etc.) are having an effect. It would be possible to transform the contexts of selected RCI questions, and to check whether there is any effect, although this would require a paired study, and so would necessitate a large sample size.

- The rigorous statistical analysis identified three pairs of questions that are robustly correlated (i.e. this correlation remained, even after correcting for student ability). A natural extension of this work is to check whether this result is reproduced with a different class, and/or different teacher. In addition, the anti-correlation between questions 7 and 8 that appeared after correcting for ability is curious - student interviews might clarify what is going on here.

- The analysis of confidence-score correlations revealed some questions in which students were conspicuously overconfident. Student interviews would be able to further probe these questions, to determine whether or not students are exhibiting strong misconceptions in their overconfidence, or if the question is merely "tricky".

- Open-ended RCI questions and interviews would serve to enhance the overall validity checking, and would be a natural focus for the next study, to complement the more quantitative focus of our analysis.

7.4 Recommendations for teachers of special relativity

We list here some recommendations for educators, based on our results:

1. In order for students to emerge from an introductory physics course thinking like true relativists, it is essential that they recognise the symmetry of relative motion. This research has made it apparent that some students leave the course harbouring the "asymmetric time dilation/length contraction" misconceptions, which are a key indicator that they have missed the point. This should be a key focus for educators.

2. A goal in teaching special relativity is to displace students' intuitive beliefs about space and time. These beliefs are often implicit, but strong, and instruction often only superifically changes student thinking - students adjust their mental model in the way that minimises cognitive dissonance, and this often results in their twisting the results of relativity to fit into their existing mental model. In particular, we have observed a tendency for some students to attribute relativistic effects to optics, or perception. One way to address this is by not only emphasising the "intelligent observer" advocated by Scherr, but also in dealing with relativistic optics explicitly. This can be achieved by using the Real Time Relativity software, and associated teaching materials [14].

Bibliography

[1] Eric Mazur. *Peer Instruction: A user's manual.* Prentice Hall, New Jersey, 1997.

[2] Edward F. Redish and Richard N. Steinberg. Teaching physics: Figuring out what works. *Physics Today*, 52(1):24–30, 1999.

[3] Carl E. Wieman. Why not try a scientific approach to science education? *Change Magazine*, 39(5):9–15, 2007.

[4] Nathaniel Lasry, Noah Finkelstein, and Eric Mazur. Are most people too dumb for physics? *The Physics Teacher*, 47(7):418–422, 2009.

[5] Frederick Reif. Scientific approaches to science education. *Physics Today*, 39(11):48–54, 1986.

[6] Edward F. Redish. Implications of cognitive studies for teaching physics. *American Journal of Physics*, 62(9):796–803, 1994.

[7] David Hammer. More than misconceptions: Multiple perspectives on student knowledge and reasoning, and an appropriate model for education research. *American Journal of Physics*, 64(10):1316–1325, 1996.

[8] Lillian C. Mcdermott. How we teach and how students learn - A mismatch? *American Journal of Physics*, 61(4):295–298, 1993.

[9] Frederick Reif. Guest Comment: Standards and measurements in physics – Why not in physics education? *American Journal of Physics*, 64(6):687–688, 1996.

[10] David Hestenes, Malcolm Wells, and Gregg Swackhamer. Force Concept Inventory. *The physics teacher*, 30(3):141–158, 1992.

[11] Richard R. Hake. Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, 66(1):64–74, 1998.

[12] Wendy K. Adams and Carl E. Wieman. Development and validation of instruments to measure learning of expert-like thinking. *International Journal of Science Education*, 33(9):1289–1312, 2011.

[13] Various Authors. HSC physics stage 6 syllabus. Technical report, N.S.W. Board of Studies, Sydney, 2010.

[14] Craig M. Savage, Dominic McGrath, Tim McIntyre, Margaret Wegener, and Michael Williamson. Teaching physics using virtual reality. Technical report, Australian Teaching and Learning Council, 2009.

[15] Peter W. Hewson. A case study of conceptual change in special relativity: the influence of prior knowledge in learning. *European Journal of Science Education*, 4(1):61–78, 1982.

[16] Rachel E. Scherr, Peter S. Shaffer, and Stamatis Vokos. Student understanding of time in special relativity: Simultaneity and reference frames. *American Journal of Physics*, 69(S1):S24–S35, 2001.

[17] Kevin Gibson. *Special relativity in the classroom.* PhD thesis, Arizona State University, 2008.

[18] David P. Maloney, Thomas L. O'Kuma, Curtis J. Hieggelke, and Alan Van Heuvelen. Surveying students' conceptual knowledge of electricity and magnetism. *American Journal of Physics*, 69(7):S12–S23, 2001.

[19] Chandralekha Singh. Assessing student expertise in introductory physics with isomorphic problems. I. Performance on nonintuitive problem pair from introductory physics. *Physical Review Special Topics - Physics Education Research*, 4(1):1–9, March 2008.

[20] Kirk Allen, Teri Reed-Rhoads, and Robert Terry. Work in progress: Assessing student confidence of introductory statistics concepts. In *Proceedings. Frontiers in Education. 36th Annual Conference*, pages 13–14. Ieee, 2006.

[21] Douglas Huffman and Patricia Heller. What does the Force Concept Inventory actually measure? *The Physics Teacher*, 33(3):138–143, 1995.

[22] Lillian C. Mcdermott and Edward F. Redish. Resource letter on physics education research. *American Journal of Physics*, 67(9):755–767, 1999.

[23] Paul Ramsden. *Learning to teach in higher education.* RoutledgeFalmer, Abingdon, 2nd edition, 2007.

[24] Frederick Reif. Millikan Lecture 1994: Understanding and teaching important scientific thought processes. *American Journal of Physics*, 63(1):17–32, 1995.

[25] Charles Henderson and Melissa H. Dancy. Impact of physics education research on the teaching of introductory quantitative physics in the United States. *Physical Review Special Topics - Physics Education Research*, 5(2):020107, 2009.

[26] Lin Ding, Ruth Chabay, Bruce Sherwood, and Robert J. Beichner. Evaluating an electricity and magnetism assessment tool: Brief Electricity and Magnetism Assessment. *Physical Review Special Topics - Physics Education Research*, 2(1):1–7, March 2006.

[27] Stephanie V. Chasteen, Steven Pollock, Rachel E. Pepper, and Katherine K. Perkins. Thinking like a physicist: A multi-semester case study of junior-level electricity and magnetism. *American Journal of Physics*, 80(10):923–930, 2012.

[28] Steve Goldhaber, Steven Pollock, Mike Dubson, Paul Beale, and Katherine Perkins. Transforming upper-division quantum mechanics: Learning goals and assessment. *AIP Conference Proceedings*, 1179:145–148, 2010.

[29] S. B. McKagan, Katherine Perkins, and Carl E. Wieman. The design and validation of the quantum mechanics conceptual survey. *arXiv preprint.*

[30] Kirk Allen, Andrea Stone, Teri Reed-Rhoads, and Teri J. Murphy. The Statistics Concepts Inventory : Developing a valid and reliable instrument. In *Proceedings, American Society for Engineering Education Annual Conference & Exposition*, 2004.

[31] Michelle K. Smith, William B. Wood, and Jennifer K. Knight. The genetics concept assessment: A new concept inventory for gauging student understanding of genetics. *Life Sciences Education*, 7(Winter):422–430, 2008.

[32] Tony Wright and Susan Hamilton. Diagnostic assessment for the biological sciences: development of a concept inventory. Technical report, Australian Learning and Teaching Council, 2011.

[33] Ibrahim Abou Halloun and David Hestenes. The initial knowledge state of college physics students. *American Journal of Physics*, 53(November):1043–1048, 1985.

[34] Richard R. Hake. The impact of concept inventories on physics education and its relevance for engineering education. *www.physics.indiana.edu/˜hake*, 2011.

[35] Richard N. Steinberg and Mel S. Sabella. Performance on multiple-choice diagnostics and complementary exam problems. *The Physics Teacher*, 35(3):150–155, 1997.

[36] Rachel E. Scherr. Modeling student thinking: An example from special relativity. *American Journal of Physics*, 75(3):272–280, 2007.

[37] John Stewart, Heather Griffin, and Gay Stewart. Context sensitivity in the force concept inventory. *Physical Review Special Topics - Physics Education Research*, 3(1):1–6, February 2007.

[38] David Hestenes and Ibrahim Halloun. Interpreting the Force Concept Inventory: A response to Huffman and Heller. *The Physics Teacher*, 33(8):502–506, 1995.

[39] Patricia Heller and Douglas Huffman. Interpreting the Force Concept Inventory: A reply to Hestenes and Halloun. *The Physics Teacher*, 33(8):503–511, 1995.

[40] Ibrahim Abou Halloun and David Hestenes. The search for conceptual coherence in FCI data. *Unpublished preprint*, 1996.

[41] Gwen Lawrie, Madeleine Schultz, and William Macaskill. Relationships between confidence, gender, high school performance, a concept inventory, and success in first year chemistry. In *Proceedings, Australian Conference on Science and Mathematics Education*, page 22, 2012.

[42] Terry F. Scott and Daniel Schumayer. Exploratory factor analysis of a Force Concept Inventory data set. *Physical Review Special Topics - Physics Education Research*, 8(2):020105, 2012.

[43] Richard R. Hake. Lessons from the physics education reform effort. *Conservation Ecology*, 5(2):28, 2002.

[44] Rachel E. Scherr, Peter S. Shaffer, and Stamatis Vokos. The challenge of changing deeply held student beliefs about the relativity of simultaneity. *American Journal of Physics*, 70(12):1238–1248, 2002.

[45] Rachel E. Scherr. *An investigation of student understanding of basic concepts in special relativity.* PhD thesis, University of Washington, 2001.

[46] Albert Einstein. On the electrodynamics of moving bodies. *Annalen der Physik*, 17:891, 1905.

[47] Edwin F. Taylor and John Archibald Wheeler. *Spacetime physics: introduction to special relativity.* W. H. Freeman & Co., New York, 2nd edition, 1992.

[48] S. Panse, J. Ramadas, and A. Kumar. Alternative conceptions in Galilean relativity: frames of reference. *International Journal of Science Education*, 16(1):63–82, 1994.

[49] J. Ramadas, S. Barve, and A. Kumar. Alternative conceptions in Galilean relativity: inertial and non-inertial observers. *International Journal of Science Education*, 18(5):615–629, 1996.

[50] A. Villani and J. L. A. Pacca. Students' spontaneous ideas about the speed of light. *International Journal of Science Education*, 9(1):55–66, 1987.

[51] Edith Saltiel and J. L. Malgrange. 'Spontaneous' ways of reasoning in elementary kinematics. *European Journal of Physics*, 1(2):73–80, 1980.

[52] Nathaniel David Mermin. *It's about time.* Princeton University Press, New Jersey, 2005.

[53] Albert Einstein. Does the inertia of a body depend on its energy content? *Annalen der Physik*, 18:639, 1905.

[54] George J. Posner, Kenneth A. Strike, Peter W. Hewson, and William A. Gertzog. Accommodation of a scientific conception: Toward a theory of conceptual change. *Science Education*, 66(2):211–227, 1982.

[55] Olivia Levrini and Andrea DiSessa. How students learn from multiple contexts and definitions: Proper time as a coordination class. *Physical Review Special Topics - Physics Education Research*, 4(1):1–18, April 2008.

[56] Sean Carroll. *Spacetime and geometry.* Addison-Wesley, San Francisco, 1st edition, 2004.

[57] Gary Oas. On the abuse and use of relativistic mass. *ArXiv preprint physics/0504110*, 2005.

[58] Lev B. Okun. The concept of mass. *Physics Today*, 42(6):31–36, 1989.

[59] Gamze Sezgin Selçuk. Addressing pre-service teachers' understandings and difficulties with some core concepts in the special theory of relativity. *European Journal of Physics*, 32(1):1–13, January 2011.

[60] Chin Long Chiang. *Statistical methods of analysis.* World Scientific Publishing, Singapore, 1st edition, 2003.

[61] Lin Ding and Robert J. Beichner. Approaches to data analysis of multiple-choice questions. *Physical Review Special Topics - Physics Education Research*, 5(2):1–17, September 2009.

[62] Robert J. Beichner. Testing student interpretation of kinematics graphs. *American Journal of Physics*, 62(8):750–762, 1994.

[63] William H. Press, Brian P. Flannery, Saul A. Tukolsky, and William T. Vetterling. *Numerical recipes: the art of scientific computing.* Cambridge University Press, New York, 1st edition, 1986.

[64] Dennis D. Wackerly, William III Mendenhall, and Richard L. Scheaffer. *Mathematical statistics with applications.* Thomson, Belmont, 7th edition, 2008.

[65] Mark H. Moulton. Rasch Demo Spreadsheet.

[66] Harry H. Harman. *Modern factor analysis.* University of Chicago Press, Chicago, 3rd edition, 1976.

[67] Craig M. Savage, Anthony Searle, and Lachlan McCalman. Real Time Relativity: Exploratory learning of special relativity. *American Journal of Physics*, 75(9):791–799, 2007.

[68] Lin Ding, Neville Reay, Albert Lee, and Lei Bao. Effects of testing conditions on conceptual survey results. *Physical Review Special Topics - Physics Education Research*, 4(1):2–7, June 2008.

[69] Manjula Sharma and James Bews. Self-monitoring: Confidence, academic achievement and gender differences in physics. *Journal of Learning Design*, 4(3):1–13, 2011.

[70] Susan Ramlo. Validity and reliability of the force and motion conceptual evaluation. *American Journal of Physics*, 76(9):882–886, 2008.

[71] Robert C. MacCallum, Keith F. Widaman, Shaobo Zhang, and Sehee Hong. Sample size in factor analysis. *Psychological Methods*, 4(1):84–99, 1999.

[72] R. D. Dietz, R. H. Pearson, M. R. Semak, and C. W. Willis. Gender bias in the Force Concept Inventory? In *AIP Conference Proceedings 1413: Physics Education Research Conference*, pages 171–174, 2011.

[73] Laura McCullough. Gender, context, and physics assessment. *Journal of International Women's Studies*, 5(4):20–30, 2004.

[74] Joseph C. Hafele and Richard E. Keating. Around-the-World Atomic Clocks: Observed Relativistic Time Gains. *Science*, 177(4044):168–170, 1972.

Appendix A

Relativity Concept Inventory

This is the version of the RCI that was used in the post-test, and differs from our final version only in its inclusion of question 24, which we recommend be removed in future studies using the RCI (see section 7.2).

Instructions:

- *Some of the questions are multiple choice, with an additional confidence scale similar to the example below. For each of these questions, circle the answer that you agree most with, and mark on the scale how confident you are in your choice.*

Rate how confident you are in your answer:

○·········○·········○·········○·········○

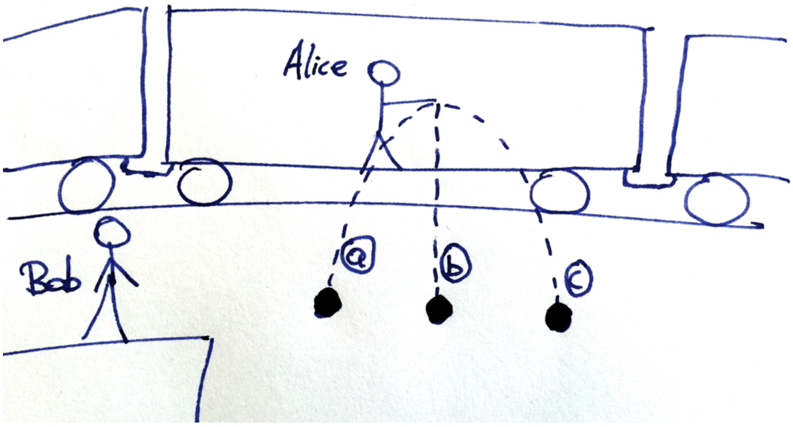
guessing unconfident neutral confident certain

- *Some of the questions are in the form of statements with which you may agree or disagree. Circle the response that most closely corresponds to your position on the question.*

- *In all of the following questions, the symbol c represents the speed of light in a vacuum,* 3×10^8 *m/s.*

- *Answer all of the questions to the best of your knowledge.*

In the following two questions, Alice is standing in a train moving at velocity v from **left to right** relative to Bob, who is standing on a platform. As Alice passes Bob, she drops a bowling ball out of the train's window:



1. Ignoring air resistance, which path of the ball would Bob observe, standing on the platform?

   (a) Path (a)

   (b) Path (b)

   (c) Path (c)

Rate how confident you are in your answer:

○ · · · · · · · · · ○ · · · · · · · · · ○ · · · · · · · · · ○ · · · · · · · · · ○

guessing unconfident neutral confident certain

2. Ignoring air resistance, which path of the ball would Alice observe, standing in the train?

   (a) Path (a)

   (b) Path (b)

   (c) Path (c)

Rate how confident you are in your answer:

○ · · · · · · · · · ○ · · · · · · · · · ○ · · · · · · · · · ○ · · · · · · · · · ○

guessing unconfident neutral confident certain

3. True or false: "In principle, it is possible for an observer following a pulse of light at a constant high speed to observe the light to be almost stationary."

   (a) True

   (b) False

Rate how confident you are in your answer:

○·········○·········○·········○·········○

guessing unconfident neutral confident certain

4. Consider a spaceship travelling from Earth towards a distant star at a constant high velocity v relative to Earth. The spaceship sends a light pulse back to Earth. On Earth, the speed of this pulse is measured to be:

   (a) c

   (b) $c + v$

   (c) $c - v$

Rate how confident you are in your answer:

○·········○·········○·········○·········○

guessing unconfident neutral confident certain

In the following two questions, Abbey is in a spaceship moving at high speed relative to Brendan, who is standing on an asteroid (a very small piece of rock floating in space). She flies past him so that at $t = 0$, she is momentarily adjacent to Brendan.

5. At the instant that Abbey's ship passes Brendan, she sends two light pulses to him from her ship. If the light pulses are emitted a nanosecond (10^{-9} seconds) apart according to Abbey's clock, what will be the time interval between the pulses according to Brendan?

   (a) Greater than one nanosecond

   (b) Equal to one nanosecond

   (c) Less than one nanosecond

Rate how confident you are in your answer:

○·········○·········○·········○·········○

guessing unconfident neutral confident certain

6. Also while Abbey's ship passes Brendan, Brendan sends two light pulses to Abbey. If Brendan sends the light pulses a nanosecond (10^{-9} seconds) apart according to his clock, what will be the time interval between the pulses according to Abbey?

   (a) Greater than one nanosecond

   (b) Equal to one nanosecond

   (c) Less than one nanosecond

Rate how confident you are in your answer:

○·········○·········○·········○·········○

guessing unconfident neutral confident certain

7. It is known that our galaxy is of the order of 100,000 light-years in diameter. True or false: “Travelling at a constant speed that is less than, but close to, the speed of light, in principle it is possible for a person to cross the galaxy within their lifetime.”

   (a) True

   (b) False

Rate how confident you are in your answer:

○·········○·········○·········○·········○

guessing unconfident neutral confident certain

8. The Olympic Games is a two-week long sports competition. An interested alien astronomer watches the Olympics from a distant planet moving at high speed relative to Earth. *If the alien were to compensate for the time the light from Earth takes to reach them*, they would measure the length of the Olympics to be:

   (a) Greater than two weeks

   (b) Equal to two weeks

   (c) Less than two weeks

Rate how confident you are in your answer:

○·········○·········○·········○·········○

guessing unconfident neutral confident certain

In the following two questions, the scenario is as follows: Alex and his friend Bianca decide to set off on separate voyages in identical spaceships. They each speed away from Earth in opposite directions - Alex at $v = 0.75c$ to the left, and Bianca at $v = 0.75c$ to the right, relative to an observer on Earth.

9. If Alex measures the rate at which his distance to Bianca is increasing, he will obtain a value that is:

   (a) Equal to $1.5c$

   (b) Greater than c but less than $1.5c$

   (c) Equal to c

   (d) Greater than $0.75c$ but less than c

   (e) Equal to $0.75c$

Rate how confident you are in your answer:

○·········○·········○·········○·········○

guessing unconfident neutral confident certain

10. If Cameron, an observer on Earth, measures the rate at which the distance between Alex and Bianca is increasing, he will obtain a value that is:

(a) Equal to $1.5c$

(b) Greater than c but less than $1.5c$

(c) Equal to c

(d) Greater than $0.75c$ but less than c

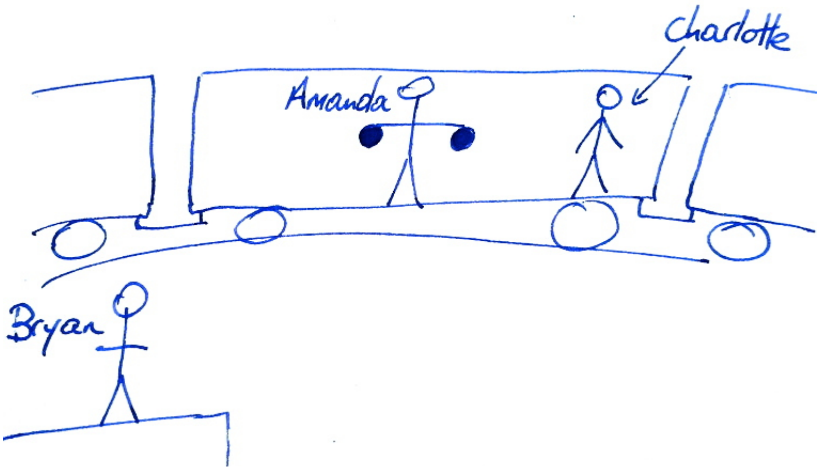
(e) Equal to $0.75c$

Rate how confident you are in your answer:

○ · · · · · · · · · ○ · · · · · · · · · ○ · · · · · · · · · ○ · · · · · · · · · ○

guessing unconfident neutral confident certain

In the following four questions, Amanda is standing on a train travelling at high speed past Bryan, who is standing on a platform. As she passes Bryan, she drops two bowling balls out of the window at the same time (Amanda's time), and from an arm's span apart.



11. Bryan stands on the platform and watches the balls fall to the ground. *If he compensates for the time that the light from the impacts takes to reach him*, in what order does Bryan measure the balls hitting the ground?

(a) At the same time

(b) One ball before the other

Rate how confident you are in your answer:

○ · · · · · · · · · ○ · · · · · · · · · ○ · · · · · · · · · ○ · · · · · · · · · ○

guessing unconfident neutral confident certain

12. Charlotte is another passenger on the train with Amanda. *If she compensates for the time that the light from the impacts takes to reach her*, in what order does Charlotte measure the balls hitting the ground?

(a) At the same time

(b) One ball before the other

Rate how confident you are in your answer:

○·········○·········○·········○·········○

guessing unconfident neutral confident certain

13. Amanda has an arm span of D meters at rest. If Bryan performs a measurement of Amanda's arm span as she passes him, he will obtain a value:

    (a) Greater than D

    (b) Equal to D

    (c) Less than D

Rate how confident you are in your answer:

○·········○·········○·········○·········○

guessing unconfident neutral confident certain

14. Amanda also has a height of H meters at rest. If Bryan performs a measurement of Amanda's height as she passes him, he will obtain a value:

    (a) Greater than H

    (b) Equal to H

    (c) Less than H

Rate how confident you are in your answer:

○·········○·········○·········○·········○

guessing unconfident neutral confident certain

15. Two separate light bulbs emit flashes of light, distant from an observer. This observer receives the light from both flashes at the same time. From this alone it is possible to conclude that:

    (a) The flashes occurred at the same time for all observers

    (b) The flashes occurred at the same time for the observer at that location

    (c) The flashes occurred at the same time if the observer is not moving relative to the light bulbs

    (d) It is not possible to make any of the above conclusions

Rate how confident you are in your answer:

○·········○·········○·········○·········○

guessing unconfident neutral confident certain

16. In the following thought experiment, you are in a high speed train travelling along a railway. True or false: "If you measure the dimensions of the train compartment, you will obtain different values than if the train were at rest."

    (a) True
    (b) False

Rate how confident you are in your answer:

○·········○·········○·········○·········○

guessing unconfident neutral confident certain

17. Consider a futuristic space station that specialises in constructing fast spaceships. Once the ships are built, they leave the station at high speed for testing. As they leave the station at speed, a serial number is stamped instantaneously on the side of the ship by a machine on the station. This serial number has length D as measured by a builder on the space station. After the ship has finished its test run, it returns to the station and is parked in the garage. What is the length of the serial number now, as measured by the builder on the space station?

    (a) Greater than D
    (b) Equal to D
    (c) Less than D

Rate how confident you are in your answer:

○·········○·········○·········○·········○

guessing unconfident neutral confident certain

18. Adam is in a spaceship moving at $v = 0.99c$ relative to our galaxy. Adam wants to measure the mass of his ship by observing how resistant the ship is to acceleration. If Adam exerts a force on the ship (by turning on a rocket engine, for example) and measures (with an accelerometer inside the ship) the acceleration that results, he will obtain a value that is:

    (a) Greater than what he would measure if his ship were at rest relative to the galaxy.
    (b) Equal to what he would measure if his ship were at rest relative to the galaxy.
    (c) Less than what he would measure if his ship were at rest relative to the galaxy.

Rate how confident you are in your answer:

○·········○·········○·········○·········○

guessing unconfident neutral confident certain

19. In the following thought experiment, you are in a high speed train travelling along a railway. True or false: "If you measure the rate at which your watch is ticking, you will obtain a different value than if the train were at rest."

(a) True

(b) False

Rate how confident you are in your answer:

○·········○·········○·········○·········○

guessing unconfident neutral confident certain

20. You are in a well equipped physics lab without windows or ways of interacting with the outside world. It is known that the lab is in uniform motion. How do you determine the velocity of the lab?

(a) You throw a ball across the lab and measure its change in velocity

(b) You shine a laser beam across the lab and measure its change in velocity

(c) Either (a) or (b)

(d) It is not possible to determine the lab's velocity by experiment

Rate how confident you are in your answer:

○·········○·········○·········○·········○

guessing unconfident neutral confident certain

21. You observe a set of distant, spatially separated clocks that are synchronised in their rest frame. You are at rest relative to the clocks, and you observe (through a telescope) that the times read on the clocks are different. This is due to:

(a) Time dilation

(b) Length contraction

(c) Relativity of simultaneity

(d) None of the above

Rate how confident you are in your answer:

○·········○·········○·········○·········○

guessing unconfident neutral confident certain

22. If two events are separated in such a way that an observer can be present at both events, which relationship(s) between the two events are the same for all observers?

(a) The time between the two events

(b) The distance between the two events

(c) The order in which the events occur

(d) None of these relationships are the same for all observers

Rate how confident you are in your answer:

○·········○·········○·········○·········○

guessing unconfident neutral confident certain

23. If two events are separated in such a way that **no** observer can be present at both events, which relationship(s) between the two events are the same for all observers?

    (a) The time between the two events

    (b) The distance between the two events

    (c) The order in which the events occur

    (d) None of these relationships are the same for all observers

Rate how confident you are in your answer:

○ · · · · · · · · · ○ · · · · · · · · · ○ · · · · · · · · · ○ · · · · · · · · · ○

guessing unconfident neutral confident certain

24. Consider a closed box, containing an equal amount of matter and antimatter. The total mass of this box and its contents is initially M. The matter and antimatter are then allowed to annihilate inside the box, turning into photons in the process. What is the total mass of the box and its contents *after* the annihilation?

    (a) Greater than M

    (b) Equal to M

    (c) Less than M

Rate how confident you are in your answer:

○ · · · · · · · · · ○ · · · · · · · · · ○ · · · · · · · · · ○ · · · · · · · · · ○

guessing unconfident neutral confident certain

Appendix B

Expert Survey

At the Physics Education Centre at the Australian National University, we are developing a "Relativity Concept Inventory", which is a survey instrument designed to probe student understanding of basic concepts in special relativity. The purpose of this instrument is to help academics to improve the quality of their teaching in special relativity.

The purpose of this survey is to collect expert opinion in the field - we want to get your feedback on which concepts in special relativity are important to teach (and hence to test) for a first course in special relativity. This will help us to arrive at a consensus on which concepts are relevant and appropriate to include in the conceptual survey.

On the following page is a list of concepts we propose to include on the test. Please indicate whether or not you think each of these concepts are relevant and/or appropriate. Additional space for commentary is provided. At the end of the survey is a comment box where we would like you to suggest any concepts that you believe are relevant but that don't feature on our list. If you wish to see a draft of our concept inventory and make further suggestions for improvements, please leave your name and email address.

This survey should take no more than 10 minutes. In addition, you can return to the survey to update your responses after submitting, if need be. Thank you for participating!

Contact information:
John Aslanides (Honours student)
Department of Quantum Science and Physics Education Centre
The Australian National University
E: u4520779@anu.edu.au
T: 02 6125 1156

Expert opinion: Which concepts in special relativity should be included in a concept inventory?

1. The first postulate: the laws of physics are the same in all inertial reference frames.

   (a) Agree

   (b) Neutral

   (c) Disagree

2. The second postulate: the speed of light in a vacuum is the same in all inertial reference frames.

   (a) Agree

   (b) Neutral

   (c) Disagree

3. Time dilation

   (a) Agree

   (b) Neutral

   (c) Disagree

4. Length contraction

   (a) Agree

   (b) Neutral

   (c) Disagree

5. The relativity of simultaneity

   (a) Agree

   (b) Neutral

   (c) Disagree

6. Inertial reference frame: a coordinate system in which a free particle will move at constant velocity - in particular, the concept that all inertial frames are equivalent.

   (a) Agree

   (b) Neutral

   (c) Disagree

7. Velocity addition: Velocities transform between frames such that no object can be observed travelling faster than the speed of light in a vacuum.

   (a) Agree

   (b) Neutral

   (c) Disagree

8. Events are independent of reference frame: if X happens in one reference frame, then X happens in all reference frames (distinct from the first postulate).

   (a) Agree

   (b) Neutral

   (c) Disagree

9. Causality: if two events are time-like separated, then the ordering of the events is fixed for all inertial reference frames.

   (a) Agree
   (b) Neutral
   (c) Disagree

10. Non-inertial frames: Specifically the concept that the acceleration of one observer breaks the symmetry between two observers in relative motion - a key insight for resolving the 'twin paradox'.

   (a) Agree
   (b) Neutral
   (c) Disagree

11. Mass-energy equivalence

   (a) Agree
   (b) Neutral
   (c) Disagree

12. Invariance of the space-time interval

   (a) Agree
   (b) Neutral
   (c) Disagree

13. Invariance of rest mass

   (a) Agree
   (b) Neutral
   (c) Disagree

14. The operational definition of time interval and space interval measurements

   (a) Agree
   (b) Neutral
   (c) Disagree

15. Please list any other concepts that you believe are relevant/appropriate.

16. If you would like your input to be acknowledged in John's thesis, please state your name and university in the box below.

17. If you would like to see a copy of our draft concept inventory to give more feedback, please type your email address in the box below and we will email you a copy.

Appendix C

Mid-semester exam

The mid-semester exam was administered on the last day of the first teaching period, after the class had done three weeks of electromagnetism and three and a half weeks of special relativity. The exam was 2 hours long, with two equal sections on electromagnetism and special relativity, and was worth 10% of the overall course grade. Below we present the special relativity section.

In the following you may use the Lorentz transformation formulas:

$$t' = \frac{t - xv/c^2}{\sqrt{1 - \frac{v^2}{c^2}}}, \quad x' = \frac{x - vt}{\sqrt{1 - \frac{v^2}{c^2}}} \tag{C.1}$$

And the inverse Lorentz transformation formulas:

$$t = \frac{t' + x'v/c^2}{\sqrt{1 - \frac{v^2}{c^2}}}, \quad x = \frac{x' + vt'}{\sqrt{1 - \frac{v^2}{c^2}}} \tag{C.2}$$

Amanda is standing in a train that is moving at a constant high speed v from left to right with respect to Bryan, who is standing on a platform. Amanda lets two bowling balls drop out of the train window from an arm's span apart, and simultaneously, in her reference frame. Let Amanda be at rest in the S' frame, and let Bryan be at rest in the S frame (standard configuration) and let E_L be the event "Left ball drops" and E_R be "Right ball drops". Let Amanda drop the balls just as her left hand is level with Bryan, so that the coordinate origins for S and S' coincide at the event E_L. Amanda's arm span at rest is D meters. The length of Bryan's platform at rest is L meters. Express all of your answers in terms of D and L, and show all of your working.

To get you started, here are the coordinates for the events E_L and E_R in Amanda's reference frame: $t'_L = 0$, $t'_R = 0$, $x'_L = 0$, $x'_R = D$.

Each ball falls for t'_{fall} seconds before hitting the ground, according to Amanda's stopwatch. Ignore air resistance.

(a) (3 marks)

How long do the balls take to fall in Bryan's frame? Is this time logner, shorter, or the same as that measured by Amanda? Give the physical reason why this is the case.

(b) (1 mark)

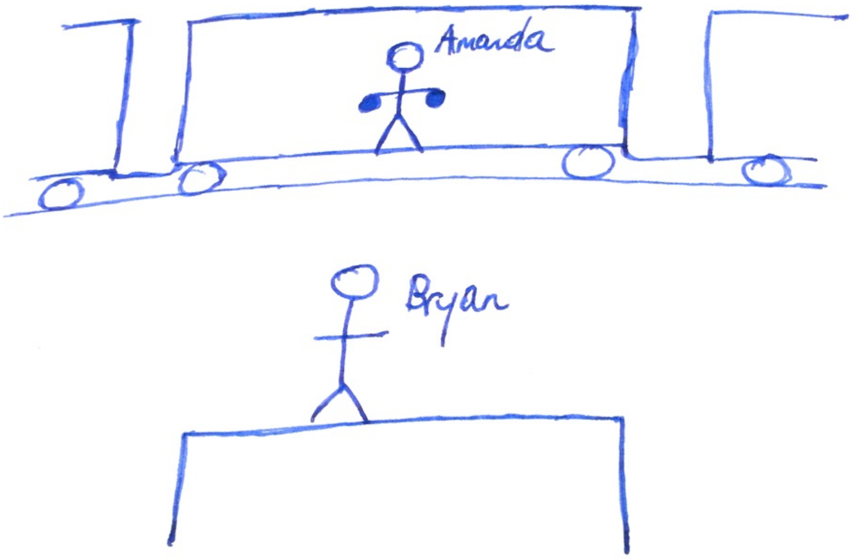


Figure C.1: Exam diagram.

What is the the time between the ball impacts in Amanda's frame?

(c) (3 marks)
What is the time between the impacts in Bryan's frame? Is this time difference longer, shorter, or the same as that measured by Amanda? Give the physical reason why this is the case.

(d) (1 mark)
If Bryan measures Amanda's arm span as she passes, what will he measure?

(e) (3 marks)
If Amanda measures the length of Bryan's platform as she passes, what will she measure? Is this longer, shorter, or the same as the length of the platform as measured by Bryan? Give the physical reason why this is the case.

(f) (1 mark)
What is the distance between the impacts in Amanda's frame?

(g) (3 marks)
What is the distance between the impacts in Bryan's frame? Is this distance longer, shorter, or the same as that measured by Amanda? Give the physical reason why this is the case. Think about this one carefully!

(h) (8 marks)
The relativity of simultaneity is a key result in relativity. It is involved in understanding the train and lightning video shown in class, and in the resolutions of the so-called

“pole and barn paradox” and “twins paradox”.

Outline one of these scenarios, or a similar one of your choice, and explain how the relativity of simultaneity resolves any apparent conundrums. In particular, analyse the scenario first from the frame of one observer, and then from the frame of the other. Explain what your analysis implies for the meaning of time measurements in different reference frames. Use equations and at least one diagram in your answer.

(i) (2 marks)

Can the relativity of simultaneity be described simply as a light delay effect, i.e. can the lack of agreement on the simultaneity of separate events for different observers be account for just by signal delays? Explain briefly.

Appendix D

Additional results and data

D.1 Expert comments

Here I hightlight the most pertinent, or most frequent comments, from the well known experts whose opinion carries the most weight, with regard to some of the "controversial" concepts. Commenter names are reproduced with permission.

Consistency (Events are independent of reference frame - if X happens in one reference frame, then X happens in all reference frames)

Reaction from experts was actually quite positive, despite the relatively low amount of agreement (23 out of 31). One question on this concept was included on the pre-test, but was dropped for the post-test. Examples:

- Edwin Taylor: *"VERY important! 'Events are the nails on which physics hangs."'*

- Susan Scott: *"This is the most important assumption of special relativity - the premise from which so many other things follow."*

Non-inertial frames

- Edwin Taylor: *"I asked John Wheeler why we did not cover this topic* [in *Spacetime Physics*]. *He replied, 'It is not very important.' I would omit this in a first treatment."*

- Susan Scott: *"I am in favor of Taylor & Wheeler's treatment of this in Spacetime Physics, in which 'acceleration' is not the issue."*

- Don Koks: *"In an introductory course it's good to confine discussion to inertial frames."*

Invariance of the space-time interval

- Joe Hope: *"Not a conceptual requirement for a first course."*

- Susan Scott: *"I like the Taylor & Wheeler approach, in which the invariance of the spacetime interval is a key premise."*

Since the interval and its invariance is an abstract and fairly technical concept, the RCI doesn't test it.

D.2 High Point University results

Aaron Titus kindly administered the post-test version of the RCI as a post-test to his class at Highpoint University, in the United States. Unfortunately, his class size was too small (3) for any analysis of the results to yield anything significant. However, it is included, because it is some evidence that the RCI is useful and meaningful in different environments - in particular, since the results are reasonable (none of the three student marks were very high or low), this is further evidence that the RCI is set at a reasonable level of difficulty. The mean score was 0.57.

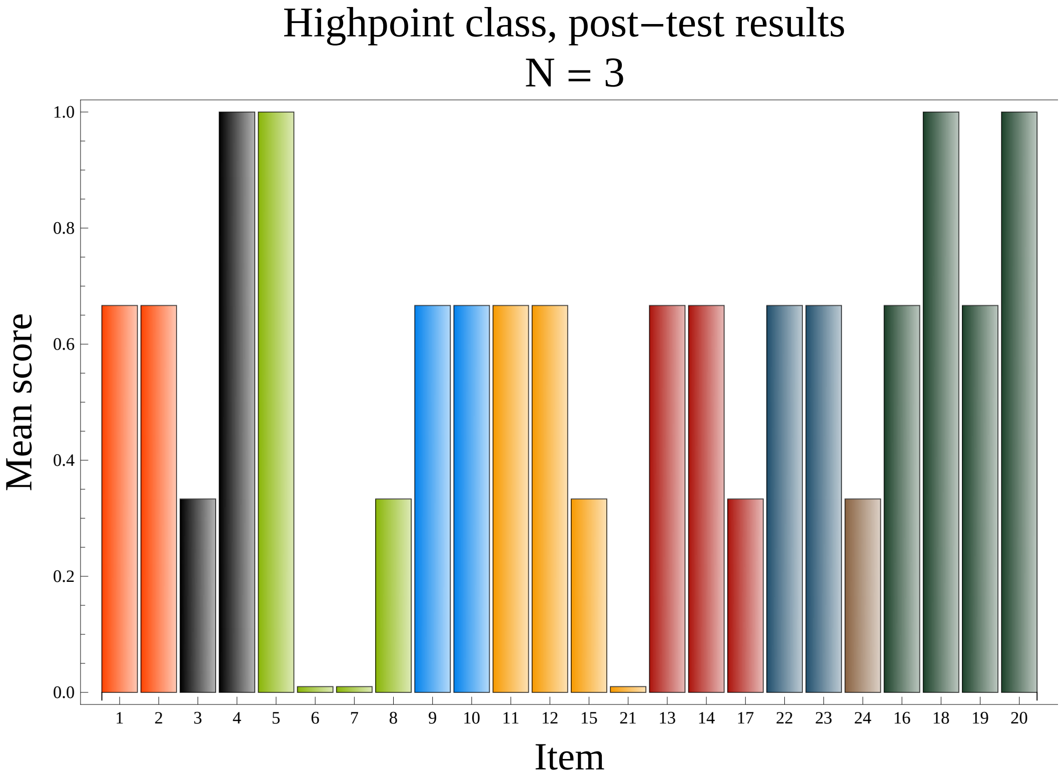


Figure D.1: Item difficulties for Aaron Titus's class, post-test.

D.3 Confidence results

The mean pre-test confidence was 0.50, and mean post-test confidence was 0.68. The correlation between total score and confidence, by student, was 0.25 in the pre-test and 0.30 in the post-test, indicating that in general, the better students are more confident.

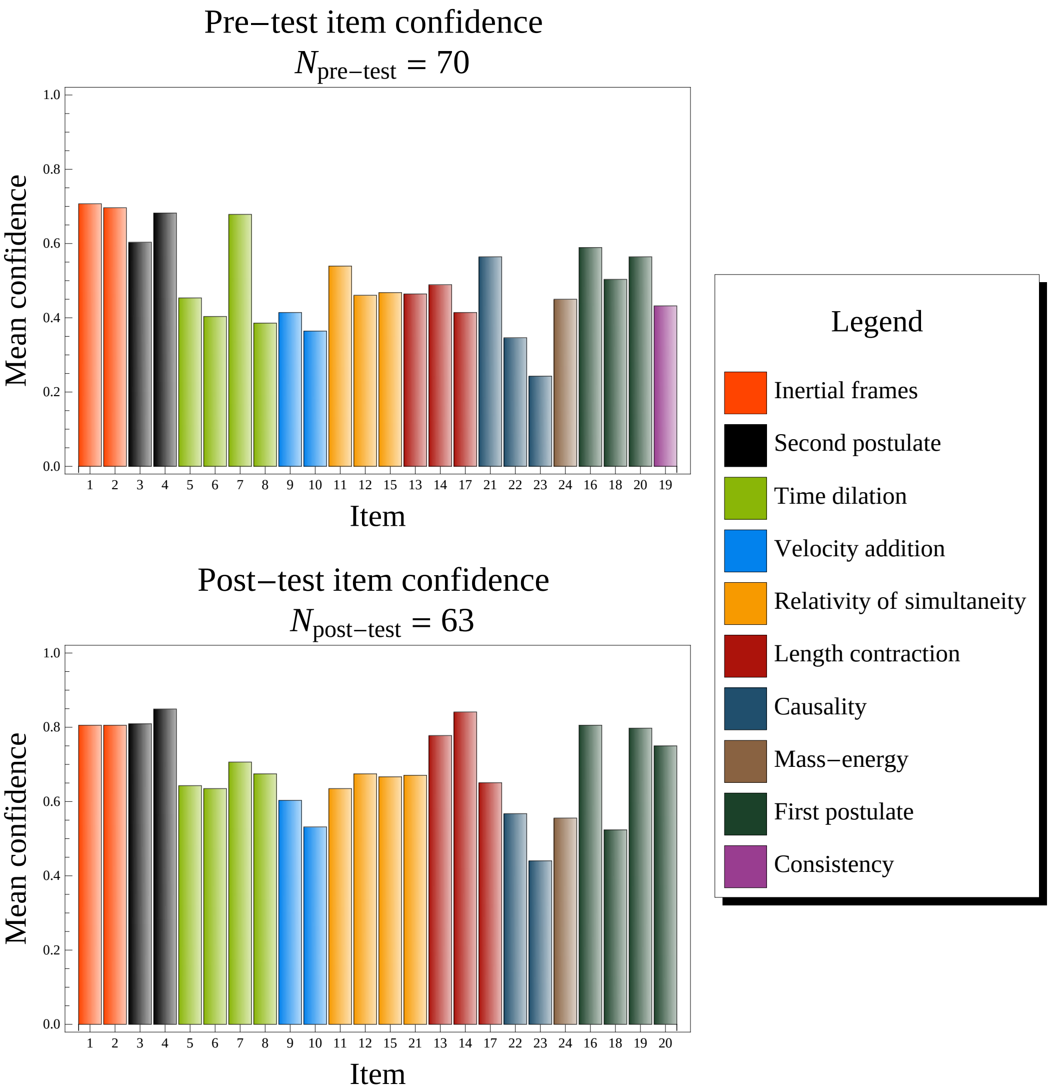


Figure D.2: Student self-assessed confidence, pre-test and post-test.

D.4 Post-test item data and answer key

Here we present the raw data for the post-test RCI, including the item difficulties (mean proportion correct!), discrimination, point-biserial coefficient, and our correct answers, for the use of future researchers.

Question	Discrimination	Point-biserial coefficient	Difficulty	Correct answer
1	0.19	0.32	0.86	C
2	0.38	0.48	0.65	B
3	0.44	0.54	0.78	B
4	0.063	0.32	0.90	A
5	0.13	0.29	0.83	A
6	0.063	0.21	0.78	A
7	0.38	0.46	0.54	A
8	0.25	0.30	0.71	A
9	0.63	0.64	0.67	D
10	0.44	0.39	0.49	A
11	0.19	0.38	0.67	B
12	-0.13	0.060	0.87	A
13	-0.13	0.20	0.98	C
14	-0.13	0.12	0.97	B
15	0.51	0.53	0.41	D
16	0.063	0.39	0.90	B
17	0.25	0.34	0.70	A
18	0.38	0.42	0.44	B
19	0.13	0.34	0.90	B
20	0	0.20	0.89	D
21	0.57	0.50	0.57	D
22	0.51	0.57	0.51	C
23	0.57	0.53	0.54	D
24	0	0.14	0.52	B

D.5 Lecture questions

This data is included for archival purposes. No major conclusions were drawn from them, but they are included for further work.

Non-assessed questions were given in-class, using the poll functionality in the www.piazza.com class forum. Student participation was generally around 80% of those present in the class. Below, we detail the student performance in the questions. The bar in green indicates the correct answer. The students that answered the lectures are a subset of the class, and since we're using this data to estimate the understanding of the class population, we will estimate the standard error with:

$$\Delta x = \sqrt{\frac{(1-x)\,x}{N}} \tag{D.1}$$

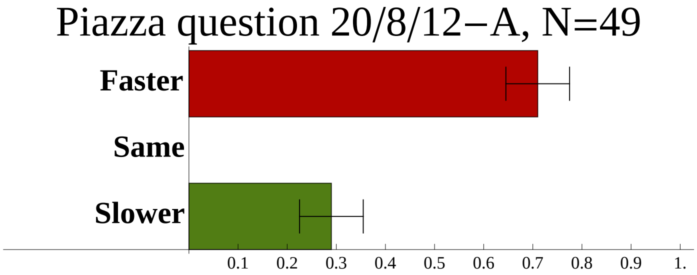
where N is the total number of respondants to the question and $(1-x)\,x$ is an estimator of the variance of x [65].

Another important point to note is that the students that attend lectures are a subset of the class, and the students that attend lectures *consistently* are a smaller subset still. Based on the Piazza response numbers, and given that the class enrolment is $N = 99$, lecture attendance was usually around 50%. It is reasonable to assume that the students that attend lectures are generally more studious, and are getting more out of the course. Thus, the results obtained from the students that respond to Piazza questions in lectures should be taken as an upper bound on the level of understanding of the class as a whole.

20/8/12: Third lecture: questions on length contraction and time dilation.

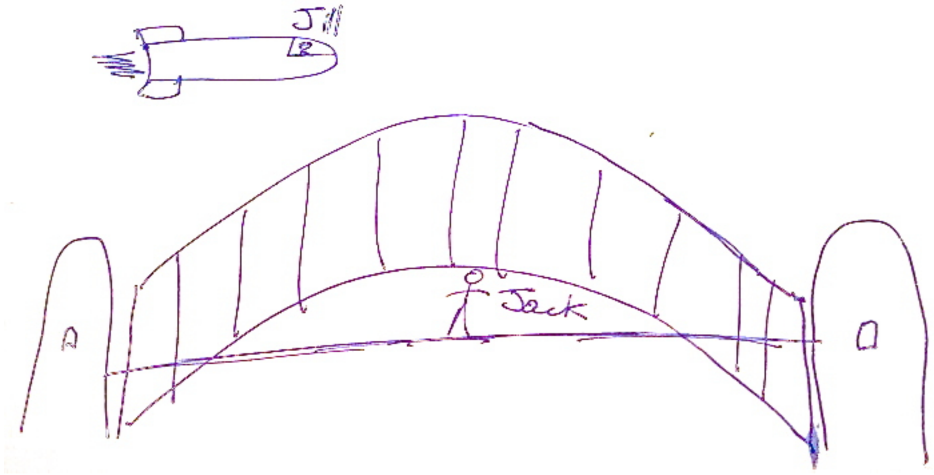
The students were shown a video about the Hafele-Keating experiment [74], in which atomic clocks were flown around the world in commercial jets and compared with clocks on the ground, in an empirical verification of time dilation. Ideally, we would have asked a pair of questions on time dilation analogous to the pair we gave on length contraction, to test asymmetry in time dilation on equal footing - however, because the purpose of the lectures is primarily pedagogical, we had to compromise, so as to not divert the lecture too much - in addition, once we addressed the asymmetry issue explicitly in one situation, presenting a situation that is completely isomorphic to it will not be indicative of students' spontaneous reasoning, since we would have *just* told them the answer - this was already more or less the case in the Sydney Harbour Bridge question, below. **:S**

A. An atomic clock is flown at just under a thousand kilometers per hour relative to the ground. An astronaut in a space station orbits the Earth ten times faster relative to the ground. If they were to measure the rate that the plane's clock is ticking and compare it to their own atomic clock, they would find, according to special relativity:

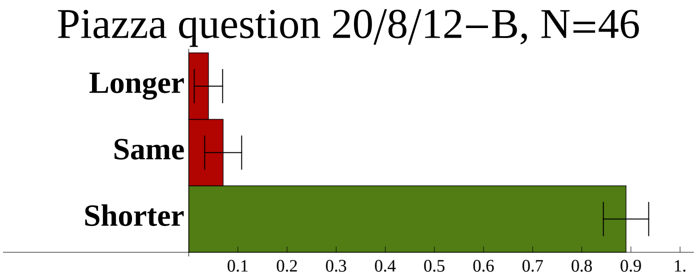


Of course, in this question, gravitational effects are ignored - this was emphasised to the students verbally. The performance is quite poor, with over 70% of the responses indicating that the astronaut would see the plane's clock ticking faster than their own. This can be attributed to one of two things: (1) the students were mis-reading the question, and were making a comparison between the rates of the plane's clock and the astronaut's clock as viewed by an observer on Earth, and may or may not hold the "asymmetric time dilation" misconception, or (2) the students interpreted the question correctly, and they do in fact hold the "asymmetric time dilation" misconception. The symmetry of the situation (ignoring gravity) was emphasised, and the class moved on, to talk about length contraction.

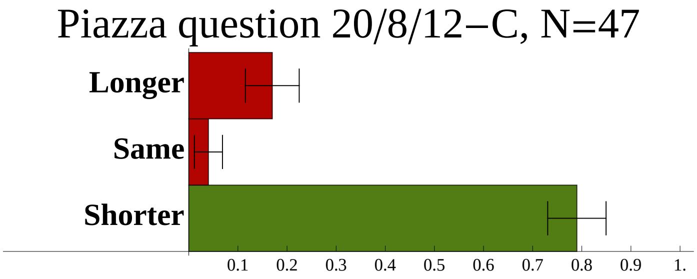
B. Jack is standing on the Sydney Harbour Bridge while Jill flies past in a spaceship at high speed (imagine that for the sake of the thought experiment, Sydney is in a vacuum).



Jill's spaceship has a length of 25 metres in her reference frame. If Jack measures the length of Jill's ship he will obtain a value that is:



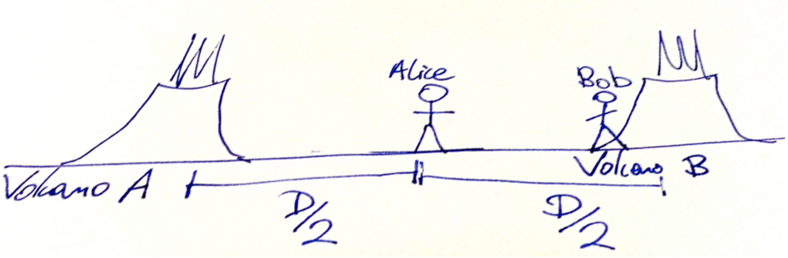
C. The Sydney Harbour Bridge has a length of about 500 metres in Jack's reference frame. If Jill measures the length of the Harbour Bridge she will obtain a value that is:



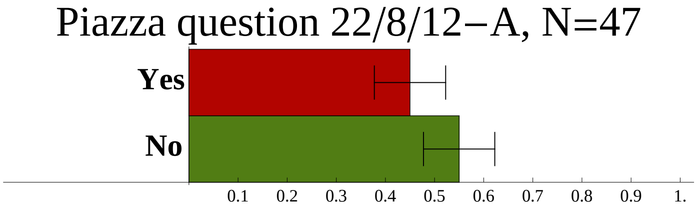
Performance on this question pair was quite good in both instances, with performance slightly worse on the second of the two questions (89% vs 79%). In the absence of finer-grained data, we cannot say what the proportion of students are that got the first of the two right, but the second wrong, or visa-versa.

22/8/12: Fourth lecture: questions on relativity of simultaneity. Distinction between "seeing" and "measuring".

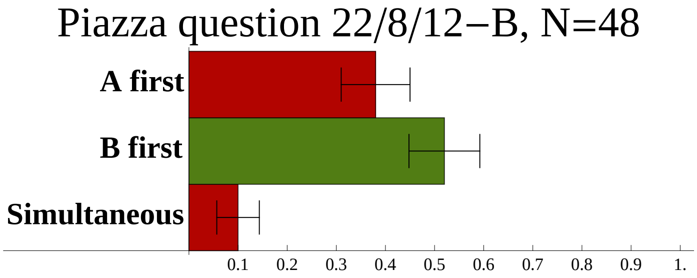
A. Two volcanoes (A and B) are a distance D apart in the Earth's reference frame. A seismologist (Alice) is standing at rest relative to the volcanoes, halfway between them. The two volcanoes erupt, and Alice sees the light from the eruptions at the same time.



A second seismologist (Bob) is standing next to volcano B, and is at rest relative to Alice. Bob doesn't receive the light from each eruption simultaneously. Are the eruptions simultaneous for Bob?



B. Charlie is an astronaut flying past in his spaceship at velocity v relative to the Earth. In the Earth's frame, the ship is directly over volcano A when both volcanos erupt. In what order do the volcanos erupt in Charlie's frame? Lorentz transformation for time: $t' = \gamma\left(t - \frac{vx}{c^2}\right)$.



27/8/12: Fifth lecture: questions on twin paradox.

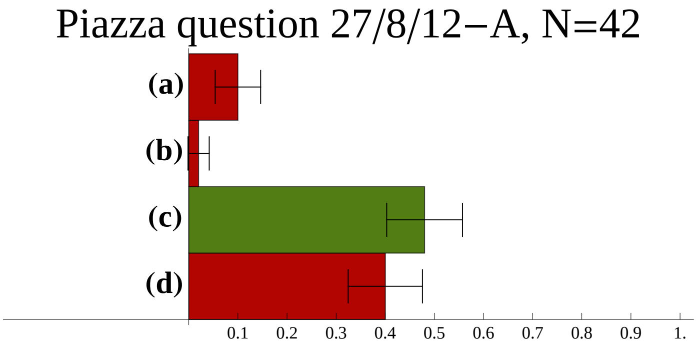
A. In the year 2600 high-speed space travel becomes available to the public. Emirates are trying to sell seats on their new high-speed spaceship. Their sales pitch is that because of time dilation, going on their space flight will prolong your lifespan. What, if anything, is wrong with their argument? *(answer alternatives written out in full):*

(a) Nothing, it's true that travelling on the ship will make you age slower than someone on Earth.

(b) Their argument is the wrong way around - travelling in the ship will actually make you age faster than someone on Earth.

(c) You age no differently than you would if you stayed on Earth.

(d) How much you age depends on how long the spaceship spends accelerating and decelerating.

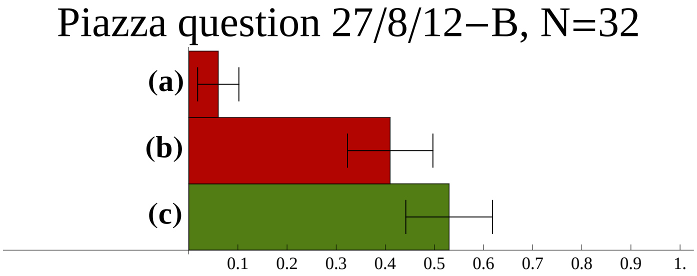


B. The relativity of simultaneity means: *(answer alternatives written out in full):*

(a) Two observers at different positions will measure different times for pairs of distant events, because light moves at a finite speed and takes different amounts of time to reach the different observers.

(b) Two observers moving relative to each other will measure different times for a pair of distant events, because in the time that light from the events takes to reach them, their position has changed.

(c) Neither (a) or (b).



Appendix E

Information statement for students

Study title: Depth of learning in special relativity

Australian National University

1. **What is this study about?**

   This study aims to develop a conceptual survey as a formative assessment to help improve teaching in special relativity. If you choose to participate in this study, you will be helping us create a tool to evaluate the effectiveness of instruction, so that students like you can have an enhanced learning experience.

2. **Who is carrying out the study?**

   The study is being conducted by John Aslanides, who is a fourth year Honours student. This project will be a part of his thesis in physics education research. His supervisor is Prof. Craig Savage, your teacher.

3. **What does the study involve?**

   The study involves responding to questions on a multiple choice conceptual survey both before and after the three-week special relativity topic. The survey amounts to a multiple choice quiz on basic concepts in relativity. You may also complete a questionnaire about your learning, and you may be asked for a voluntary follow-up interview. The surveys will take no more than 20 minutes, and interviews no more than 50 minutes.

4. **How will this affect my grades?**

   None of the surveys or questionnaires are assessable. Data produced from this study will be used to improve the conceptual survey, and is independent of course assessment. This study and the teaching method used in the course are founded on previous physics education research, which has been proven to improve students' learning in physics. Completing the survey will also be good practice for the exam, so participating in this study should help you to learn the material.

5. **What happens to the results?**

All aspects of this survey are confidential, as are your assessment results, in accordance with ANU policy. Any publication of the results of the study will be in aggregate, so individuals won't be identifiable. The only people who will see your results for this part of the course are the researchers (John and Craig).

6. **Can I withdraw from the study?**

Being in this study is completely voluntary - you are under no obligation to consent. If you wish to, you can privately opt out of the study by selecting an option on Wattle.

7. **Where can I get more information?**

If you have any questions about the study, please feel free to contact John Aslanides (u4520779@anu.edu.au or 6125 1156), or Craig Savage (craig.savage@anu.edu.au or 6125 4202). If you have any concerns about the way the research is being done, please contact the Secretary of the Human Research Ethics Committee, Research Office, Chancelry 10B, ANU (human.ethics.officer@anu.edu.au or 6125 3427).