Efficacy of a Blended Learning Mastery Progression Cycle on Student Achievement and Attitude in High School Science

Sam Peter Roberson

This thesis is presented for the Degree of
Doctor of Science Education
of
Curtin University

July, 2020.

Author's Declaration

To the best of my knowledge and belief this thesis contains no material previously published by any other person except where due acknowledgement has been made.

This thesis contains no material which has been accepted for the award of any other degree or diploma in any university.

The research presented and reported in this thesis was conducted in accordance with the National Health and Medical Research Council National Statement on Ethical Conduct in Human Research (2007) — updated March 2014. The proposed research study received human research ethics approval from the Curtin University Human Research Ethics Committee (EC00262), Approval Number # **HRE2017-0265**

Sam Peter Roberson

25th June 2020

Abstract

This study was conducted to examine the effect of a Blended Learning Mastery Progression Cycle (BLMPC) on student achievement and attitude in a High School Physics context, specifically through the use of the Minds on Physics (MOP) application for the formative assessment and corrective activity components of the Mastery Learning cycle.

The sample (N = 199) consisted of mixed gender classes from Year 10 cohorts in a single Queensland high school. Classes were randomly assigned to the treatment or control condition. An experimental pretest–posttest approach was used to measure any changes in students' understanding of the Newtonian Force concept, measured using the Force Concept Inventory (FCI), and Attitudes toward Science, measured using the Test of Science Related Attitudes (TOSRA). All students were exposed to the same initial learning activities; the control group then continued through the course content in a linear manner followed by working through non-personalized revision material, whilst the treatment group completed the relevant MOP module at the end of each subtopic.

Data were analyzed in terms of FCI and TOSRA mean pre- and post-unit scores, the distribution and standard deviation of scores, a t-test comparison of the pre- and post-unit scores, and the FCI normalized change and effect size. When comparing the control and treatment group FCI scores, the latter demonstrated significantly more improvement in the raw score, normalized gain and effect size, demonstrated a larger improvement in all dimensions of the Newtonian Force Concept, and showed greater stability in correct responses from the pre to post unit test. An analysis of TOSRA results showed there was no significant difference between the control and treatment groups.

It was concluded that the use of the MOP platform in a BLMPC led to improvements in understanding of the Newtonian force. These findings indicate that the use of Blended Learning activities as correctives is an effective way of improving students' understanding of the Newtonian Force Concept.

Acknowledgements

I would like to take this opportunity to thank the large number of mentors, colleagues and students who have guided, supported and inspired me throughout the Doctoral process.

Professor Robert Cavanagh's guidance and support have been invaluable throughout, from study design to analysis. I am privileged to be one of his last Doctoral students and am grateful for his continual guidance, feedback and encouragement throughout the research process.

Dr David Henderson's guidance and flexible support has been invaluable in maintaining my progress throughout the project.

Thank you to all my school-based colleagues, past and present, for allowing me access to your ideas and innovation. I am inspired every time I see you teach or interact with students.

A special thank you to the students involved in this study. Your willingness to try something new is a testament to your love of learning.

Finally, thank you to Ruth, without whom this endeavour would not have been possible.

Dedication

To Jen, who set me on the path of learning.

To Paul, whose calm guidance kept me on the path.

To Ruth, whose support and perseverance inspired me to keep moving along the path.

To Elowyn and Isla, whose infectious joie de vivre makes the path worth travelling.

Contents

List of Figures

List of Tables

List of Abbreviations

ACARA	Australian Curriculum and Reporting Authority
ACSSU	Australian Curriculum Standards Science Understanding
BL	Blended Learning
BLMPC	Blended Learning Mastery Progression Cycle
CAI	Computer Aided Instruction
EW5	Eaton and Willoughby 5 Dimension Model
FCI	Force Concept Inventory
HWS6	Hestenes, Wells, & Swackhammer Six-Dimension Model
ICT	Information and Communications Technology
ITS	Intelligent Tutoring Systems
LFM	Learning for Mastery
MBT	Mechanics Baseline Test
ML	Mastery Learning
MLA	Multimedia Learning Activities
MOP	Minds on Physics application
OECD	Organisation for Economic Co-operation and Development
PhET	Physics Educational Technology
PISA	Programme for International Student Assessment
SD	Standard deviation from the mean
SSG5	Scott, Schumayer, & Gray Five-Factor Model
STEM	Science Technology Engineering Mathematics
TOSRA	Test of Science Related Attitudes

Chapter 1: Introduction

1.1 Chapter Overview

This chapter provides an overview of the context of the study, the relevant conceptual frameworks, a statement of the problem, including the purpose of the study, and the research questions and limitations. It concludes with a summary of the structure of the thesis.

1.2 Context of the Study

The achievement and uptake of students in Science, Technology, Engineering, and Mathematics (STEM) subjects at school has recently received a renewed focus (Education Council, 2018). This drive to improve outcomes in the STEM subjects comes from the predicted workforce requirements of a technologically driven economy. However, due to decreasing outcomes and engagement in STEM subjects, concerns have been raised regarding Australia's international positioning in this area (Education Council, 2018). Much attention has been given to the decline in Australian students' performance in international assessments, such as the Organisation for Economic Co-operation and Development (OECD) Programme for International Student Assessment (PISA) rankings for Mathematics and Science, in which Australia's ranking has fallen from 8th of 57 in 2006 to 14th of 72 in 2015 (Thompson et al., 2017). Whilst there are many complex societal, psychological and educational issues that may affect the uptake and achievement of STEM subjects (Aschbacher et al., 2009; Fraser & Kahle, 2007), some of the issues may be related to pedagogical practice in this area. For example, the hierarchical nature of many concepts in STEM subjects means that if students do not develop a foundational understanding, future study becomes less likely to be successful

The proliferation of ICT, and individual student provided devices, has seen an increasing trend driven by a move towards Blended Learning (BL) and the flipped

classroom, where content is delivered asynchronously, in schools (Chandra, 2004). Much research reports the benefits of a Mastery approach to learning (Bloom, 1968; Kulik et al., 1990). More recently, software developers (atomi.com.au, Pearsonmylabandmastering.com, physicsclassroom.com, Mathletics.com.au, mathspathway.com) have begun producing Mastery-based learning programs; however, there is little empirical evidence regarding student performance outcomes in this context.

1.3 Conceptual Underpinning of the Study

This study aims to assess the efficacy of combining the approaches of Mastery and Blended Learning in a Science classroom context.

The term 'Mastery Learning' is frequently used to describe approaches that emphasize sequential learning with regular checks and feedback to students, combined with corrective activities to address the identified gaps in learning (Guskey, 2007). The key aspect of Mastery Learning is that students are required to show mastery of a concept before moving on to the next concept in the topic. Mastery Learning approaches are frequently modelled on Bloom's Learning for Mastery (LFM) four-stage process:

Stage 1 – Group instruction using a range of teaching and learning activities.

Stage 2 – Students complete a formative assessment task.

Stage 3 – Students complete corrective or enrichment activities based on areas identified in Stage 2.

Stage 4 – Students have a further opportunity to demonstrate mastery, with further corrective activities available if required.

It is the directed application of specific corrective activities that characterizes effective Mastery Learning programs (Guskey, 2010). These correctives are learning activities that present the same content as originally taught in the group phase of instruction but in a different format, allowing students to develop their understanding

using a variety of cognitive processes. In this study, the corrective activities were based on Blended Learning approaches.

Blended Learning is a term used to describe approaches to learning activities that use a combination of web-based technologies to accomplish an educational goal through a combination of pedagogical approaches, including some face-to-face instruction (Driscoll, 2002). The increased availability of IT solutions in schools has led to a heightened interest in Blended Learning approaches (Chandra & Fisher, 2009; Crook et al., 2014). In a meta-analysis of over 4,500 studies (Hattie, 2009) it was shown that the most effective use of IT occurs when it is used for a diverse range of teaching strategies, there are multiple learning opportunities, the student has control of the learning process, and feedback is optimized. These features align closely with the tenets of Mastery Learning.

There are limited studies that have rigorously assessed the impact of Blended Learning approaches on student outcomes (Means et al., 2010) and no studies of Mastery Learning approaches in Australian high schools since 2008 (Melbourne Graduate School of Education, 2018). Therefore, a study using the combination of these two approaches, in an Australian high school Science context, fills a gap in the research literature. The theoretical proposition of this study is that the benefits of Mastery Learning approaches are dependent on the directed application of effective corrective activities and that Blended Learning approaches may be an effective means of delivering these correctives due to the alignment of the features of effective Blended Learning activities with the tenets of Mastery Learning approaches.

As this study aims to assess the efficacy of a novel approach, the measurement of the impact of the approach is a key aspect of the research. Efficacy was measured in terms of attainment and attitudes. In this context, the term ‘attainment’ was taken to

mean the ability of students to respond to questions and problems through the application of the concepts studied in the topics.

There is a wide range of methodologies that can be used to measure a student's understanding of Science topics, including: detailed teacher-led questioning, teacher devised assessments, external assessments (e.g., New South Wales High School Certificate and International Competitions and Assessments for Schools exams), and validated instruments (e.g., Mechanics Baseline Test (MBT), Force Concept Inventory (FCI)). For this study, the FCI was used as a pre- and post-unit assessment of students' understanding of the Newtonian concept of force. Concept inventories are research-based assessment instruments that measure conceptual understanding, as opposed to rote learning of definitions or the application of algorithms. The FCI is a 30-item multiple choice inventory developed to measure six conceptual dimensions of the concept of force—Newton's First, Second, and Third Laws, Kinematics, Types of Forces, and Superposition of Forces—with the aim of building an overall picture of a student's understanding of Newtonian concepts.

A student's attitude towards a subject can have a significant impact on their application to study, and hence their achievement (Kind et al., 2007; Osborne et al., 2003). Schommer (1994), amongst others (Kind et al., 2007; Siegel & Ranney, 2003), showed a positive relationship between science attitude and science achievement; hence it is important that any novel approach does not have a negative impact on students' attitudes towards Science. For this study, scales from the Test of Science Related Attitudes (TOSRA) were used to measure students' attitudes towards scientific enquiry (Scale I), enjoyment of Science lessons (Scale E), and the adoption of scientific attitudes, such as open-mindedness and willingness to revise opinions (Scale A). The TOSRA scales were used pre- and post-unit to measure students' attitudes and determine if the approach led to any changes in attitudes.

1.4 Statement of Problem

A predicted increase in the number of STEM-related jobs will require a scientifically literate workforce. Australian students' performance in OECD PISA assessments indicates that this may be a cause for concern (Thompson et al., 2017). It may be that the current underachievement is due to poor pedagogy in the teaching of STEM subjects in Australian high schools; in particular, the hierarchical and sequential nature of Physics subjects means that students must develop mastery of a range of concepts to access the higher order concepts required for further study.

This study used a combination of Mastery and Blended Learning approaches to attempt to address this problem by improving students' understanding of the Year 10 Australian Curriculum and Reporting Authority (ACARA) Physics curriculum. This approach will be referred to as a Blended Learning Mastery Progression Cycle (BLMPC) in this study. The Minds on Physics (MOP) suite of apps was used as a method for providing formative assessment and correctives in the Mastery Learning cycle.

1.4.1 Purpose of the Study

The purpose of this study is to investigate the efficacy of combining Blended and Mastery Learning approaches in a Blended Learning Mastery Progression Cycle (BLMPC). There is limited research into the use of Mastery Learning approaches in this context, particularly with the use of Blended Learning activities as correctives. One barrier to the implementation of Mastery Learning in classrooms is the significant planning required to develop a range of suitable and effective corrective activities (Morgan, 2011). However, the use of commercially available programs may allow teachers to efficiently use the assessment and correctives cycles of mastery approaches.

This study aims to provide educators with evidence as to the effectiveness of using Blended Learning approaches in a Mastery Learning cycle and may provide improved pedagogical approaches to the use of ICT in the classroom environment.

1.4.2 Research Questions

The research questions (RQ s) for this study are based on three foci; the effect of a Blended Learning Mastery Progression approach on student attainment, student attitudes toward Science and associations between achievement and attainment.

RQ 1 Student Attainment

RQ 1a: Are pre-unit FCI scores for the treatment and control groups statistically significantly different?

RQ 1b: Are there statistically significant differences between the pre- and post-unit FCI scores for the control group?

RQ 1c: Are there statistically significant differences between the pre- and post-unit FCI scores for the treatment group?

RQ 1d: Are post-unit FCI scores for the treatment and control groups statistically significantly different?

RQ 2 Student Attitudes

RQ 2a: Are pre-unit TOSRA scores for the treatment and control groups statistically significantly different?

RQ 2b: Are there statistically significant differences between the pre- and post-unit TOSRA scores for the control group?

RQ 2c: Are there statistically significant differences between the pre- and post-unit TOSRA scores for the treatment group?

RQ 2d: Are post-unit TOSRA scores for the treatment and control groups statistically significantly different?

RQ 3 Associations

RQ 3 Are there associations between FCI scores and student attitudes towards Science, as measured by the TOSRA scales?

1.4.3 Limitations

The focus of this study is the use of a Blended Learning Mastery Progression Cycle (BLMPC) on student achievement and attitudes in a Physics class in an Australian high school. Therefore, the study is limited in context to the ACARA curriculum elaboration ACSSU229 (‘The motion of objects can be described and predicted using the laws of physics’) (ACARA, 2020), and Year 10 Australian high

school students. Achievement was limited to the measure of conceptual understanding as determined by the FCI, and students' attitudes towards Science were limited to those measured by the TOSRA Scales A, E and I.

Additional limitations existed because the study involved students from one school, which had a high level of ICT infrastructure and student competency. It may be that the findings cannot be generalized to other school contexts.

1.4.4 Assumptions

It was assumed that students who participated in the study completed all the learning activities, corrective assignments, and pre- and post-unit FCI and TOSRA to the best of their abilities. Any differences between the treatment and control groups were due to the random assignments of the groups to either the treatment or control condition. It was also assumed that the general learning activities and teacher guidance were comparable across both groups.

It was assumed that instruments were valid and reliable in their measurement of student understanding and attitudes towards Science—specifically, that the FCI was an valid and reliable measure of student understanding of aspects of the Newtonian Force concept as described by the ACARA syllabus, and that the TOSRA Scales A, E and I reliably measured students' adoption of scientific attitudes, enjoyment of science lessons, and attitudes towards scientific inquiry.

A classical test theory approach was taken to data analysis. This implies assumptions regarding the relationship between test score and true score. An assumption was made that differences in test scores are only dependent on the ability or attitude being investigated (Van der Linden & Hambleton, 2013)—for example, that a score on the FCI is only dependent on a student's understanding of the Newtonian Force concept. Other sources of variation in the testing applications are held to be constant or to have a random error on test score; these sources may include internal or external

conditions of the students taking the assessment. Classical test theory assumes that each individual has a true score that would be achieved if there were no errors in measurement. Due to imperfections in test instruments there may be a difference (the measurement error) between the test score and the true score, the latter being the measure of the student's true ability or attitude. Classical test theory uses the standard deviation of the observed score and the reliability of the test to provide an estimate of the error in the measurement and the boundaries of the confidence intervals approximate to the value of the true score (Kaplan & Saccuzzo, 1997). These approaches are widely used, but have the limitation that the statistics generated are sample dependent. This issue can be mitigated when successive samples do not vary over time, such as in the use of pre- and post-unit approaches (Hambleton, 2004).

1.5 Organization of the Thesis

This thesis is organized into five chapters, with supplementary information presented in the appendices.

Chapter One provides an introduction to the study, including an overview of the problem, theoretical frameworks, research questions, and limitations of the study.

Chapter Two provides a review of the research literature relevant to the topics of the study. It explains: the approaches and effectiveness of Mastery Learning; the concept and variety of approaches in Blended Learning; methods of measuring student understanding, with a particular focus on the FCI; and methods of measuring student attitudes towards Science with a particular focus on the TOSRA scales.

Chapter Three outlines the research methodology used in the study. It explains the research focus and questions that were investigated. The intervention is described as the use of a Blended Learning Mastery Progression Cycle, utilizing the Minds on Physics (MOP) platform for formative assessments and corrective activities. The

procedures for measuring student attainment and attitudes are explained, including the methodologies used for the analysis of data from the FCI and TOSRA.

Chapter Four presents the results of the study, including the characteristics of the cohort, inferential statistics from the analysis of the FCI and TOSRA scale data, and associations between the FCI and TOSRA scale scores.

Chapter Five presents a discussion of the results in relation to the individual research questions of the study. The results are then compared with pertinent material in the research literature. This places the empirical findings in the context of other studies, and enables a conclusion to be drawn regarding the overarching question of the study. This is followed by a discussion of the implications of the findings of this study for use in the science classroom and curriculum development. Finally, the limitations of the study are discussed and areas for further research are highlighted.

The appendices contain: sample questions from the FCI; a taxonomy of the concepts probed by the FCI; scales A, E and I from the TOSRA; examples of the formative questions and corrective activities from the MOP app; and a copy of the consent letter sent to participants and parents.

1.6 Chapter Review

This chapter has provided an introduction to the study, including: the context of the study in terms of underachievement in STEM subjects in Australian high schools; an introductory explanation of the concepts of Mastery and Blended learning; a statement of the purpose of the study and the research questions to be investigated; the limitations due to cohort characteristics; and assumptions regarding the participants and instruments. Finally, the content of each chapter has been outlined.

The importance of student performance in STEM subjects has been explained in terms of the predicted workforce requirements and the current perceived underachievement in the area of Australian students. Possible causes of this

underachievement have been suggested in relation to the nature of the subjects and pedagogical approaches. The increased use of ICT in Australian high school classrooms has been identified as a possible source for improvement in pedagogical approach.

Mastery Learning has been explained in terms of a 4-point cycle with a focus on sequential learning, regular checks and feedback to students, corrective activities, and repeated opportunities to demonstrate success. Blended Learning has been explained as the use of approaches that blend face-to-face and ICT-based learning activities.

A gap in the research literature has been identified due to limited rigorous studies into the impact of Blended Learning approaches and no recent studies of Mastery Learning in Australian high schools.

The purpose of this study was to investigate the efficacy of a Blended Learning Mastery Progression Cycle approach to learning in the Year 10 ACARA Physics curriculum. Hence, the study aimed to provide educators with evidence as to the effectiveness of using Blended Learning approaches in a Mastery Learning cycle and may inform pedagogical approaches for the use of ICT in the classroom environment.

Chapter 2: Literature Review

2.1 Chapter Overview

This chapter initially provides a review of the current research literature in the focus areas of the study of Mastery Learning and Blended Learning. The discussion of the Mastery Learning approach focuses on the historical development of the approach and a range of strategies used in its implementation. A review of the effectiveness of various Mastery Learning approaches in a range of contexts is discussed. Blended Learning is discussed in the context of the increased availability of ICT in schools. Definitions from the literature are used to frame the use of Blended Learning within the study, and a range of approaches are discussed in terms of application and effectiveness, with a particular focus on Science education.

The remainder of the chapter focuses on the use of tools to measure student attainment in Science, and attitudes towards Science. The structure, use and effectiveness of the Force Concept Inventory (FCI) as a method for measuring student understanding of the Newtonian concept of force is discussed. Finally, use of the Test of Science Related Attitudes (TOSRA) as a measure of students' attitudes is discussed, along with a presentation of its alignment with other attitudinal scales.

2.2 Search Strategy and Inclusion Criteria

Mastery Learning has been applied to a broad range of educational applications from vocational to university, and in various school settings (Guskey, 2010). This has led to a large quantity of research across this range of applications. Due to the large amount of literature relating to student achievement in a Mastery Learning context, a review was conducted of the literature most relevant to the focus of this study. The works of Bloom (1968), Block and Anderson (1975), and Guskey (1980) were used to develop an understanding of the historical development and key aspects of Mastery

Learning approaches. The effectiveness of Mastery Learning approaches was reviewed through the use of a range of meta-analyses to give the broadest interpretation of findings (Hattie, 2009; C. Kulik et al., 1990), with a particular focus on Science applications. Research identified in these meta-analyses was used to review more fine-grained details of the results obtained in subject areas most relevant to this study.

Literature regarding approaches to Blended Learning was reviewed with an initial focus on the defining parameters of these approaches (Driscoll, 2002). This was supported by the use of database searches and citation tracing to develop an understanding of the wide range of strategies used in these approaches. A more focused review of the literature relevant to Science-specific Blended Learning activities, simulations and multimedia presentations, was conducted to determine both their effectiveness and the scope of research in this area.

Quantitative methods to measure student attainment and attitudes were investigated, with a focus on concept inventories and attitudinal scale measurements.

2.3 Introduction

A range of Mastery Learning approaches have been shown to have positive educational outcomes in academic performance in formative and standardized tests, student engagement, and the long-term retention of skills and information (Dillashaw & Okey, 1983; Guskey, 2010; Inui, 2015; C. Kulik et al., 1990; J. A. Kulik et al., 1976; J. A. Kulik et al., 1974; Zimmerman & Dibenedetto, 2008). However, there is a lack of recent research in the Australian context (Hattie, 2009; Melbourne Graduate School of Education, 2018). More recently, research has focused on the refinement of the individual aspects of Mastery Learning approaches and its use in combination with other pedagogies (Guskey, 2010). The effective design and use of corrective activities has been one such area of study (DeWeese & Randolph, 2011).

The availability of ICT in schools has increased the interest in, and application of, Blended Learning in the classroom (Chandra & Fisher, 2009; Crook et al., 2013). Research into the effectiveness of various Blended Learning approaches, such as the use of simulations (Finkelstein et al., 2005), multimedia learning activities (Chandra & Watters, 2012), and computerized tutoring (Bayraktar, 2001), has shown them to be effective in a Science education context (Graham, 2005; Güzer & Caner, 2014).

This study investigated the combination of Mastery and Blended Learning by using Blended Learning activities as the correctives in a Mastery Learning cycle.

2.4 Mastery Learning

Achieving mastery of a subject being studied seems to be a reasonable goal for all students; however, this outcome is not always achieved. Many students, across a range of subjects and educational settings, fail to achieve such mastery, leading to a range of detrimental effects (Darling-Hammond, 2004; Masters, 2016). For example, 40,000 Australian 15-year-olds failing to achieve an international baseline proficiency in Reading and 57,000 failing to achieve the baseline in Mathematics each year (Masters, 2016). There are numerous factors that may contribute to a student's inability to master a subject, including lack of interest, the concept being beyond their zone of proximal development, poor quality learning activities, poorly trained teachers, poor teacher–student relationships, and social and emotional influences (Churchill et al., 2013). Nevertheless, numerous studies have indicated that it is often a lack of time and appropriate learning activities, rather than aptitude, which leads to students failing to achieve mastery (Bloom, 1984; Guskey, 2007; C. Kulik et al., 1990). The concept of allowing students differing amounts of time to develop knowledge and skill acquisition is not new (Guskey, 1986; Keller, 1968), and is perhaps most formalized in the work of Bloom (1968) in his Learning for Mastery model (LFM) and in Keller's Personalized System of Instruction (PSI) (Keller, 1968).

The attributes of a student or course of study which determine the student's ability to master a unit of work, and hence lead to variability in academic achievement, have long been subject to research. Carroll (1989) outlined five classes of variables that, in his model, account for this variance:

- Aptitude, which is defined as the amount of time a student needs to learn a given task or concept to an acceptable level of mastery, when under optimal conditions of instruction and motivation. A high aptitude indicates a short amount of time is required.
- Opportunity to learn, which is defined as the amount of time allowed for learning, frequently determined by the course provider.
- Perseverance, which is defined as the amount of time a student is willing to spend on learning the task or concept and is a measure of the motivation for learning.
- Quality of instruction. Although the model does not define characteristics of high-quality instruction, it does outline the need for the sharing of clear learning objectives with carefully planned learning steps, and opportunities for teacher feedback.
- Ability to understand instruction, which is described in terms of both language comprehension and a learner's ability to determine the most effective way to complete a learning task.

It is the interplay of these factors that determines the ability of a student to be successful in achieving mastery. For example, if a student lacks the perseverance to spend the time on learning as determined by their aptitude, then they will not achieve mastery, or if a course is constructed so that students cannot understand the instructions, or the time allocated for learning is insufficient, then mastery cannot be achieved (Carroll, 1989).

Research into the attributes that determine student achievement has continued to the present day, with the recent works of Dweck (2007) refocusing the public's and educators' attention on this area (Hopkins, 2015). Dweck (2007) refers to a 'growth mindset' as being the determining factor for perseverance and hence mastery of work, specifically referring to students that are 'mastery orientated' as being more successful than those who are 'grade orientated'. Dweck concludes that the development of this mindset is key to ensuring student success (Dweck, 2013). Students who focus on mastery goals, as opposed to extrinsic performance goals, are more likely to persist at academic tasks and use more effective metacognitive strategies (Harackiewicz et al., 2000; Wolters, 2004).

2.4.1 Historical Development of Mastery Learning Approaches

The term 'Mastery Learning' is frequently used to describe a range of approaches that emphasize sequential learning, regular checks and feedback to students, corrective activities, and individually determined progression times (Guskey, 2007). Guskey (1980, 2007) charts the development of Mastery approaches from Washburne and Morrison in the 1920s through to Bloom's Learning for Mastery (LFM) model (Bloom, 1968). However, there is some discrepancy in the exact terminology used to describe the variety of methods of personalized learning utilized in Mastery Learning approaches (Guskey, 1997; J. Kulik et al., 1976) and some conflation of LFM with Keller's (1968) Personalized System of Instruction (PSI) (Slavin, 1987). The following section aims to clarify the terminology and techniques used by charting the development of various Mastery Learning approaches to the present day. It should be noted that Bloom originally referred to his model as Learning for Mastery (LFM) but later changed this to Mastery Learning (Guskey, 1997); however, to distinguish between Bloom's approach and the wider field of Mastery Learning, the original name will be used in this review.

2.4.1.1 Learning for Mastery Model

The rationale behind Bloom's development of the Learning for Mastery (LFM) model came from his dissatisfaction in observing that only a third of students within a class would adequately learn what was required of them in the course (Bloom, 1968), and the subsequent negative impacts on both the student and society. Using results from individual tutoring as his reference point, Bloom (1984) opined that over 90% of students can master the concepts within a course, and stated that it is the task of the teacher to determine the level of mastery required by a course and the methods that will enable students to reach this level.

Bloom (1968) argued that that traditional grading, on a normal distribution curve, does little to measure students' actual understanding or competency in a concept; rather, it ranks students in comparison to their cohort, in an attempt to distinguish between performances (Guskey, 2007). This ranking is used in many present-day education systems—for example, the newly adopted Australian Tertiary Admission Rank system in Queensland is a method of ranking students for tertiary study that is based on comparative performance rather than demonstrated criteria (Matters & Masters, 2014). Bloom built on the work of Carroll (1989) to conclude that aptitude for learning is based on the time required for the student to master a concept, and mastery of a concept is available to most if enough time is allocated or appropriate learning activities are provided. Hence, it is the limitations of course time and activities that lead to students being unable to master a concept (Bloom, 1968). In a system where all students are taught unit content in the same manner, are tested on this content, and then move on to the next unit, time and the provision of alternative learning activities are limited and so students may fail to achieve mastery (Guskey, 2007).

Bloom further identified the sequential nature of concepts in many subjects, such as Mathematics and Physics, as causing issues in the longer term mastery of the

subject, because students who do not master the initial concepts do not have the prerequisite knowledge or skills required for mastery of the later concepts (Bloom et al., 1971). This links to the ideas of Vygotsky (Zone of Proximal Development), Pestalozzi, and others (in Churchill et al., 2013) in that students must have an understanding of the underpinning theory of a new concept to fully understand it. This lack of prior competency can lead to cumulative effects as students move through successive units of work, with some researchers indicating that only about 20% of the students in a class understand all the material by the end of the school year (Guskey, 1997).

In developing the LFM model, Bloom identified individual tuition as the ideal instructional strategy for developing mastery in all students, and attempted to replicate the benefits of tutoring in a group setting (Bloom, 1984). An effective tutor will identify when a student makes an error (feedback) and follows up with an alternative explanation and remedial activities (correctives). In LFM, this is achieved by extending teaching beyond the unit test, through a series of feedback and corrective instruction cycles with enrichment activities when mastery has been demonstrated. Guskey (1980) identified these cycles and the alignment of instructional material with assessment criteria as the most fundamentally important aspects of LFM. It is important that the feedback given for formative assessments is specific (Goss et al., 2015; Hattie, 2009) and that the corrective instruction activities are aligned to the identified issues and involve a range of pedagogical approaches (Bloom, 1984; Guskey, 2007; Levin, 1979).

The LFM instructional model can be described as a four-stage process (Guskey, 2007). Stage 1 involves group instruction, using a range of teaching and learning activities, at a pace determined by the teacher, usually for one to two weeks. Stage 2 involves the students completing a formative assessment based on the learning outcomes for the topic, and being given feedback on areas that require focus. Stage 3 involves students completing corrective or enrichment activities based on the areas

identified in Stage 2. In Stage 4, students are given a second chance to demonstrate mastery of the topic by taking a parallel assessment. Students who still fail to demonstrate sufficient mastery can be assigned further corrective activities (Block, 1977; Guskey, 2007).

Although it is apparent that Bloom built on Carroll's idea of time being one of the limiting factors for mastery (Guskey, 2007), his LFM approach does not allow limitless time or opportunities for students to develop and demonstrate mastery. Rather, it is the focused application of corrective activities that is the essential form of differentiation in this approach (Guskey, 1997). This leads to a reduction in the amount of time required for mastery to be achieved (Carroll, 1989).

LFM has been applied in a range of contexts, from postgraduate medical training (Inui, 2015) to massive open online courses in a range of subjects (Said & Zainal, 2017), and primary school Mathematics courses (Throndsen & Turmo, 2013), with a high level of success in many situations (C. Kulik et al., 1990).

2.4.1.1.1 Correctives in Learning for Mastery

It is the directed application of specific corrective activities that differentiates LFM from other Mastery Learning approaches (Block, 1977; Guskey, 1997, 2007, 2010). The specific nature of the correctives is achieved through the use of formative assessments to identify the specific areas of focus for a student. Such formative assessment allows the feedback to students to contain a clear goal, have evidence of the student's current understanding, and include a method for the student to achieve the goal. These features have been identified as key parameters of good feedback (Black & William, 2010), and many authors have stressed the importance of effective feedback on assessments (Guskey, 2007).

Correctives, as described by Block and Anderson (1975), are learning activities that present the same content as originally taught in the group phase of instruction, but

in a different format, thus allowing an alternative method for developing student understanding through a range of cognitive processes (Block, 1977; Guskey, 2007). Due to the importance of corrective activities in the LFM process, it is important they are carefully designed to meet the learning needs of students (Block & Anderson, 1975; Guskey, 1997). Guskey (1997, 2007) outlines three fundamental requirements of effective corrective activities: they must present the concepts in a different format from the original instruction; they must involve different learning activities; and they must provide students with opportunities for successful learning as aligned with the objectives.

Correctives may be conducted in small groups or individually, and include activities such as alternative textbooks, workbooks, multimedia activities, academic games, computer-based activities, tutoring, and small group study sessions (Guskey, 2007). It is apparent from the literature (Bloom, 1984; Guskey, 2007) that the quality of corrective activities used is fundamental in ensuring that students have the opportunity to achieve mastery.

2.4.1.2 Personalized System of Instruction

Although originally developed as a method for delivering an undergraduate Psychology course at the University of Brasilia and Arizona State University in 1965 (Keller, 1968), the Personalized System of Instruction (PSI) has been adapted and applied to numerous scenarios. Sometimes referred to as the Keller Plan (J. Kulik et al., 1974), the approach is defined by being individually paced and mastery orientated.

PSI builds on the Winnetka Plan of individualized instruction whereby students worked in small groups with tutors followed by repeated testing until mastery was demonstrated (Motamedi & Sumrall, 2000). Developments in individual instruction up to the 1960s were focused on secondary and elementary schools, these were applied to

create the PSI for use higher education where it received considerable attention (J. Kulik et al., 1974).

Keller (1968) outlined five main features of a PSI course, to distinguish it from traditional lecture-based teaching approaches. He explained that courses should:

- be individually paced
- be mastery orientated
- be student tailored
- include printed study guides for communication
- include some face-to-face lectures for stimulation.

At the start of a PSI course, students receive a guide that explains the unit structure of the course, each unit's objectives, study procedures and questions. In a contemporary setting, the study guides, lectures and mastery tests are often available in a digital format (McRae, 2015). Students are then free to study the unit contents at their own pace, and when they feel they have met the objectives they are required to complete a short examination to demonstrate mastery. Once mastery has been demonstrated, the student receives access to the next unit, thus moving through the entire course at an individual pace. Courses are frequently monitored by tutors or proctors who are available to provide remedial activities and support for students who fail to demonstrate mastery. A further feature of the original Keller course was that the final course grade was largely based on the success in the unit examinations, as opposed to a final exam covering all units (Keller, 1968; J. Kulik et al., 1974).

PSI style courses have been widely used in formal education settings, workplace training, and self-help courses (Grant & Spencer, 2003; Viness et al., 2017) and have been shown to have a variety of positive effects (Hattie, 2009).

2.4.1.3 A Comparison of Learning for Mastery and Personalized System of Instruction

Table 2.1 *A Comparison of LFM and PSI Approaches*

Feature	Learning for Mastery	Personalized System of Instruction
Pace	Group lessons paced by teacher, correctives paced by student within overall time-frame set by course requirements.	Students work through materials at their own pace within the confines of the course deadlines.
Learning Materials	A range of pedagogical approaches used in the group learning phase; different approaches used in the correctives.	Written or multimedia presentation of information in the initial learning phase, some support and alternative activities provided by proctors after the initial formative assessment.
Mastery Requirement	Students required to demonstrate mastery before moving on to the next unit; specific corrective interventions applied. Due to the group nature of the initial learning phase, issues arise if students do not have sufficient time to reach mastery before cohort progresses.	Students are (required to demonstrate mastery before having access to the next unit material.
Timing	Some students may require extra time outside of class to achieve mastery.	Students may require additional time outside of class to achieve mastery.

Table 2.1 shows a comparison of the features of LFM and PSI. While each method has the same focus on mastery of topics before progression, the delivery and approach are disparate.

2.4.2 Positive Educational Impacts of Mastery Learning Approaches

A meta-analysis of historical research has shown positive educational impacts of Mastery Learning approaches, both in LFM and PSI formats, with improvements showing a significant effect size (C. Kulik et al., 1990). In their extensive meta-analysis of prior studies, C. Kulik et al. (1990) summarized the findings of 108 PSI style programs and 36 LFM approaches, with the majority of the PSI studies set in a college

environment and with the LFM studies mostly being conducted in a high school context. By using examination performance as a measure of success, statistically significant improvement (Mean Effect Size = 0.52, which is above the 0.4 threshold (Hattie, 2009)) was demonstrated in 67 out of 108 of the mastery programs when compared with control groups; however, there was a wide range of results between studies. The most significant improvements were found in mastery programs that: exhibited some group-based aspects rather than being self-paced; had high requirements for mastery to be demonstrated; had assessments based on program-specific exams compared to external standardized tests; and provided increased feedback. Although differences were reported between the parameters of PSI and LFM studies, both approaches produced positive results of similar effects (PSI $d = 0.48$, LFM $d = 0.59$, both above the 0.4 threshold). The study also reported positive benefits towards students' attitudes towards the subject and the instructional method.

C. Kulik et al. (1990) attribute these gains to the mastery approach rather than any other factors, such as increased time, course completion rates, or differences in feedback, and conclude that mastery approaches are effective in improving student performance when compared with a range of other interventions. Similar conclusions have been reached by Block and Anderson (1975) and Guskey (2010). Bloom (1984) claimed that LFM leads to gains of one full standard deviation when compared with conventional instruction and that it leads to the improved engagement of learners of all abilities due to the nature of correctives and enrichment activities. Studies by Davis and Sorrell (1995) and Miles (2010) have reported a range of positive impacts on student attainment in high schools in the United States across several subjects and age ranges.

In an Australian context, there is limited research in the field of Mastery Learning. Hattie (2009) relied on nine pre-1990 studies for his meta-analysis (although these appear to be meta-analyses themselves, of US studies), and found an effect size of

d = 0.58, which is significantly above the 0.4 threshold. Melbourne Graduate School of Education (2018) reports that there have been no new studies of Mastery Learning approaches in Australian high schools since 2008. There have been some recent studies in Australian Universities, with a study by Shafie et al. (2010) demonstrating improved attainment in a university Mathematics course and research at the Australian National University (Francis et al., 2009) showing some positive results in a first-year undergraduate Physics course.

2.4.3 Criticisms of Mastery Learning Approaches

Some criticisms of Mastery Learning are focused on a disagreement with the fundamental premise that mastery can be achieved by all students and the behaviourist foundation of the theoretical frameworks for such approaches (Block & Burns, 1976). Criticisms aimed at Mastery Learning are frequently delineated by the approach being considered PSI or LFM.

In Slavin's (1987) analysis of LFM, he outlines the oft-cited criticism of the increased time required for mastery leading to a lack of breadth in the curriculum. Bloom (1987) counters that there is little benefit to having breadth if the foundational knowledge is not secure. Slavin (1987) further comments on students having accomplished mastery wasting time waiting for classmates to catch up; this highlights the importance of effective enrichment activities being available for such students.

The allocation of time for students to perform correctives is also seen as problematic because of the envisaged need for either extra time outside of class or the lack of time available to cover future curriculum content (Arlin, 1984), with much discussion in the literature around how this problem is most appropriately managed (Slavin, 1987). The issue has led to criticism of the experimental methodology of some studies due to the unequal time allocated to experimental and control groups. However, studies have reported that the amount of additional time required for correctives

decreases as courses proceed, especially in hierarchical subjects such as Mathematics, Science and languages (Anderson, 1976; Block, 1972).

Further criticisms have been made of studies that use tests developed by teachers or researchers specifically to evaluate the effectiveness of the approach, as this may create bias towards the experimental group when compared to standardized testing (Slavin, 1987), in response to which Bloom (1987) highlights the increased alignment to unit objectives of such assessments. Although Slavin (1987) highlighted some valid concerns with LFM, his meta-analysis only analyzed seven studies. This was due to the restrictions placed on inclusion which specified that there should be no use of feedback-corrective cycles, and interventions should be at least four weeks in length, both counter to the fundamental features of the LFM model (Hattie, 2009).

Considerations of PSI approaches identify the negative impact they can have on course completion rates (C. Kulik et al., 1990), with rates decreasing by a small but significant amount. Researchers have identified the negative impact of student procrastination on the effective use of PSI courses (Eyre, 2007), and significant efforts have been made to reduce its impact, including the use of contracts and pacing requirements. One of the fundamental features of the PSI approach is the use of proctors to aid students in completing the learning activities and provide support when required. The quality of proctor feedback is therefore vital to the success of the student, and issues arise where this feedback is insufficient (Eyre, 2007).

In summary, the criticisms of Mastery Learning approaches tend to focus on the allocation of time and the quality of feedback and correctives required, these factors may be addressed by careful planning and management.

2.4.5 Mastery Learning in a Science Context

Mastery Learning approaches have been used in a variety of Science subject areas and across a range of educational settings (C. Kulik et al., 1990). Ngozi and

Chinedum (2012) found a positive relationship between a Mastery approach and achievement when compared to a lecture-based approach for an Electricity unit in Physics in Nigerian high schools. Whilst the study showed an improvement in mean gain of Mastery approach students, 37.15% compared to 22.88%, the sample size was small (N= 40), the test instrument may have been subject to bias (Slavin, 1987), and the comparison with a lecture-based approach means the findings may not be transferable to Australia, where a more diverse range of pedagogical approaches are used. A Kenyan study (Wambugu & Changeiywo, 2008) found significant improvements in achievement, mean scores of 54% compared to 24%, with $p<0.05$ showing strong confidence in the causality of the improvement, for units involving equilibrium and centre of gravity topics when using an adapted standardized test. However, it is unclear what the format of the regular teaching method involved.

Dillashaw and Okey (1983) investigated the effects of a range of mastery approaches to achievement, attitude and on-task behaviour in a high school Chemistry context. They found positive gains in achievement in Mastery groups (difference in mean scores of up to 12%), even when specific corrections were not applied, concluding that the use of regular diagnostic tests may be sufficient to improve achievement. Similar gains were found by Damavandi and Kashani (2010), especially in the performance and attitudes of weaker students. However, the sample size (n=40) makes it difficult to extrapolate the findings to other situations.

The study of PSI approaches appears to be mainly focused on college level science courses. A meta-analysis by J. Kulik et al. (1974) evaluated courses across a range of subjects from introductory Psychology to Fluid Mechanics and Biology. They found: positive course reviews by students, mainly related to self-pacing and tutor contact; an overall improvement in achievement; and a reported increase in learning,

time applied and effort. Negative relationships were found with course completion rates and procrastination.

In summary, there is a lack of research into the effect of Mastery Learning approaches in an Australian high school Science setting, but the wider research shows generally positive effects of both LFM and PSI in students' achievements in Science subjects.

2.5 Blended Learning

The term 'Blended Learning' is used to describe a variety of methods of teaching with technology. However, it may be more appropriately defined in terms of the learning experience (Oliver & Trigwell, 2005). A more learning-centred definition describes Blended Learning as being a combination of different concepts, including:

- the combining of web-based technology to accomplish an educational goal.
- the combining of pedagogical approaches, such as constructivism, behaviourism and cognitivism, to produce an optimal learning outcome, with or without instructional technology.
- the combining of any form of instructional technology with face-to-face instructor-led training (Driscoll, 2002).

This three-point definition will be used as the basis for the Blended Learning approaches discussed in this study. Other definitions place greater emphasis on the features of Blended Learning course structures and a mix of face-to-face and online activities (Stein & Graham, 2014), as opposed to the pedagogical focus adopted by the definition above.

Interest in Blended Learning approaches has increased with the availability of cost-effective IT solutions in schools (Chandra & Fisher, 2009; Crook et al., 2013). Due to the Australian government laptop scheme, many schools now have an abundance of computers available for student use (Crook et al., 2013). If these IT resources are to

have a significant positive impact on student learning, they must be used for more than a simple replacement for handwriting and textbooks, as too often they are poorly integrated into classrooms (Songer, 2007) and fail to transform pedagogical practice (Cuban, 2001). A number of studies have been conducted into the impact of computers on student outcomes—for example, Hattie (2009) identifies nearly 4,500 studies across 4 million students, and a wide variety of IT-based approaches. The generalized findings show that the most effective use of IT occurs when: it is used for a diverse range of teaching strategies; there are multiple learning opportunities; the student has control of the learning process; and feedback is optimized. These results were more recently supported by a meta-analysis of the impact of technology on learning in elementary school students, which found a significant average effect size of 0.546 when compared to the 0.4 threshold (Chauhan, 2017). It is interesting to note the alignment of these factors with the features of LFM: a variety of learning activities (instruction and correctives), student pace of correctives, and feedback from formative assessments.

It is important for teaching methodologies to adapt if these resources are to have a positive impact on student learning (Crook et al., 2013). Blended Learning at its most effective requires each student to have an individualized learning path utilizing various online and teacher-developed resources (Stockwell et al., 2015). Learning experiences can then be tailored to reflect a student's prior knowledge and preferred learning style.

There are limited studies that have rigorously assessed the impact of Blended Learning on student outcomes in a P-12 context (Means et al., 2010; Sparks, 2016; Stockwell et al., 2015). In a meta-analysis of studies across a range of educational settings, Means et al. (2010) found that it was the blending of face-to-face with online instruction that led to the largest advantage for student outcomes when compared with face-to-face or online only programs. With an effect size = +0.35, $p < .001$, the impact fell below Hattie's 0.4 threshold (Hattie, 2009), but still indicated an improvement in

outcome with high confidence in the correlation with the approach and the outcome. The study also showed that the largest effect was found where Blended Learning activities involved students interacting with media rather than being passive recipients of information (Means et. al., 2010), although this still fell below the 0.4 threshold. The definition of Blended Learning used in this report was more focused on a description of course type than pedagogical approach but the features of Driscoll's definition (2002) were met. However, the authors indicate that the lack of studies in the P-12 context makes it difficult to extrapolate findings to this group; they also identify the impact of uncontrolled variables, such as time spent, instructional material and pedagogy, as problematic in determining the differences between the treatments.

It is evident that Blended Learning approaches have the potential to positively affect student achievement but that programs need to be carefully designed, managed, and studied. In particular, the use of instructional technologies needs to be carefully considered (Patchan et al., 2016), because, whilst they can provide positive benefits in terms of pace flexibility, information formats, immediate and targeted feedback, and student participation, significant teacher training is required (Hattie, 2009). The range of instructional technologies available is vast, from YouTube videos, to PowerPoint presentations, to online encyclopedias, to simulations, to digital quizzes, and sophisticated computerized tutoring systems (Chandra & Watters, 2012; Conole et al., 2008; Driscoll, 2002; Patchan et al., 2016). It is beyond the scope of this study to investigate all these approaches, so a focus will be placed on Computer Aided Instruction (CAI), Multimedia Learning Activities and Simulations, as these have demonstrated the highest effect size in previous studies (Means et. al. 2010)

2.5.1 Computerized Tutoring

Much effort has been put into the development of systems to allow computers to teach, and more specifically tutor, students (Hattie, 2009; VanLehn, 2011). The goal of

these systems has been to achieve the same impact as human tutoring which is often reported to have an effect size of 2 (Bloom, 1984). However, this goal fails to account for the varieties in human tutoring types such as face-to-face, online, small group, and individual sessions, and when considering the variety of human tutoring approaches, there is a much wider range of outcomes (Bloom, 1984; VanLehn, 2011).

The use of computer tutoring is normally separated into two approaches; Computer Aided Instruction (CAI), in which students receive immediate feedback and hints regarding their answer, and Intelligent Tutoring Systems (ITS), in which students complete tasks in steps, allowing the system to give hints and feedback on each step. In a meta-analysis of the effectiveness of these two approaches in STEM subjects VanLehn (2011) concluded that CAI has an effect size of 0.31 and ITS had an effect size of 0.76, the latter being comparable to the human tutoring studies reviewed, and significantly above the 0.4 threshold indicating an above average impact on learning.

The effectiveness of human tutoring is frequently explained by several theories, including detailed diagnostic assessments, individualized task selection, student dialogue, broad content knowledge of the tutor, motivation, feedback, scaffolding, and moderation of student behaviours (VanLehn, 2011). These instructional characteristics are frequently interdependent, and tutoring that uses effective scaffolding and feedback to engage students in interactive learning activities has been found to be the most beneficial (Chi, 2001; Graesser et al., 2010). Developers of CAI and ITS approaches need to include these characteristics in their systems if they are to move towards the 2.0 effect target.

CAI approaches can be further separated into three categories: drill and practice, tutorial, and simulation (Bayraktar, 2001). In a meta-analysis of 42 studies of the effect of various CAI approaches in Science education, Bayraktar (2001) found an overall d = 0.273, with the most effective approach being simulations (d = 0.391), and the least

effective being drill and practice (d =-0.107). The latter result is contradictory to the findings of other researchers (J. Kulik et al., 1983; Niemiec & Walberg, 1985), who found drill and practice to be beneficial with effect size= 0.47 (>0.4 threshold) across a range of subjects. This may be due to the cognitive requirements of Science courses being unsuited to rote learning as promoted by drill and practice approaches. When approaches were separated into subject areas, it was found that Physics was most impacted by CAI approaches (effect size=0.555 which is significantly >0.4 threshold) (Bayraktar, 2001). However, the limited number of studies in this area indicates that further research is required to substantiate the findings and determine why CAI appears so effective in a Physics context. Findings also showed larger effect sizes when CAI was used to supplement regular instruction compared with solely using CAI (Bayraktar, 2001). This indicates that CAI could be used to enhance learning activities rather than to replace traditional instruction. Other important findings were that individual computer use proved significantly more effective than group use (effect size = 0.368, effect size = 0.096, both < 0.4 threshold), and that the length of CAI intervention was best limited to short sessions < 4weeks (perhaps related to a novelty or Hawthorne effect) (Bayraktar, 2001; Güzer & Caner, 2014).

It is important to note that these studies do not suggest replacing explicit group teaching activities; rather, they should be used to replace traditional homework such as question reviews.

2.5.2 Multimedia Learning Activities

Multimedia Learning Activities (MLA) are defined by their use of a variety of media types, including text, images, videos, animations and sound (Mayer, 2009). However, Mayer (2009) argues that simply using multiple types of media does not lead to an improvement in learning without an understanding of a cognitive theory of multimedia learning. This Cognitive Theory of Multimedia Learning (Mayer, 2009) is

based on three principles of learning: there are dual channels for processing visual and auditory information; there is a limit to the capacity of each channel; and learning requires the active completion of a set of coordinated processes involving audio and visual representations (Mayer & Moreno, 2003). These principles guide the development of effective MLA and highlight some issues that need to be considered in activity design to avoid cognitive overload—for example, using techniques such as reducing text content, removing unnecessary content, and ensuring diagrams are clearly labelled. Further, by considering the cognitive aspects of the MLA, the focus of the activity is placed on learning, rather than the media being used. This is important, as the literature regarding the benefits of various media types is large and often contradictory and media types continually undergo rapid change (Muller et al., 2005).

The approach of tailoring multimedia learning activities has a well-documented history of success in the Science curriculum. For example, a study of the impact of using a computer simulation on the development of students' understanding of the topic of buoyancy found an increase in students' understanding of both the buoyancy concept and the experimental process (Zhang et al., 2004). In a study utilizing a web-based Physics resource, Chandra and Fisher (2005) reported positive impacts on both student attainment and engagement. They accounted for these findings in terms of the variety of resources (pedagogical approaches) available on the website, and the ability of students to work at their own pace, both of which are important aspects of the Mastery approach.

2.5.3 Simulations

It is generally acknowledged that students who construct their own understanding of scientific concepts achieve improved learning; this has traditionally been accomplished using experimentation and laboratory work (Bransford et al., 2000). However, there are numerous examples where such laboratory work are beyond the reach of students. This may be due to safety, time or financial constraints, and it is in

these situations that simulation applications can be useful (Akpan, 2002). In this context, simulations are defined as computer generated dynamic models of real-world phenomena and processes, which engage students in inquiry-based activities (Smetana & Bell, 2012). They may take the form of interactive experiments, animations, visualizations or interactive models. Some research has shown that simulations can be more effective than hands-on experiments, as students are able to focus on the concept, rather than completing the experiment to a high standard (Akpan, 2002; Finkelstein et al., 2005; Wieman et al., 2008), however they are limited in the development of some of practical skills required by the ACARA curriculum (ACARA, 2020). For simulations to be effective, they must be engaging for students, involve students interacting with their content, and not cause cognitive overload in their design (Adams et al., 2004). Further research has shown that computers have the most significant impact on academic outcomes in Physics when students interact with simulations (Crook et al., 2014).

The University of Boulder Physics Education Technology (PhET) library of simulations consists of a range of over 100 simulations across the range of Science disciplines. Their research has shown improvements in student outcomes when using these simulations to replace traditional hands-on experimental work and lecture-based courses (Wieman et al., 2008), and they report positive impacts on student engagement. While there is obvious potential for bias in these reports (as they are written by PhET staff), these claims appear to be supported by a range of studies (Finkelstein et al., 2005; University of Colorado Boulder, 2019). Smetana and Bell (2012) reviewed 61 studies into the use of simulations in Science education and found that they were effective at promoting content knowledge, improving learner perceptions, developing Science process skills, and challenging student preconceptions (i.e., conceptual change). They

also reported that simulations are most effectively used to supplement, rather than replace, other learning activities.

The use of simulations may help students to behave cognitively like scientists, by testing hypothesis and manipulating variables, and hence improve their learning. However, there are some issues identified with student perceptions of the alignment of simulations to real life (Wieman et al., 2008), and this could be especially important when students' preconceptions are challenged, as they may disbelieve the simulation rather than challenge their preconceived models. Other issues relate to teacher training in the effective use of simulations and the associated IT (Smetana & Bell, 2012).

The use of simulations can model aspects of the Mastery approach, as students are free to work at their own pace, receive feedback from the simulation, and perform corrective activities until their understanding aligns with the learning objectives of the simulation.

2.6 Student Attainment

The term 'student attainment' can be used to describe a variety of student academic outcomes, from results on standardized and internal assessments to qualitative teacher judgments or future earnings (Hanushek, 1997). The measurement of student attainment is an increasingly important part of many education systems, with high stakes public examinations being used to determine pathways available to students after high school (Queensland Tertiary Admission Centre, 2016; University Admissions Centre, 2016). Perhaps the most important measurement of student attainment is their ability to demonstrate an understanding of the subject over a range of scenarios (Mayer & Moreno, 2003). Therefore, the design of any assessment technique or instrument is critical to its validity in making judgments about student attainment (Jackson et al., 2008). For this study, 'student attainment' will be taken to mean a student's ability to apply the concepts of the topic to respond to questions and problems.

There is a wide range of methodologies that can be used to measure a student's understanding of Science topics, including detailed teacher-led questioning, teacher devised assessments, external assessments (e.g., New South Wales High School Certificate and International Competitions and Assessments for Schools exams), and validated instruments (e.g., Mechanics Baseline Test (MBT), Force Concept Inventory (FCI)).

2.6.1 Validated Instruments in Science and Physics Education

When using assessment instruments to measure student achievement and progression, it is important that the instrument is measuring the correct objectives and knowledge; this alignment is referred to as content validity. It is also important to ensure the instruments are population appropriate in their use of language and technical detail. Cohen and Wollak (2006) describe instrument validity as being dependent on a combination of instrument content and scope and alignment with the test population.

There is a range of instruments that have been validated for use in assessing student understanding in a wide variety of Science concepts for different populations. For example, the University of Pittsburgh Science Education Research Center lists over 20 instruments in Biology, including general Biology concepts ('Biology Concept Inventory'), Genetics ('Genetics Concept Assessment'), Diffusion and Osmosis ('Diffusion and Osmosis Diagnostics Test'), and 10 instruments in Chemistry, including general Chemistry ('Chemical Concept Inventory'), Thermodynamics ('Thermodynamics Concept Inventory') and Bonding ('Bonding Representation Inventory') (Discipline Based Science Education Research Center, 2018). PhysPort lists 87 validated instruments that assess various aspects of Physics, including Newtonian Force concepts ('Force Concept Inventory'), Quantum mechanics ('Quantum Mechanics Concept Assessment'), and Electrostatics ('Electricity and Magnetism

Conceptual Assessment'). These have been validated across a range of academic levels from middle school to graduate level (PhysPort, 2018).

2.6.1.1 Force Concept Inventory

The Force Concept Inventory (FCI) is a collection of 30 multiple-choice questions specifically developed to test the understanding of Newtonian mechanics and the 'Force Concept' (Hestenes et al., 1992a, 1995), which is widely used in Physics Education Research (Morris et al., 2012) (see Appendix 1 for sample questions). Concept inventories are research-based assessment instruments that measure conceptual understanding as opposed to rote learning or the application of algorithms. They are frequently delivered in multiple-choice form and differ from standard multiple-choice examinations in that the alternative responses are selected to conform to common misconceptions of the topic (Madsen et al., 2017). These distractor responses have been developed after many hours of interviews with students to determine the most common areas of misconception and the responses such misconceptions can lead to (Jackson et al., 2008). The inclusion of students' everyday 'common-sense misconceptions' (Hestenes et al., 1992b) and language in these distractors requires students to have a sophisticated understanding of the topic to select the correct response and minimizes the inflation of scores by guessing (Madsen et al., 2017). Concept inventories can be used both pre- and post-unit and therefore may be used to measure conceptual gain. Hake (1998) suggests that in this case it is useful to calculate a normalized gain; this methodology allows a comparison of the gains of courses at different levels and students of differing prior attainment. In this manner, gains can be used to compare students' progress on a variety of course types and hence evaluate the effectiveness of the course and interventions (Madsen et al., 2017).

The FCI is designed to measure six conceptual dimensions of the 'Force Concept'—Newton's First, Second, and Third Laws, Kinematics, Types of Forces, and

Superposition of Forces—as these are considered to be fundamental in developing and demonstrating an expert understanding of the ‘Force Concept’ (HWS6 model) (Hestenes et al., 1992a). Each dimension consists of a range of interlinked concepts which are assessed on multiple occasions, through a range of question types, to build an overall picture of a student’s understanding of Newtonian concepts.

Table 2.2 *Newtonian Concepts in the Revised Force Concept Inventory*

Dimension	Newtonian Concept	Inventory Item & correct response (1995 Version)
0. Kinematics	0.1 Velocity discriminated from position	19E
	0.2 Acceleration discriminated from velocity	20D
	0.3 Constant acceleration entails parabolic orbit	12B, 14D, (21E)
	0.4 Constant acceleration entails changing speed	25B
	0.5 Vector addition of velocities	9E
1. First law	1.1 With no force	6B, 7B, 8B, (11D)
	1.2 With no force velocity direction is constant	23B
	1.3 With no force speed is constant	10A, 24A
	1.4 With cancelling forces	17B, 25C
2. Second Law	2.1 Impulsive force	(8B), (9E)
	2.2 Constant force implies constant acceleration	21E, 22B, 26E
3. Third Law	3.1 For impulsive forces	4E, 28E
	3.2 For continuous forces	15A, 16A
4. Superposition Principle	4.1 Vector sum	(8B), (9E)
	4.2 Cancelling forces	(11D), (17B), (25C)
5. Kinds of Force	5.1 Solid contact – passive	11D, 29B
	5.2 Solid contact – impulsive	5B, 18B
	5.3 Solid contact – friction opposes motion	27C
	5.4 Fluid contact – air resistance	30CC
	5.5 Gravitation	3C, (5B), (11D), (12B), 13D, (17B), (18B), (29B), (30C)
	5.6 Gravity – acceleration independent of weight	1C, 2A
	5.7 Gravity – parabolic trajectory	12B, 14D

(No.) indicates item probes a range of concepts (Hestenes & Jackson, 2010)

Hestenes et al. (1992b) identified six ‘common-sense categories’ in which students’ understanding did not align with Newtonian thinking:

- Kinematics: problems distinguishing between position, displacement, velocity, and acceleration. Vectorial nature of displacement, velocity, and acceleration.
- Impetus: the belief that objects have an intrinsic motive force that keeps them moving along a path.

- Active Force: the belief that an active agent provides a force that is maintained by the object even when contact is lost, leading to the concept that motion implies an active force.
- Action/Reaction Pairs: confusion of action/reaction pairs with superposition principle and the need for a ‘winning’ force.
- Other influences on Motion: confusion regarding the impact of mass on falling objects and use of the term ‘centrifugal force’.

The responses to these ‘common-sense misconceptions’ form the basis for the distractor responses in the FCI, and this allows the source of the misconception to be diagnosed and challenged. Appendix 2 provides a taxonomy of naïve concepts probed in the FCI.

The FCI is an extension of the Mechanics Diagnostics Test (MDT*)* (Hestenes & Halloun, 1995; Madsen et al., 2017) which was developed through a process of collating students’ ideas on open-ended questions through a series of student interviews, and expert physicists using this data to refine the questions and responses. Statistical analysis was then performed on 1,000 student responses, which found the MDT test to be highly reliable (KR-20= 0.86 (Hestenes & Halloun 1985)). The same process was followed in the development of the remainder of the questions used in the FCI. There have been many studies (>50) using FCI data at high school and college level which include data on over 6,500 students (PhysPort, 2018). The FCI was reviewed and refined in 1997, to become the current 30-question iteration (Henderson, 2002). The wide-ranging use of the FCI provides the ability to compare instructional techniques and curriculum reforms using the comparative data available, although much of these data are focused on the college level in the United States (Henderson, 2002).

2.6.1.1.1 Critical Appraisal of the Force Concept Inventory

Although there is a large body of work that supports the use of the FCI as a diagnostic and measurement tool (Savinainen & Scott, 2002; Scott et al., 2012), there have been some issues identified with using the pre- and post-method, the conceptual coherence of the FCI (Henderson, 2002; Huffman & Heller, 1995; Scott et al., 2012), the use of normalized gain as a reporting metric (Miller et al., 2010) and with concept inventories more generally (Smith & Tanner, 2010; Wallace & Bailey, 2010).

Henderson (2002) identifies the influence of the pre-test on post-test performance as being a common area of concern amongst course teachers; this concern stems from students on a course being more conscious of the focus on the FCI assessment and so paying closer attention to these topics. However, a study at Western Michigan University found there to be no significant difference in post-test performance when a pre-test was given. Further, when using the FCI to conduct a comparative study, the pre-test is given to all groups, so any advantage is common amongst all students (Henderson, 2002).

The purpose of the FCI is to assess students' understanding of the Newtonian view of forces, their effects, and interactions; Hestenes et al. (1992a) refer to this as the 'Force Concept', which breaks down into the six dimensions discussed . However, in a factor analysis to determine the relationship between students' responses to the item dimensions, Huffman and Heller (1995) found little coherence between the responses to items in each dimension, suggesting that misconceptions do not fall neatly into the dimensions. Huffman and Heller suggest that this is due to students having an ill-defined Force Concept, with their understanding being formed from pieces of correct and incorrect ideas that are used to assess individual situations. Hestenes and Halloun (1995) countered that the data provided by Huffman and Heller supported their original analysis of the validity of the FCI, and due to the interlinked nature of the dimensions of

the FCI, it would be expected to see issues across a range of dimensions. Further, they point out that there was no factor analysis performed to find coherence amongst the incorrect responses which would have allowed a holistic view of a student's misconceived Force Concept.

Scott et al. (2012) performed a similar factor analysis on the FCI responses of over 2,000 Health Science university students and found coherence amongst five factors (SSG5 model) that did not completely align with the original six dimensions. They also determined that there was an overarching factor that showed conceptual coherence across the entirety of the FCI; this could be considered the 'Force Concept' as originally intended, the SSG5 factors are shown in table 2.3.

Table 2.3 *The Assignment of FCI Items to Factors*

Factor Number	Factor Classification	Item
1	Identification of forces	5, 11, 13, 18, 30
2	Newton's First Law with zero force	6, 7, 8, 10, 12, (16), 24, (29)
3	Newton's Second Law and Kinematics	19, 20, 21, 22, 23, 27
4	Newton's First Law with cancelling forces	(16), 17, 25
5	Newton's Third Law	4, 15, 28
0	Unassigned questions	1, 2, 3, 9, 14, 26

(No.) indicates low loading factor Adapted from Scott et al. (2012)

In this factor analysis, correlation between the responses to items was determined. This is distinct from the correlation between item questions and Newtonian concepts as written by expert Newtonian thinkers (Scott et al., 2012). As such, the commentary by Hestenes and Halloun (1995) on earlier factor analyses by Heller and

Huffman (1995) can be applied to this analysis, in that the variation of responses from non-Newtonian thinkers is likely to be due to incomplete and incoherent models of the 'Force Concept' that are not only incorrect but lead to contradictory explanations of events.

Evidence from Scott et al. (2012) shows that student responses can be aligned with identified aspects of the 'Force Concept', as shown in Table 2.3. Factor 1 was found to relate to the identification and classification of forces in a variety of situations, with a particular emphasis on the force of gravity. This is a shift of focus from the 'kinds of forces' grouping used by the authors of the FCI. This factor correlates highly to the overall achievement of students in the FCI, as the ability to identify forces underpins the other concepts in the FCI. As Scott et al. (2012) comment, it is unlikely a student would be able to determine the trajectory of an object if they are unable to identify the forces acting upon it. Factor 2 aligns closely with the First Law Dimension Concepts 1.1, 1.2 and 1.3 (Table 2.2), with the addition of Items 12 and 29, which rely on an understanding of these concepts to determine the path of an object. While Items 16 and 29 are found in this factor, they have a small correlation, and their inclusion may be due to the interpretation required for the situation. The third factor aligns a range of items from different FCI Dimensions (19, 20, 21, 22, 23, 27); Kinematics (0), Newton's Second Law with constant force (2.2), Newton's First Law with zero force (1.1), and Kinds of Force (5.3). This may be due to Items 19 and 20 both examining position plots, and 21, 22 and 23 relying on an understanding of the same situation. These two groups both rely on a clear understanding of position, velocity and acceleration and the ability to describe these concepts visually; hence students who can correctly answer Items 19 and 20 are more able to successfully respond to Items 21, 22 and 23. Item 27 is distinct from the others in this factor as it probes the understanding of an unbalanced force and so relies on an understanding of Newton's Second law. Newton's First Law with

cancelling forces (1.4) is the defining concept for Factor 4. Although Item 16 is not classed as such in Table 2.2, it requires a similar analysis of cancelling forces to the other items in the factor. Further, the loading factor is the lowest in this group, which leads the authors to classify Item 16 independently as a conflation of Newton's First and Third Laws, in which students may attain the correct response with incorrect reasoning. The Dimension of Newton's Third Law (concepts 3.1 and 3.2) is clearly aligned with Factor 5, as both HWS6 and SSG5 group items 4, 15, and 28 together.

The analysis identified that responses to Items 1, 2, 3, 9, 14 and 26 did not correlate with other items in the five-factor analysis. However, they did correlate with the other items in a single factor analysis, indicating that they measure some aspect of the 'Force Concept'. The authors attribute this to a combination of the difficulty of the items and the level of sophistication required to understand the situation in which the item is set, or some underlying unidentified misconception. From the factor analysis of responses, there appears to be no overarching Kinematics factor; the items in the Kinematic Dimension (0) are spread across Factors 2 and 3, indicating that students' ideas relate kinematics concepts more closely to Newton's First and Second Law concepts than to each other.

In a confirmatory factor analysis, Eaton and Willoughby (2018) investigated the alignment of 20,822 student responses on the dimensions determined by Hestenes et al. (1992b) (HWS6), the factors determined by Scott et al. (2012) (SSG5), and a model of their development (EW5). Their model was developed from consideration of the FCI items to form expert-like models. The factors developed are shown in Table 2.4.

Table 2.4 *FCI Factor Classification EW5*

Factor Number	Factor Classification	Item
1	First Law and Kinematics	6, 7, 8, 10, 20, 23, 24.
2	Second Law and Kinematics	9,12, 14, 19, 21, 22, 27
3	Third Law	4, 15, 16, 28
4	Force identification	5, 11, 13, 18, 30
5	Mixed	17, 25, 26

(Eaton & Willoughby, 2018)

Both the EW5 and HWS6 models are developed through experts grouping the items in line with Newtonian concepts, whereas the SSG5 model is developed through an exploratory factor analysis of students' responses. In the EW5 model, Factors 1 and 2 relate to Newton's First and Second Laws respectively, with associated Kinematics that relate to path identification and changes in speed resulting from zero and resultant forces. These factors align with the HWS6 model Dimensions 1 and 2 with the associated Items from 0. The EW5 Factor, described as Newton's Third Law, aligns with the HWS6 Dimension 3, except for Item 16, as this was identified as a First Law concept by the SSG5 model. All three models group Items 5, 11, 13, 18 and 30 together as Force Identification (SSG5 & EW5) or Kinds of Force (HWS6). These items have shown strong correlation amongst student responses and describe forces in a constant velocity situation (Eaton & Willoughby, 2018). The Mixed Concepts factor from EW5 aligns exactly with HWS6 Dimension 4, the former choosing the term to describe the mixture of concepts required to correctly respond to these Items. The EW5 model omits Items 1, 2, 3, 7 and 29 from analysis; this is due to their being unassigned in the SSG5 model. As the EW5 model uses some of the findings from Scott et al. (2012), the

authors consider it a hybrid model that allows the expert model to better fit the data received from student responses.

In their analysis of these three models, Eaton and Willoughby (2018) found that all provide an acceptable model of student responses to the FCI, with the EW5 model having the closest correlation for both large and small cohorts. As the expert factor model of the EW5 aligns with student responses, they suggest that the factors can be used to indicate the areas of misconception for students post-testing and the FCI can be graded in chunks to determine a student's understanding of each factor.

In an analysis of the relationship between overall achievement and the response to individual items, Wang and Bao (2010) developed item response curves for each item, which give the probability of a student achieving a correct response to an item in relation to their proficiency. They determined that Items 5, 13, and 18 are the most discriminatory between high and low proficiency students, Items 1 and 6 were the easiest, Items 25 and 26 the most difficult, and that low proficiency students have a 1 in 3 probability of correctly responding to Item 16. These results generally conform to the findings of Scott et al. (2012), who found Items 1 and 6 were amongst the most common correctly responded to, Items 25 and 26 were amongst the most challenging, and that almost 80% of students responded correctly to Item 16.

The internal consistency of an assessment item is a method of measuring the consistency of different parts of the item with each other in their measurement of the objective—in the case of the FCI, the consistency of individual items with each other in their measurement of the dimension or factor. Lasry et al. (2011) reported that the FCI exhibits a high internal consistency ($r = 0.8$), and hence measures a unique construct, such as the 'Force Concept'. They also investigated the stability of student responses between test and retest for a cohort with no instruction between tests. They indicated that students were more likely to change their response if initially incorrect than if

initially correct, but showed that 8% of responses changed from correct to incorrect. There is no information provided on the overall proficiency of these students, or the factors that were particularly susceptible to these changes. It may be that students obtained the correct responses with flawed reasoning (Lasry et al., 2011) or an ill-defined conceptual model that was subject to fluctuation, which would also account for the high level of wrong-to-different-wrong variations. The stability of student responses between pre- and post-test, was further investigated by Miller et al. (2010), with a focus on the effect of losses on the use of normalized gain as an effective reporting metric—i.e., students who score lower in the post-test than the pre-test (Miller et al., 2010). illustrate four possible transitions of responses between pre- and post-test, as shown in Table 2.5.

Table 2.5 *Possible Transitions of Answers between Pre- and Post-test*

Pre-test	Post-test	Transition	Average % Transitions 1991–1996
Right	Right	RR	66
Wrong	Wrong	WW	10
Right	Wrong	RW	3
Wrong	Right	WR	21

(Miller et al., 2010)

The desired outcome of any course would be to achieve a high level of WR transitions and 0% WR transitions, as this would indicate students had developed a better understanding of the concepts and reinforced their previously correct models.

When using a pre- and post-test research methodology in education research, a common goal is to measure the improvement or academic gain of students. In the use of the FCI this is commonly achieved through the use of a normalized gain (g) (Hake, 1998; Miller et al., 2010); however, the calculation and use of g as a reporting and comparative metric of academic gain has received some refinement and criticism in the literature (Madsen et al., 2017; Marx & Cummings, 2007; Miller et al., 2010). Hake introduced the average normalized gain as a metric for comparing the effectiveness of

different forms of course instruction in improving conceptual understanding of the Newtonian Force Concept. Hake's average normalized gain (<g>) is defined as the ratio of the difference in total score between the pre- and post-test to the maximum possible increase in score or the ratio of the actual average gain (<G>) to the maximum possible average gain (<G_{max}>).

$$< g > = \frac{\% < G >}{\% < G_{max} >} = \frac{(\% < S_f > - \% < S_i >)}{(100 - \% < S_i >)}$$

Where $< S_f >$ is the class average final (post unit) score and $< S_i >$ is the class average initial (pre-unit) score (Hake, 1998). Hake defined high gain courses as those with <g> ≥ 0.7, medium gain as those with 0.7> <g> ≥ 0.3, and low gain courses as those with <g> <0.3.

This equation is frequently presented as:

$$g_{ave} = < (Post - Pre) \div (100 - Pre) >$$

(Coletta & Phillips, 2005; Madsen et al., 2017; Miller et al., 2010).

Issues with the use of normalized gain have been raised around a number of features, including the independence of normalized gain scores from students' initial knowledge or pre unit score (Coletta & Phillips, 2005), issues with the handling of student dropout (i.e. students who do not complete the post-unit test) (Madsen et al., 2017), and the impact of students whose performance on the post-test is worse than the pre-test or transition from correct to incorrect responses (Miller et al., 2010).

Coletta and Phillips (2005) showed a positive correlation between pre-instruction score and normalized gain amongst the cohorts they audited. They conducted a further study to try to determine the reason behind the correlation, and suggested that a higher level of operational thinking gives students a better pre-unit FCI score and also allows them to develop their understanding through improved engagement with, or understanding of, the learning activities of the course. These

findings highlight the importance of considering the initial characteristics of any cohort being studied, particularly when a comparative methodology is being used.

The issue of students not completing both the pre-unit and post-unit course is particularly relevant to courses with high drop-out rates. To counter this effect, Madsen et al. (2017) suggest only using matched student data when performing cohort analysis. This is the methodology adopted by PhysPort and this study.

In a meta-analysis of studies using normalized gain in reporting FCI results, Von Korff et al. (2016) reported some variation in the calculation of the average normalized gain. The two approaches used were classified as the gain of averages $\langle g \rangle$ and the average of gains g_{ave}. The former is the original method proposed by Hake; however, the average of gains is commonly used and presents some advantages, including a more meaningful relation between individual student gains and comparative cohort average gains (Von Korff et al., 2016). In situations where students have a negative gain, i.e., they perform more poorly on the post- than pre-test, Marx and Cummings (2007) suggest the use of normalized change. The normalized change approach is summarized below:

$$c = \begin{cases} \dfrac{post - pre}{100 - pre} \; when \; post > pre \\ drop \; when \; pre = 100 \; or \; post \; 0 \\ 0 \; when \; post = pre \\ \dfrac{post - pre}{pre} \; when \; post < pre \end{cases}$$

Further, the use of c_{ave} is suggested because it more accurately catches the spread of student achievements.

The use of is c_{ave} has been adopted by PhysPort, a Physics education research platform developed by the American Association of Physics Teachers to provide analysis of data gathered from a variety of concept inventories, in calculating the gains of students in the FCI.

A further metric commonly used in the measurement of the impact of teaching and learning activities is effect size (Hattie, 2009; Madsen & McKagan, 2017; Melbourne Graduate School of Education, 2018). Effect size is a measure of the difference between two results and the importance of the difference; it provides a comparison of the average raw gain to the standard deviation of individuals' raw scores. Effect size is calculated using Cohen's *d*, which is calculated from:

$$d = \frac{(< post > - < pre >)}{stdev}$$

Where *stdev* is the pooled standard deviation of the pre- and post-test scores and <*pre*> and <*post*> are the class average pre- and post-test scores.

Effect sizes can be used to measure how substantially students' knowledge of a subject has changed, and to compare different cohorts or treatments. The magnitude of the *d* value is often grouped from small to large effects. The inclusion of standard deviation (which relates to the number of students) in the calculation gives a fairer comparison between cohorts of different sizes than some other metrics (Coe, 2017).

As students respond to the FCI in a multiple choice format, DeVore et al. (2016) highlight the possible impact of 'testwiseness' on student score. 'Testwiseness' is defined as the set of strategies that students may use to improve their score when responding to multiple choice responses. This includes strategies such as avoiding 'none of the above' or 'zero' distractors. In a study of the impact of these two strategies in the FCI, DeVore et al. (2016) found the combined impact to be 0.58% from 'none of the above' distractors (Qs 13, 17, 20, 29) and 0.55% from 'zero' distractors (Qs 11, 15, 16, 29). A further issue with the use of multiple choice response formats is the possibility for false positives, whereby students achieve the correct response by incorrect reasoning or guessing (Yasuda et al., 2019). In the five-choice scale of the FCI there is a 20% chance of students guessing the correct response. The use of powerful distractor responses aims to minimize this chance, but in detailed interviews, Hestenes et al.

(1992b) found false reasoning in a number of students. By using sub-questions to determine whether a response to the FCI was a false positive, Yasuda et al. (2019) found Qs 6, 7 and 16 the most susceptible to false positives and determined that the use of Cohen's *d* is less susceptible to these systematic errors than average normalized gain.

Although these studies do raise some questions regarding the use of the FCI in analyzing student understanding, they indicate precautions that must be taken in analyzing the results, rather than negating their importance and significance, and provide opportunities for a more detailed analysis of student understanding than a simple raw score or gain. In this study, the results will be analyzed in view of the FCI dimensions, the level of right to wrong responses will be reviewed, and normalized change and effect size will both be calculated as described, and reported.

2.7 Student Attitudes

When considering a student's attitude towards any subject, it is important to define the components of attitude that are being discussed. Attitude can be considered to have three distinct components: the cognitive, which refers to perceptions, beliefs and understanding; the affective, which includes emotional reactions; and the behavioural, which relates to the predisposition to action (Angell et al., 2008; Rosenberg & Hovland, 1960).

Gardner (1975) identified that students' attitudes towards Science and scientific attitudes are two distinct areas relevant to science education. The latter refers to a mix of the desire for knowledge, a respect for logic, and a consideration of evidence. These attributes contribute to scientific thinking and are cognitive in nature (Osborne et al., 2003). A student's attitude towards Science describes a set of affective behaviours that relate to a variety of parameters. Klopfer (1971) identified and separated these parameters into six categories: attitudes to science and scientists, attitude to inquiry, adoption of scientific attitudes, enjoyment of science learning experiences, interest in

science and science related activities, and interest in a career in science. Osborne et al. (2003) reviewed the wide range of approaches used to measure student attitudes towards Science and grouped approaches in five main groups:

- Preference ranking—a comparative measure of the ranking of Science compared with other school subjects.
- Attitude scales—the most commonly used method is through a Likert scale to indicate agreement or disagreement with a statement related to an attitudinal construct.
- Interest inventories—students choose items that they are most interested in from a list, and this is then compared to a 'scientist's' choice.
- Subject enrolment—this involves the collection of data on the numbers of students enrolled in Science subjects.
- Qualitative methods—student interviews and group discussions are used to ellicit an understanding of students' attitudes towards various aspects of Science.

Student attitudes towards a subject can have a significant impact on their engagement and attainment; therefore, they have been the focus of a substantial body of research with the aim of increasing interest, performance and student numbers in science subjects (Kind et al., 2007; Siegel & Ranney, 2003). Studies have reported a positive relationship between science attitude and science achievement; this can partly be attributed to the process by which attitudes affect students' persistence and hence performance (Schommer, 1994). For these reasons, it is important to understand the impact any novel approach to teaching and learning of Science has on student attitudes towards the subject (Siegel & Ranney, 2003).

2.7.1 Test of Science Related Attitudes

The Test of Science Related Attitudes (TOSRA) instrument was developed to measure science-related attitudes amongst secondary school students (Fraser, 1982) (see

Appendix 3 for sample questions from the scales used in this study). This instrument is used to measure a student's attitude towards a range of aspects of Science and relates closely to Klopfer's categories (Fraser, 1982; Fraser & Butts, 1982). In its full form it consists of seven scales of ten items each; however, some researchers have selected specific scales from the TOSRA to focus on particular attitudes (Madu, 2010). These seven scales and their alignment with Klopfer's categories are shown in Table 2.6.

Table 2.6 *Classification of Each Scale in TOSRA*

TOSRA Scale name	Klopfer Classification
Social Implications of Science (S)	H.1 Manifestation of favourable attitudes towards science and scientists
Normality of Scientists (N)	
Attitude to Scientific Inquiry (I)	H.2 Acceptance of scientific inquiry as a way of thought
Adoption of Scientific Attitude (A)	H.3 Adoption of 'scientific attitudes'
Enjoyment of Science Lessons (E)	H.4 Enjoyment of science learning experiences
Leisure Interest in Science (L)	H.5 Development of interest in science and science-related activities
Career Interest in Science (C)	H.6 Development of interest in pursuing a career in science

(Fraser, 1982)

The TOSRA I scale aligns with Klopfer's Category H.2, and measures a student's attitude to obtaining information by scientific experimentation and enquiry, as opposed to being the passive recipient of information. This is an important parameter as it conflates with the inquiry-based model of learning that many studies have found to have a positive effect in Science education (Hattie, 2009). Scale A aligns with Category H.3, and measures a group of attitudes that have been determined as important to the work of scientists, such as open-mindedness and willingness to revise opinions (Fraser, 1982). Scale E assesses enjoyment of Science lessons, as distinct from the enjoyment of Science in general; this is an important distinction, as students may enjoy aspects of Science beyond the classroom whilst being uninterested in classroom-based practice (Angell et al., 2008).

Each scale on the TOSRA consists of 10 items which students respond to on a Likert-type five-point scale consisting of Strongly Agree (SA), Agree (A), Not sure (N), Disagree (D), and Strongly Disagree (SD). Scoring involves allocating a 5–1 score for SA–SD respectively for positive items and the reciprocal for negative items. To ensure students respond honestly to the TOSRA, Fraser (1982) suggested ensuring students are aware that no grading is attached to their responses and that it is their honest opinion that is valued, not their adherence to perceived 'acceptable' attitudes.

Initial studies of the reliability of the TOSRA were conducted across Year 7–10 students in eleven schools, of varying socioeconomic, geographic and educational types, in the Sydney metropolitan area (Fraser, 1982). Internal reliability, the extent to which items in a scale measure the same attitude, was found to be good across the year levels. It was also determined that the cross-scale correlation was such that the scales were measuring distinct attitudes from each other. The test-retest reliability was also found to be good when assessed for a group of Year 8 and 9 classes, meaning that students' attitudes maintained stability as measured by the instrument. Since this initial research,

the TOSRA has been translated into numerous languages and has been validated across a number of cultural and education settings (Ali et al., 2013; Madu, 2010).

Criticisms of attitudes scales in science education stem from a range of concerns, namely: determining a clear construct for the variety of aspects of a 'scientific attitude'; conflation of results from the measurement of different aspects of the 'scientific attitude' into one score; and lack of internal consistency or unidimensionality of constructs (Kind et al., 2007). It has also been reported that attitudes towards Science, and other subject, can be resistant to interventions (Kind et al., 2007; Madu, 2010; Siegel & Ranney, 2003). The TOSRA has been shown to be resistant to these criticisms due to its adherence to Klopfer's aspects of 'scientific attitude', reporting of individual scale attitude scores, and the internal consistency demonstrated by a range of studies. Benefits of attitudinal scales include the increased reliability achieved by using repeated questioning to measure the same construct, the fact that they are simple to distribute to students, and that it is relatively simple to collate results from them (Kind et al. 2007).

2.8 Chapter Review

This chapter has provided a review of the literature of Mastery Learning and Blended Learning. It has discussed the use of the FCI to measure students' understanding of the Newtonian Force Concept and the use of the TOSRA to measure students' attitudes towards science.

Mastery Learning was defined as an emphasis on sequential learning, regular checks and feedback to students, corrective activities, and individually determined progression times. It has been shown to have a range of positive impacts on learning outcomes, though there is a lack of evidence in Australian high school Science classroom.

Blended Learning was defined as a focus on learning activities which integrate ICT activities with face-to-face learning. There is some evidence of benefits in the use of Blended Learning approaches; in a Science context this is particularly evident with the use of simulations and multimedia presentations of complex phenomena.

The use of the FCI to measure student understanding of the Newtonian concept of force has been common in a significant amount of Physics education research, and the literature shows that the FCI provides a useful method for comparing the effectiveness of different types of instructional pedagogy. Some research has shown that the original question groupings may not represent the concepts targeted; however, this may be due to the incomplete formation, and hence application, of students' concepts.

The measurement of students' attitudes towards science through the use of the TOSRA has been shown to be reliable across a range of cohorts.

The following chapter will outline the research design and methodology. An explanation of the construct, intervention, and measurement process will be provided.

Chapter 3: Research Design and Methodology

3.1 Chapter Overview

First, this chapter will outline the quasi-experimental research methodology that was adopted by this study, including the associated ethical considerations. Second, it will explain the research focus and questions that were investigated. Third, the use of a Blended Learning Mastery Progression Cycle as an intervention will be explained through the use of the commercially available Minds on Physics (MOP) platform. Fourth, the methodologies for measuring student attainment using the FCI, and student attitudes using selected scales from the TOSRA, will be explained, including the analysis of data through the use of a variety of metrics, significance tests, and correlations.

3.2 Introduction

This study sought to investigate the effect of combining a Mastery for Learning (ML) approach with Blended Learning (BL) activities; specifically the use of the Minds on Physics (MOP) software program to monitor the attainment of mastery and for the provision of correctives when required. The association of this approach with student attainment and attitude was estimated using pre- and post-treatment FCI and TOSRA scores respectively with a variety of quantitative measures. A quasi-experimental design methodology (Creswell, 2012) was used to compare the treatment group with a control group, for a sample population of Year 10 students in an Australian high school.

The construct model for this study relies on the premise that students have misconceptions in their understanding in the Year 10 Physical Science curriculum, and that different learning activities can lead to different outcomes in terms of attainment and attitude towards a subject area. The construct, intervention process, and

measurement methods are illustrated in Figure 3.1

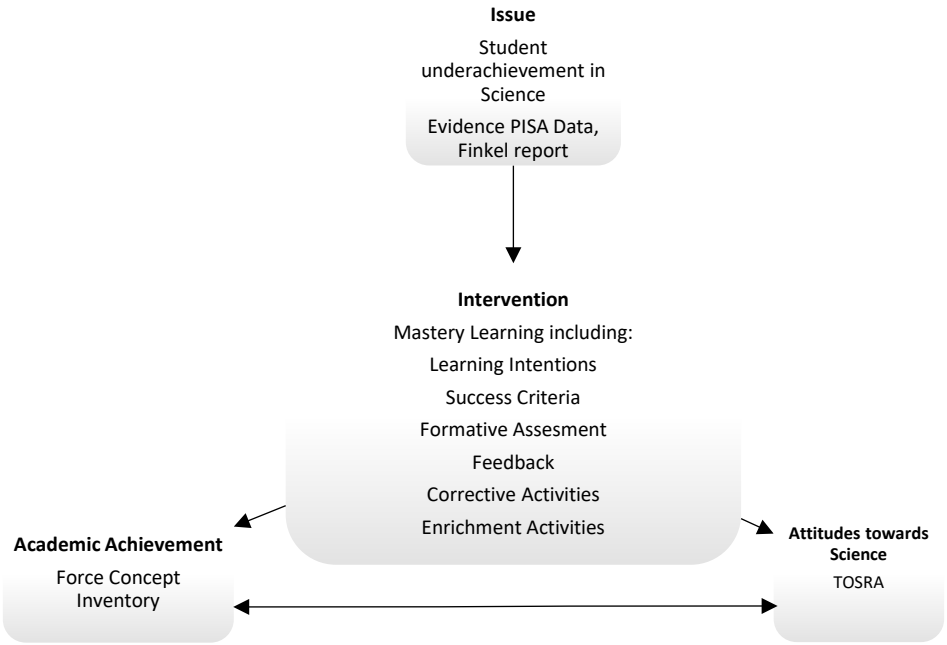


Figure 3.1 *Construct Model for Study*

3.3 Research Design and Approach

As the aim of this investigation was to determine the impact of a novel approach, it used a quasi-experimental design, with the independent variable being the treatment and the dependent variables being students' attainment and attitudinal outcomes. The investigation of correlation is a fundamental feature of experimental design. It is appropriate to class this study as a quasi-experimental methodology, as it is impossible to control all variables, such as pupils' prior experience, and emotional and developmental states. However, by comparing results from classes taught by the same teacher, using the same techniques, the impact of other variables was potentially minimized. The methodology also exhibited the quasi-feature of having a non-random assignment of participants to a group; this is because the group members were determined by the school's administration through their usual class structuring process. As a range of groups was exposed to each approach, it was assumed this had minimal effect on the results.

3.3.1 Ethical Considerations

In any educational research, there are ethical issues that must be considered. Such issues may range from the design of the study, to the conduct of the researcher, to the dissemination of findings, and many other factors. An overriding emphasis should be to do no harm, and where possible deliver benefit (Clark & Sharf, 2007). In educational research, this ethos can be extended to include not causing any disadvantage, as, for many, an ineffective education can have a lasting impact. It is with these points in mind that the ethical framework for this study was constructed.

As this investigation sought to compare different approaches to Science teaching, this could lead to ethical issues of equity if one of the approaches was considered to be, or was revealed to be, delivering improved student outcomes. This issue was addressed in two ways. Firstly, each method has both supporters and detractors in the published research and so in any school may be the standard delivery of instruction. Second, it was determined that if it became apparent in the early stages of the study that one approach was delivering improved student results, the study would be refocused to determine the reasons for the improvements, and all students would be exposed to the approach delivering the improved results.

Ethical concerns are frequently raised in the context of the allocation of participants to groups (Creswell, 2012) and groups to the treatment or control. In the school in which the study was conducted, the allocation of students to groups was based on criteria beyond the control of the researcher, this was usually based on streaming for Mathematics classes. Science was considered a core subject in this year group, and so all students were required to study the same core curriculum. Groups were allocated to approaches on a stratified sampling basis; after the study, all participants were given access to the treatment materials and provided with opportunities to further develop their subject understanding.

Participation in the study also had to be managed on ethical grounds (Creswell, 2012). Although participants were allocated to groups as part of the normal processes of the school, there may have been students who did not want to be involved in the study, or whose guardians did not want them to participate. To minimize the occurrence of this, in the first instance, participants and parents were asked for consent (see Appendix 5 for a copy of the consent letter). At this stage, they were given details of the reasons for the study and an outline of the procedures that were to be used, including the anonymity of all responses. It was made clear that participation in the extra assessment would have no impact on their grading for the subject. If participants chose not to be involved in the study, they were encouraged to stay in the group but were not involved in the analysis of performance or engagement.

An essential part of the ethical framework of a research project is in ensuring that the methodology is balanced and free from bias. It is important that the preconceived ideas of the researcher do not influence the research design process. In the case of this project, there was the possibility of bias being introduced in a number of areas—for example, in the favouring of a particular approach. To minimize this possible source of bias, it was ensured that each approach was equally well designed, resourced and delivered; this was monitored by an experienced teacher outside of the research team. To ensure consistency of delivery, the teaching phase of each unit followed the same plan and was delivered by the same teacher. Further potential bias issues arise in the reporting of findings. The MOP modules were developed by a third party, and care was taken to maintain objectivity. There were no financial or other relationships between the researcher and the MOP developers.

Of concern in any study is the reporting of the findings to participants after the study is completed. In this study, such feedback was provided at key points during the study. The most relevant information to students was their achievement on the FCI

assessments, as this was a good indicator of their understanding of the subject. Participants received their FCI pre- and post-unit scores, their normalized gain, and a list of topics that required further study. Review sessions were offered based on these misconceptions after the completion of the unit.

Ethics approval was granted by Curtin University Human Ethics Research Committee (Approval Number # HRE2017-0265), this indicated that the research plan was aligned with appropriate guidelines.

3.3.2 Selection and Description of Sample

The sample consisted of mixed gender classes from a Year 10 cohort in a Queensland high school, the control group consisted of n=104 (m= 46, f = 58) students and the treatment group consisted of n=95 (m= 45, f= 50) students, giving a total sample size N =199 (m= 91, f= 108).

This specific site was chosen due to several factors: the ease of access for the researcher; a 1–1 laptop scheme across the school; an excellent ICT infrastructure; and a high level of digital literacy amongst students. Classes were randomly allocated to the treatment or control groups. All students were taught by the same teacher and followed the same curriculum to minimize the effect of these parameters on the validity of the study. Although it perhaps would have been beneficial to increase the sample size by including a more extensive and diverse, in age and socioeconomic background, range of classes and schools, this may also have introduced a wide range of variables into the study.

3.4 Research Questions

The overarching research question for this study was:

Does a combination of the application of a Mastery for Learning (ML) approach with Blended Learning (BL) activities—specifically the use of Minds on Physics—affect student academic performance and attitudes towards science?

This question was addressed by considering the following research questions:

1 Student Attainment

RQ 1a: Are pre-unit FCI scores for the treatment and control groups statistically significantly different?

RQ 1b: Are there statistically significant differences between the pre- and post-unit FCI scores for the control group?

RQ 1c: Are there statistically significant differences between the pre- and post-unit FCI scores for the treatment group?

RQ 1d: Are post-unit FCI scores for the treatment and control groups statistically significantly different?

2 Student Attitudes

RQ 2a: Are pre-unit TOSRA scores for the treatment and control groups statistically significantly different?

RQ 2b: Are there statistically significant differences between the pre- and post-unit TOSRA scores for the control group?

RQ 2c: Are there statistically significant differences between the pre- and post-unit TOSRA scores for the treatment group?

RQ 2d: Are post-unit TOSRA scores for the treatment and control groups statistically significantly different?

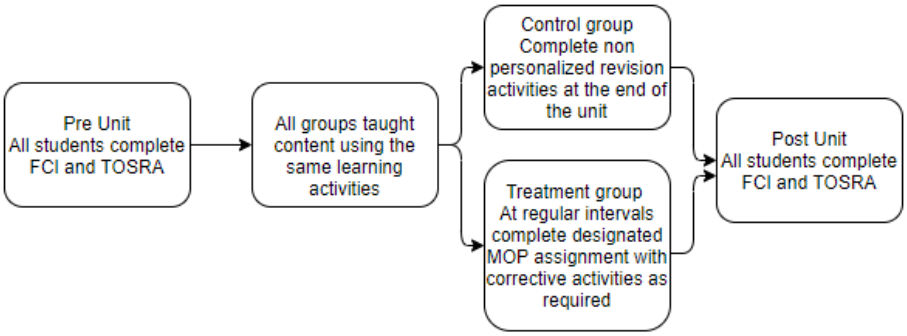
3 Associations

RQ 3: Are there associations between FCI scores and student attitudes towards Science as measured by the TOSRA scales?

3.5 Research Plan

The study involved all student groups completing a pre-unit assessment of understanding (FCI) and attitude (TOSRA). All students were then exposed to the same learning activities devised by the classroom teacher as most appropriate for that cohort. The control group continued through the course content in a linear manner followed by working through non-personalized revision material, whilst the treatment group completed the relevant MOP module at the end of each subtopic. After the unit, all students were reassessed for knowledge (FCI) and attitude (TOSRA). The methodology process is illustrated in Figure 3.2.

Figure 3.2 *Flowchart to Illustrate the Study Methodology*



3.5.1 Treatment

The treatment and control group completed the same learning activities, as devised by the teacher, in line with the ACARA curriculum requirements for Year 10 Physical Sciences content descriptor ACSSU229 ‘The motion of objects can be described and predicted using the laws of Physics’(ACARA, 2020). The pedagogical approach used was based on the Modeling Instruction framework which places an emphasis on the construction and application of conceptual models of physical phenomena as a central aspect of effective learning activities in Science (Jackson et. al, 2008). The curriculum was delivered over a 10-week period with three 50-minute sessions each week. At appropriate intervals, usually after the completion of a group of related learning intentions, the treatment group engaged in the appropriate Blended

Learning Mastery Progression Cycle (BLMPC) activities in the form of Minds on Physics (MOP) modules. The alignment of the MOP modules with ACARA curriculum statements and course learning intentions is shown in Table 3.1, students in the treatment group were directed to the appropriate MOP modules on the completion of the related learning activities.

Table 3.7 *Alignment of ACARA Curriculum with Learning Intentions and MOP Mission*

ACARA Curriculum Descriptor elaboration	Course Learning Intention	MOP Mission
Analyse everyday motions, such as measurements of:	To understand the distinction between a vector and scalar quantity.	KC1 Vectors v Scalers
Distance and time	To distinguish between distance and displacement of an object to analyse its motion.	KC2 Distance v Displacement
Speed and velocity	To distinguish between the speed and velocity of an object to analyse its motion.	KC3 Speed and Velocity
	To use the average speed equation to calculate relevant quantities to describe an object's motion.	KC6 Average Speed calculation
Acceleration	To understand the definition of acceleration and apply this to an analysis of an object's motion.	KC4 Acceleration
	To use the acceleration equation to calculate relevant quantities to describe an object's motion.	KC7 Acceleration calculation
Gather data to analyse: Distance and time	To identify the features of a Position-Time graph and relate these to the motion of an object.	KG1 Basics of Position-Time Graphs
	To calculate the distance travelled and speed of an object from a P-T graph.	KG2 Shape and Slope of P-T Graphs
		KG3 Matching Motion and Shape for P-T Graphs
		KG4 Slope Calculations for P-T Graphs
Speed	To identify the features of a V-T graph and relate these to the motion of an object.	KG5 Basics of Velocity-Time Graphs
Distance and acceleration	To calculate the distance travelled and acceleration of an object from a V-T graph.	KG6 Shape and Slope of V-T Graphs
		KG7 Matching Motion and Shape for V-T Graphs
		KG8 Slope and area Calculations for V-T Graphs
Forces	To identify the forces acting on an object.	NL4 Types of Forces
	To draw and interpret force diagrams.	NL5 Force Diagrams
Force and Mass	To distinguish between the mass and weight of an object and	NL6 Mass V Weight

	understand the relationship between the two.	
A stationary object, or a moving object with constant motion, has balanced forces acting on it.	To understand the concept of inertia, Newton's First Law and the consequences for describing the forces acting on an object.	NL1 Inertia and Newton's First Law
	To determine the forces acting on an object by considering the object's motion.	NL2 Balanced Forces
	To determine the net force acting on an object and the effect on motion.	NL3 Unbalanced Forces and Accelerations
Using Newton's Second Law to predict how a force affects the motion of an object	To determine the relationship between the resultant force, mass, and acceleration of an object.	NL7 Newton's Second Law
	To use Newton's Second Law to calculate the force, mass or acceleration of an object.	NL8 F=ma Calculations
Recognizing and applying Newton's Third Law to describe the interactions between two objects	To recall Newton's Third Law and apply it to situations to identify force pairs.	NL12 Newton's Third Law

3.5.2 Role of Researcher

The primary researcher was responsible for the development of the study parameters, the teaching of all the classes involved in the study, the development of course materials, the delivery of TOSRA and FCI assessments, the collation and analysis of data, and the writing of the report. Ethical considerations of this approach are discussed in section 3.3.1.

3.5.2.1 Minds on Physics

Minds on Physics (MOP) is a commercially available online resource that uses students' responses to questions to assess understanding of Physics topics and direct students to remedial activities as required. Teaching staff at Glenbrook South High School, Glenview, Illinois, USA, originally developed the MOP modules after receiving

a grant from the National Science Teachers Association; with the aim of improving their students' understanding of the Physics curriculum. The MOP modules were developed with Mastery Learning and Blended Learning as cornerstones of the approach (T. Henderson, personal communication, April 24, 2016). Other Mastery-based learning programs are available across a range of subjects disciplines (Green, 2016) (atomi.com.au, Pearsonmylabandmastering.com, physicsclassroom.com, Mathletics.com.au, mathspathway.com). More recently, the MOP modules have been integrated into the Physicsclassroom.com website and developed into a series of apps; this provides a more substantial range of resources for students.

MOP consists of 15 modules designed to probe student understanding of Physics concepts and subsequently correct student misconceptions. Each module contains a series of topics that are further divided into assignments with clearly specified learning objectives. The assignments consist of a bank of multiple-choice questions and problems organized into groups, based on the learning objectives. As students attempt questions, they receive immediate feedback concerning their level of success, and, if repeatedly unsuccessful on a group of questions, they are directed to a location on an instructional webpage for remediation. These Blended Learning correctional activities range from instructional text, to interactive animations, to suggested experimental activities. Students then re-attempt the assignment with a different set of questions from the bank (Henderson, 2016). This cycle reflects the four-stage process outlined by Guskey (2007). Once students have successfully demonstrated mastery of a topic, by obtaining 85% on an assignment, they are rewarded with 'medals' and success codes; these allow the class teacher to check the progress of individual students (Henderson, 2016). Exemplars of the multiple-choice questions, problems, and corrective activities are provided in Appendix 4.

MOP currently has approximately 350 registered teacher accounts and regularly receives traffic of 5–6,000 visitors per day. There is currently anecdotal evidence from teachers and students as to the effectiveness of MOP, but no formal systematic study into the efficacy has been conducted (T. Henderson, personal communication, April 24, 2016). For this study, the most appropriate MOP modules were Kinematic Concepts, Kinematic Graphing and Newton's Laws of Motion, as these most closely match the ACARA Year 10 Physics curriculum (ACSSU229) (ACARA, 2020). See Figure 3.3 for a flowchart of the MOP process.

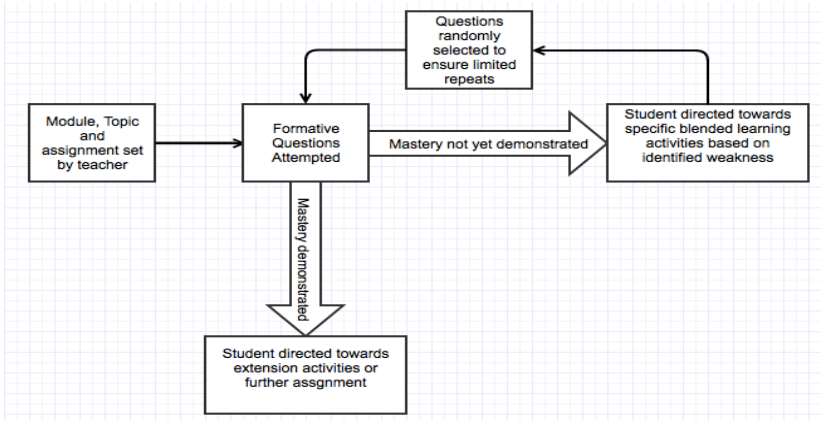


Figure 3.3 *Flowchart to Illustrate MOP Process*

The MOP system was selected for inclusion in this study due to its alignment with the underlying principles of Mastery Learning, its use of various multimedia resources as interventions, its alignment with the ACARA Physical Science curriculum, and its commercial availability.

3.6 Data Collection and Instrumentation

3.6.1 Force Concept Inventory

The FCI was chosen to ascertain student attainment by measuring the understanding of the concepts associated with Newtonian mechanics and the underlying principles required in the description of motion. The FCI was selected due to its alignment with the topics in the Year 10 ACARA curriculum and the high level of research into the efficacy of the inventory. Further, the FCI results can be subjected to various analyses to measure and compare student gain, to indicate areas of student

misconception and coherence of understanding. The normalized change was determined for each student, and a comparison between treatment and control group was made. Further associations between initial result and post-unit result were investigated, including the effect size and responses to questions clustered in the dimensions of the Force Concept. Aspects of the reliability of the FCI were measured by conducting a detailed analysis of individual and cluster item responses to measure the stability of students' responses, with a particular focus on right-to-wrong transitions.

3.6.2 Test of Science Related Attitudes

The TOSRA was chosen as it has high reliability and validity in a high school setting (Fraser, 1981) and has been used extensively in previous studies (Fraser & Butts, 1982; Madu, 2010; Welch & Huffman, 2010). The TOSRA also provides a profile of student attitudes, rather than the single score that some other instruments provide, thus allowing more detailed analysis to be performed and avoiding conflating different attitudes (Angell et al., 2008). To prevent students from selecting middle non-committal responses, as a default this option was removed, thus forcing respondents to make either a positive or negative response to the item (Cavanagh & Romanoski, 2007). The correlation of FCI score with student attitudes towards Science was also estimated using a Pearson's *r* correlation. Cronbach alpha values were also calculated to measure the internal consistency of the scales in the four-point Likert scale format.

The scales used were Attitude to Scientific Inquiry (I), Enjoyment of Science Lessons (E), and Adoption of Scientific Attitudes (A), these scales were chosen as they have been shown to have correlation with attainment in science (Madu, 2010)

Table 3.8 *Summary of Research Question, Data Instrument and Analysis*

Research Question	Instrument	When	Who	Analysis
Are pre-test FCI scores for the treatment and control groups statistically significantly different?	FCI	Pre-unit	Treatment and Control groups	t-Test to compare the mean score of the groups
Are there statistically significant differences between the pre- and post-unit FCI scores for the control group?				t-Test to compare the mean score of the groups Normalized changed Effect size
Are there statistically significant differences between the pre- and post-unit FCI scores for the treatment group?				t-Test to compare the mean score of the groups Normalized changed Effect size
Are pre-test TOSRA scores for the treatment and control groups statistically significantly different?	TOSRA Scales I, E, A	Pre-unit	Treatment and control groups	t-test to compare the mean score for each scale
Are there statistically significant differences between the pre- and post-unit TOSRA scores for the control group?		Post-unit		t-test to compare the mean score for each scale. Effect size
Are there statistically significant differences between the pre- and post-unit TOSRA scores for the treatment group?		Post-unit		t-test to compare the mean score for each scale. Effect size
Are there associations between FCI scores and student attitudes towards Science as measured by the TOSRA scales?	FCI and TOSRA	Pre- and Post-unit	Treatment and control groups	Comparison of FCI and TOSRA data through Pearson correlation

3.6.3 FCI

The data obtained from pre- and post-unit FCI was analyzed to determine correlations between treatment and a range of outcomes; raw pre-unit and post-unit score, average normalized gain, effect size, individual question response, and dimension response. The following sections will outline the methods for determining these factors.

3.6.3.1 Raw Score

The raw FCI score is reported as a score on a 30 point scale. Pre- and post-unit scores were compared to give a measure of any change, and these were then compared between control and treatment groups. Group mean scores were calculated and a series of t-tests were conducted to measure the statistical significance of any difference between groups and between pre-unit and post-unit scores, with p values <0.05 indicating that any difference in scores is significant.

3.6.3.2 Normalized Change

Normalized change is used to measure the effectiveness of a course in promoting conceptual understanding (Hake, 1998; Marx & Cummings, 2007). The average of gains was calculated by uploading data to PhysPort for analysis (Madsen et al., 2017). Average of gains was calculated using:

$$c_{ave} = < (Post - Pre) \div (100 - Pre) >$$

In cases where a negative gain was achieved, or students scored 100% on the pre-test, the following procedure was used:

$$c = \begin{cases} \dfrac{post - pre}{100 - pre} \; when \; post > pre \\ drop \; when \; pre = 100 \; or \; post \; 0 \\ 0 \; when \; post = pre \\ \dfrac{post - pre}{pre} \; when \; post < pre \end{cases}$$

A comparison of c_{ave} was made between control and treatment groups.

3.6.3.3 Effect Size

The effect size measures the change between the pre-unit and post-unit test results for a group and gives a measure of the difference between the two. One method of measuring effect size is to use Cohen's *d*, and this was calculated using:

$$\boldsymbol{Effect\ size\ (d) = \frac{< post > - < pre >}{Standard\ deviation}}$$

(Hattie, 2009; Madsen et al., 2017).

Effect sizes were calculated for both the control and treatment groups and a comparison performed. In an educational context, Hattie (2009) describes effect sizes of $d = 0.2$ as small, $d = 0.4$ as medium, and $d > 0.6$ as large. It is further suggested that interventions that have positive effect sizes of $d > 0.4$ should be considered useful in improving student achievement (Hattie, 2009).

3.6.3.4 Individual Question and Dimension Response

The number of correct responses per question was recorded and calculated as a percentage of the group. This allowed a comparison on an item-by-item basis of the treatment and control groups. In the FCI, questions are grouped into five dimensions that together measure the Newtonian Concept (Hestenes et al., 1995). By combining responses into these dimensions, an analysis of student adoption of the various aspects of the Newtonian concept was made between control and treatment groups. This grouping was performed by uploading group data to PhysPort for analysis.

Further analysis was performed to investigate the stability of student responses; this involved an analysis of students who selected the correct response in the pre-unit test but selected an incorrect response in the post-unit test, a right-to-wrong (RW) transition (Miller et al., 2010).

The percentage of students who responded correctly in the pre-unit test and incorrectly in the post-unit test was calculated using:

$$\%RW = \frac{Number\ of\ RW\ Students}{Number\ of\ initial\ correct\ responses} \times 100$$

This lack of consistency may indicate a range of issues, including good guessing, incompletely formed Force concept, and regression of understanding (Miller et al., 2010). A comparison was made between treatment and control groups in terms of the consistency of correct response.

3.6.4 TOSRA

TOSRA data, scores for each of the three scales used, was analyzed to measure any changes in student attitudes between the start and completion of the unit. Comparisons were then made between control and treatment groups. A two-tailed *t* test was performed to test the statistical significance of any difference between pre- and post-unit responses to each of the TOSRA scales being used. *P* values were calculated to determine the statistical significance of any differences, with values of <0.05 considered as indicating a statistically significant difference in scores (Muijs, 2013). Cronbach alpha values were calculated to test internal consistency of the TOSRA scales for both treatment and control groups, and these were compared to published scores from other applications of the TOSRA.

3.6.5 Correlations between FCI and TOSRA

Relationships between student responses to TOSRA scales and the FCI were investigated by performing two-tailed Pearson tests to assess correlations between pre-unit TOSRA and pre-unit FCI, and pre-unit TOSRA and post-unit FCI. This was conducted to determine any correlations between student attainment and attitudes towards science.

3.7 Chapter Review

This chapter has outlined the research methodology adopted by this study and explained the research focus and the questions that were investigated. The quasi-experimental design methodology was explained in terms of the application of an

intervention and measurement of its impact through pre- and post-test design, and the ethical implications of intervention design were discussed. The implementation of the intervention was explained as the use of a Blended Learning Mastery Progression Cycle, in which students in the treatment group completed a computer-based quiz on nominated ACARA curriculum descriptors, followed by a range of Blended Learning corrective activities as required. This cycle was provided by the use of the commercially available MOP platform. Outcomes were measured in terms of student attainment using the FCI, and student attitudes using selected scales from the TOSRA. The analysis of data was outlined through the use of a variety of metrics, significance tests, and correlations designed to investigate the research questions.

Chapter 4: Results

4.1 Chapter Overview

This chapter will first describe the characteristics of the cohort involved in the study. Descriptive statistics of the results of the FCI and TOSRA scales will be presented for the control and treatment groups. This consists of mean pre- and post-unit scores, the distribution of scores, and FCI normalized change and effect size. Inferential statistics will then be presented for the control and treatment groups, arranged in order to respond to the research questions. This includes the standard deviation in the FCI score and a t-test comparison of the pre- and post-unit control and treatment group scores, the standard deviation of each TOSRA scale, and t-test comparison of the pre- and post-unit control and treatment group scores. Associations between the FCI and TOSRA scale scores are shown through the use of a Pearson's correlation.

4.2 Introduction

The purpose of this study is to investigate the effect of a combining a Mastery for Learning (ML) approach with Blended Learning (BL) activities—specifically, the use of the MOP software program to measure the attainment of mastery, and the provision of correctives when required. The results presented in this chapter were obtained using a quasi-experimental design methodology (Creswell, 2012) in which pre- and post-treatment FCI and TOSRA instruments were used to determine correlations between the study factors. The study population was a mixed cohort of Year 10 Students in an Australian Catholic high school (n = 199).

The results were used to answer the research questions:

RQ 1 Student Attainment

RQ 1a: Are pre-unit FCI scores for the treatment and control groups statistically significantly different?

RQ 1b: Are there statistically significant differences between the pre- and post-unit FCI scores for the control group?

RQ 1c: Are there statistically significant differences between the pre- and post-unit FCI scores for the treatment group?

RQ 1d: Are post-unit FCI scores for the treatment and control groups statistically significantly different?

RQ 2 Student Attitudes

RQ 2a: Are pre-unit TOSRA scores for the treatment and control groups statistically significantly different?

RQ 2b: Are there statistically significant differences between the pre- and post-unit TOSRA scores for the control group?

RQ 2c: Are there statistically significant differences between the pre- and post-unit TOSRA scores for the treatment group?

RQ 2d: Are post-unit TOSRA scores for the treatment and control groups statistically significantly different?

RQ 3 Associations

RQ 3: Are there associations between FCI scores and student attitudes towards Science as measured by the TOSRA scales?

To answer the first research question, regarding the academic outcomes of students, analysis of FCI data obtained from 199 Year 10 Science students was performed. This analysis included the determination and comparison of raw score change, normalized gain and effect size. Further analysis of individual question

response and dimension responses was performed to give finer detail on students' progress. The second research question, regarding students' attitudes towards Science, was addressed through the analysis of pre- and post-unit TOSRA responses. Finally, the third question was investigated through the comparison and correlation of TOSRA and FCI data.

4.3 Descriptive Characteristics of the Cohort

The initial cohort consisted of all Year 10 students (N = 228) within the study location. All students in the school complete a course of Core Science in line with the ACARA curriculum requirements, and this course is divided into four rotations: Physics, Chemistry, Biology, and Earth Science. This rotation allows all students to have access to specialist teachers, and, for the purposes of this study, provided consistency in the delivery of learning activities. The cohort was divided into classes of between 20 and 30 students. Due to school based timetable constraints, the groups were streamed based on mathematical progress. A stratified sampling technique was used (Creswell, 2012) to allocate groups to either control or treatment, based on streaming level. As the school assigned students to groups based on mathematical progress in a three-tier hierarchy, the stratification attempted to ensure an even spread of notional ability across the control and treatment groups. Groups 1, 3, 4, 5 and 8 were allocated to the control procedure, and Groups 2, 6, 7 and 9 were allocated to the treatment procedure.

The study was conducted during the Physics rotation for each group. Following the completion of the unit and associated instruments, the data collected was cleansed by the removal of the data of students who:

- had not provided consent for inclusion in the study.

- had not completed all aspects of the study (i.e. pre- and post-test of the FCI and TOSRA instruments, MOP analysis modules and corrective activities where appropriate).
- had demonstrated intentional patterning of the multiple-choice response formats (e.g. selecting all As or ABCD repeat responses).

Following this procedure, the control group consisted of *n*=104 (m= 46, f = 58) students and the treatment group consisted of *n*=95 (m= 45, f= 50) students, giving a total sample size *N* =199 (m= 91, f= 108).

Data were prepared for analysis in Excel and PhysPort by structuring the data by individual class group and by control or treatment allocation. Individual responses to questions and scores were maintained to allow detailed analysis of individual and group data.

4.4 Presentation of Results

The presentation of results shows the data and analysis associated with each research question. The FCI data can be analyzed in a variety of ways. The normalized change and effect size are both measures of the change in the scores of students between pre- and post-unit tests; both these metrics are reported for the treatment and control groups. A more detailed analysis of individual and cluster item response was performed to measure the stability of students' responses, with a particular focus on right-to-wrong transitions. This data was used to assess Research Questions 1a, 1b and 1c.

Scales I, E and A were selected from the TOSRA to determine students' attitudes towards Science. Pre- and post-unit TOSRA scores were determined for each scale and compared for any variation, and the statistical significance of any differences between pre- and post-unit scores were investigated using a T-test. This data was used to assess Research Questions 2a, 2b and 2c.

Research Question 3, regarding associations between FCI and TOSRA data, was investigated by utilizing a two tailed Pearson test to assess correlations between the FCI scores and TOSRA scales.

4.5 Descriptive Statistics

The descriptive statistics show a summary of the data set collected, grouped by the instrument used. Student attainment data is presented in the form of FCI results, and student attitude data is presented in the form of TOSRA scale results.

FCI summary results are shown in Table 4.1, including pre- and post-unit raw scores, normalized change and effect size. The distribution of pre- and post-unit FCI scores is illustrated in Figure 4.1, for both control and treatment groups.

Table 4.9 *Summary of Treatment and Control Pre- and Post-Unit FCI Results*

Group	Pre-Unit		Post-unit		Mean Normalized Change		Effect Size	Number of students		
	Mean score /30	*SD*	Mean score /30	*SD*	$< C_{ave} >$	*SD*	d	male	female	n
Control	7.6	2.4	9.8	3.4	0.08± 0.02	0.16	0.81	46	58	104
Treatment	8.2	2.9	12.1	4.0	0.19± 0.01	0.13	1.47	45	50	95

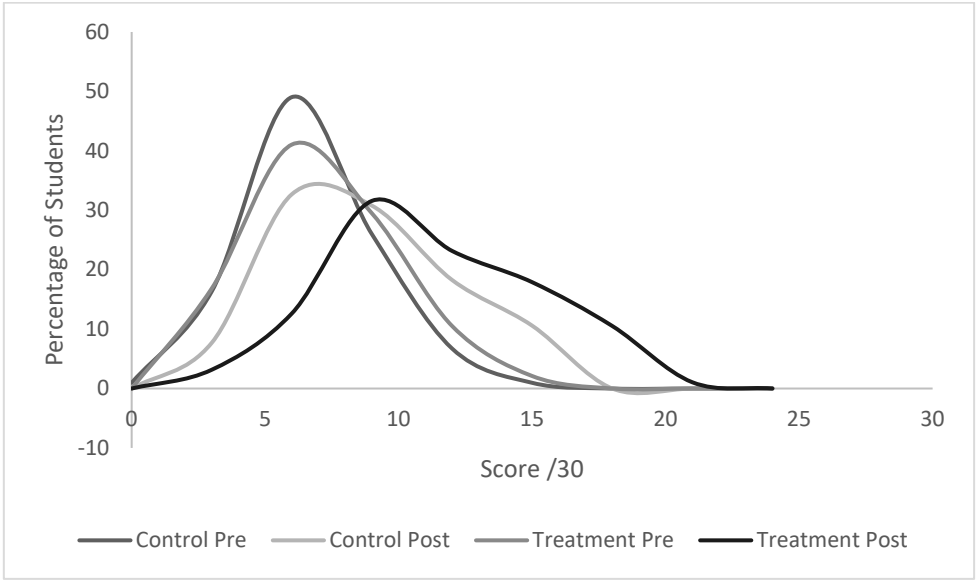


Figure 4.4 *Distribution of Scores in the Pre- and Post-Unit test for Control (n= 104) and Treatment (n=95) Groups*

TOSRA summary results are shown in Table 4.2, including pre- and post-unit mean scores and standard deviation. The distribution of pre- and post-unit TOSRA scores is illustrated in Figure 4.2 for the treatment group and Figure 4.3 for the control group.

Table 4.10 *Summary of Treatment and Control Group Pre- and Post-Unit TOSRA Scale Results*

Group	Scale I				Scale A				Scale E			
	Pre		Post		Pre		Post		Pre		Post	
	Score (/40)	*SD*	Score (/40)	*SD*	Score (/40)	*SD*	Score (/40)	*SD*	Score (/40)	*SD*	Score (/40)	*SD*
Control	27.9	3.4	28.5	4.2	28.7	3.6	29.4	3.8	28.0	3.7	28.7	4.1
Treatment	28.84	4.2	29.42	4.4	29.85	4.2	30.26	3.9	28.0	4.1	28.9	3.9

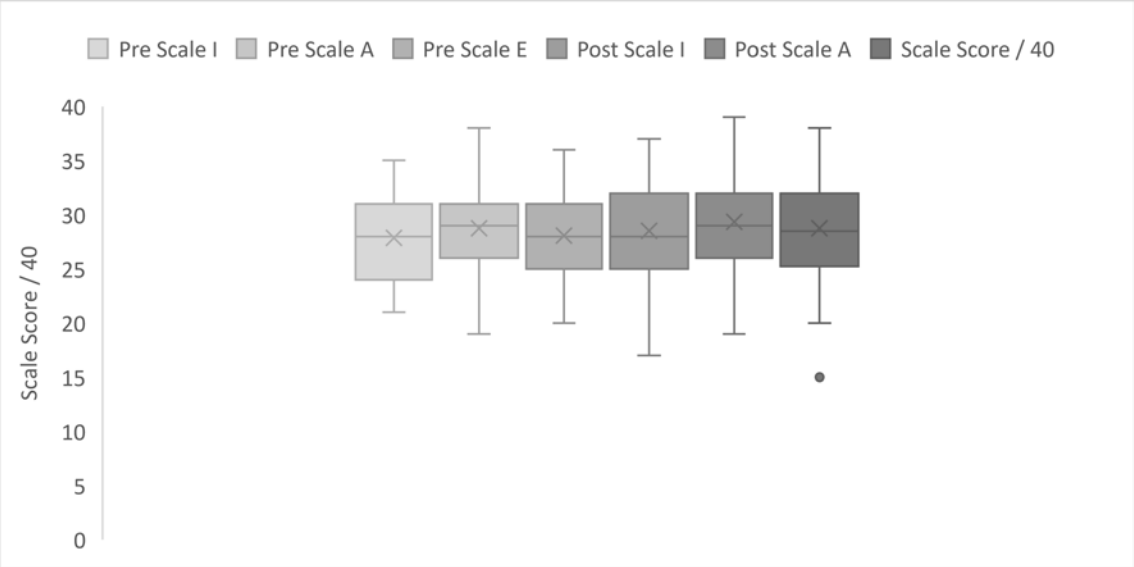


Figure 4.5 *Distribution of Scores in the Pre- and Post-Unit TOSRA Scales for Control (n= 104) Group.*

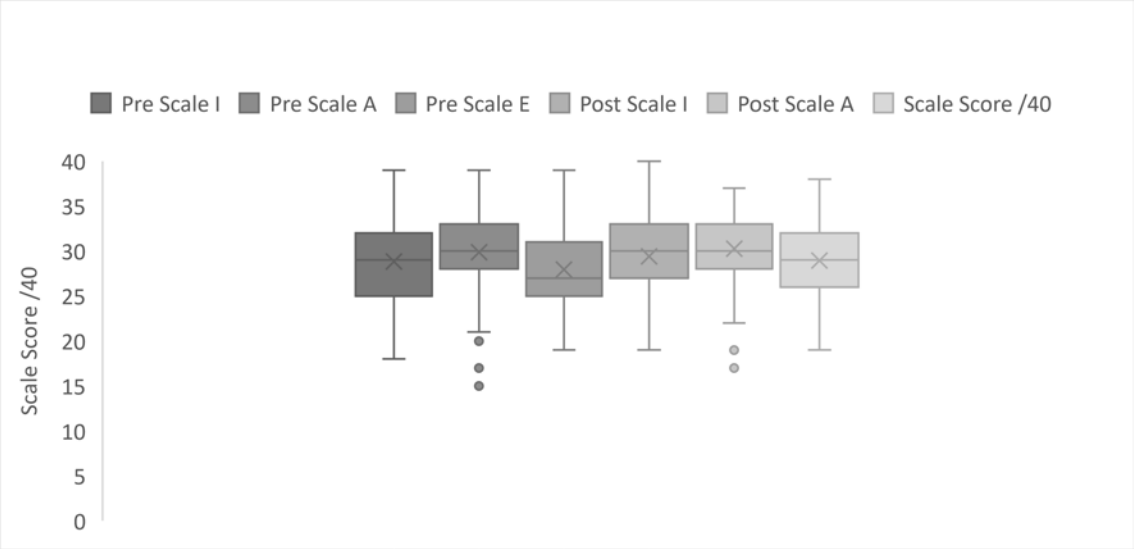


Figure 4.6 *Distribution of Scores in the Pre- and Post-Unit TOSRA Scales for Treatment (n= 96) Group.*

4.6 Inferential Statistics

4.6.1 Student Achievement

RQ 1a: Are pre-test FCI scores for the treatment and control groups statistically significantly different?

It was found that the difference in pre-unit FCI scores for treatment and control groups was not statistically significant. This data is presented in Table 4.3. The control group had an average pre-unit FCI score of 7.6/30 SD = 2.4 compared with the treatment group average of 8.2/30 SD = 2.9. A two-sample independent t test for unequal variance, with a 95% confidence interval, showed no significant difference in the control and treatment pre-unit FCI scores, with $t(197) = 1.66$ $p = .097$. This gave t below the critical value of 1.97 and $p > 0.05$, indicating that the mean difference of 1.4 was not significant.

Table 4.11 *Pre-Unit FCI Score Treatment and Control Group t-Test: Two-Sample Assuming Unequal Variance*

Characteristic	Control score	Treatment score
Mean	7.6	8.2
Variance	6.3	8.6
Sample size N	104	95
Pooled Variance	7.37	
Hypothesized Mean Difference	0	
Df	197	
t Stat	-1.66	
P(T<=t) two-tail	0.098	
t Critical two-tail	1.97	

This implies that there is no significant difference between the pre-unit FCI scores of the two groups, control and treatment. This data shows that the treatment and control groups have comparable pre-unit understanding of the Newtonian Concept, as measured by the FCI.

RQ 1b: Are there statistically significant differences between the pre- and post-test FCI scores for the control group?

It was found that there was a statistically significant difference in the mean pre- and post-unit FCI scores for the control group. This data is shown in Table 4.4. The pre-unit mean FCI score was 7.6/30, SD = 2.4, compared with a post-unit mean of 9.8/30, SD = 3.3. A paired two-sample t-test, with a 95% confidence interval, showed a significant difference in the pre- and post-unit mean scores, with t (103) = 8.47 and $p <$ 0.01. This gave t above the critical value of 1.98 and $p < 0.05$, indicating that the mean difference of 2.2 was significant. A Pearson's r value of 0.63 shows a moderate correlation (Muijs, 2013) between pre-unit and post-unit FCI raw score. This implies that students who achieved a higher grade on the pre-unit FCI also achieved a higher grade on the post-unit FCI.

Table 4.12 *Pre- and Post-Unit FCI Score Control Group t-Test: Paired Two-Sample for Means*

Characteristic	Pre-Unit	Post-Unit
Mean	7.6	9.8
Variance	6.27	11.41
Observations	104	104
Pearson Correlation	0.63	
Hypothesized Mean Difference	0	
Df	103	
t Stat	-8.47	
P(T<=t) two-tail	<0.01	
t Critical two-tail	1.98	

Variation in the pre- and post-unit FCI score can also be used to calculate normalized change and an effect size. In the control group this may be related to the effect of the learning activities in the unit. Table 4.5 shows a summary of the FCI data for the control group, including calculated values for the normalized change and effect size.

Table 4.13 *Summary FCI Data for Control Group*

Group	Mean Normalized Change	Standard Deviation in Normalized Change	Effect Size	Mean Pre-Unit score /30	Mean Post-Unit score /30	Number of students n
Control	0.08 ± 0.02	0.16	0.81	7.8	9.9	104

The normalized change from pre- to post-unit for the control group was 0.08 ± 0.02 ($SD = 0.16$). This data gives an effect size of 0.81, which is considered a large effect. This data shows a significant difference in pre- to post-unit academic achievement for the control group, as measured by the FCI.

RQ *1c: Are there statistically significant differences between the pre- and post-test FCI scores for the treatment group?*

It was found that there was a statistically significant difference in the mean pre- and post-unit FCI scores for the treatment group. This data is shown in Table 4.6. The

pre-unit mean FCI score was 8.2/30, *SD* = 2.9, compared with a post-unit mean of 12.1/30, *SD* = 3.9. A paired two-sample t- test, with a 95% confidence interval, showed a significant difference in the pre- and post-unit mean scores, with *t* (95) = 13.70 and p= < 0.01. This gave *t* above the critical value of 1.99 and $p < 0.05$, indicating that the mean difference of 3.9 was significant. A Pearson's *r* value of 0.71 shows a moderate correlation (Muijs, 2013) between pre-unit and post-unit FCI raw score. This implies that students who achieved a higher grade on the pre-unit FCI also achieved a higher grade on the post-unit FCI.

Table 4.14 *Pre- and Post-Unit FCI Score Treatment Group t-Test: Paired Two-Sample for Means*

Characteristic	Pre-Unit	Post-Unit
Mean	8.2	12.1
Variance	8.6	15.6
Observations	95	95
Pearson Correlation	0.71	
Hypothesized Mean Difference	0	
Df	94	
t Stat	-13.70	
P(T<=t) two-tail	<0.01	
t Critical two-tail	1.99	

Variation in the pre- and post-unit FCI score can also be used to calculate normalized change and an effect size. In the treatment group this may be related to the effect of the learning activities and the BLMP cycle used in the unit. Table 4.7 shows a summary of the FCI data for the treatment group, including calculated values for the normalized change and effect size.

Table 4.15 *Summary FCI Data for Treatment Group*

Group	Mean Normalized Change	Standard Deviation in Normalized Change	Effect Size	Mean Pre-Unit score /30	Mean Post-Unit score /30	Number of students N
Treatment	0.19 ± 0.01	0.13	1.47	8.1	12.3	95

The normalized change from pre- to post-unit for the treatment group was 0.19 ± 0.01 ($SD = 0.13$). The measured effect size was $d = 1.47$. This is considered a large effect.

This data shows a significant difference in pre- to post-unit academic achievement for the treatment group, as measured by the *FCI.*

1d: Are post-test FCI scores for the treatment and control groups statistically significantly different?

It was found that the difference in post-unit FCI scores for treatment and control groups was statistically significant. This data is shown in Table 4.8. The control group had an average post-unit FCI score of 9.8/30, $SD = 3.3$, compared with the treatment group average of 12.1/30, $SD = 3.9$. A two-sample independent t test for unequal variance, with a 95% confidence interval, showed significant difference in the control and treatment post unit FCI scores, with $t(186) = 4.48$ and $p < 0.001$. This gave t below the critical value of 1.97 and $p < 0.05$, indicating that the mean difference of 2.3 was significant.

Table 4.16 *Post-Unit FCI Score Treatment and Control Group t-Test: Two-Sample Assuming Unequal Variance*

Characteristic	Control	Treatment
Mean	9.8	12.1
Variance	11.4	15.7
Sample size N	104	95
Hypothesized Mean Difference	0	
Df	186	
t Stat	-4.48	
P(T<=t) two-tail	<0.01	
t Critical two-tail	1.97	

4.6.1.1 Individual Question Response

Whilst the use of normalized gain (Hake, 1998) and effect size (Hattie, 2009) is common in Physics education research (Miller et al., 2010) in the comparison of group data sets, such analyses can hide detail in the responses to individual questions and clusters of questions. Further, as the FCI is designed to assess a range of dimensions and has been shown to consist of a number of factors, these can be used to give more detail in the comparison of treatment and control groups. The percentage of correct responses per question in the FCI is shown in Table 4.9. Inspection of this data shows improvement in the percentage of correct responses for the majority of questions for both groups. Questions 3 and 15 are the only exception for the treatment group, and Questions 3, 12, 26 and 27 are exceptions for the control group. The question with the most substantial improvement in response was Question 1 for the control group with an increase of 32.7%. The treatment group showed the largest improvement for Questions 10, 19 and 28, with an increase of 31.6% correct responses. The treatment group showed the largest score change for 20 of the 30 questions. This data is illustrated in Table 4.9.

Table 4.17 *A Comparison of Correct Responses per Question in the FCI*

FCI Question Number	% Correct Response						Comparison
	Pre-Unit		Post-Unit		Post–Pre		Difference T–C
	Treatment	Control	Treatment	Control	Treatment	Control	
1	32.63	20.19	48.42	52.88	15.79	32.69	-16.90
2	21.05	20.19	28.42	32.69	7.37	12.50	-5.13
3	48.42	49.04	48.42	43.27	0.00	-5.77	5.77
4	54.74	38.46	57.89	43.27	3.16	4.81	-1.65
5	12.63	5.77	13.68	11.54	1.05	5.77	-4.72
6	52.63	44.23	62.11	53.85	9.47	9.62	-0.14

7	42.11	41.35	60.00	53.85	17.89	12.50	5.39
8	44.21	43.27	50.53	47.12	6.32	3.85	2.47
9	27.37	25.96	42.11	27.88	14.74	1.92	12.81
10	10.53	7.69	42.11	18.27	31.58	10.58	21.00
11	6.32	1.92	16.84	13.46	10.53	11.54	-1.01
12	67.37	72.12	74.74	68.27	7.37	-3.85	11.21
13	6.32	4.81	9.47	17.31	3.16	12.50	-9.34
14	26.32	15.38	48.42	26.92	22.11	11.54	10.57
15	15.79	5.77	12.63	25.00	-3.16	19.23	-22.39
16	18.95	21.15	40.00	26.92	21.05	5.77	15.28
17	10.53	4.81	20.00	13.46	9.47	8.65	0.82
18	8.42	11.54	15.79	17.31	7.37	5.77	1.60
19	28.42	24.04	60.00	36.54	31.58	12.50	19.08
20	26.32	38.46	56.84	46.15	30.53	7.69	22.83
21	41.05	24.04	46.32	29.81	5.26	5.77	-0.51
22	21.05	30.77	37.89	35.58	16.84	4.81	12.03
23	29.47	28.85	37.89	29.81	8.42	0.96	7.46
24	34.74	28.85	54.74	46.15	20.00	17.31	2.69
25	8.42	10.58	16.84	12.50	8.42	1.92	6.50
26	8.42	8.65	16.84	5.77	8.42	-2.88	11.31
27	35.79	41.35	46.32	41.35	10.53	0.00	10.53
28	11.58	7.69	43.16	18.27	31.58	10.58	21.00
29	63.16	68.27	82.11	69.23	18.95	0.96	17.99
30	4.21	9.62	18.95	10.58	14.74	0.96	13.78

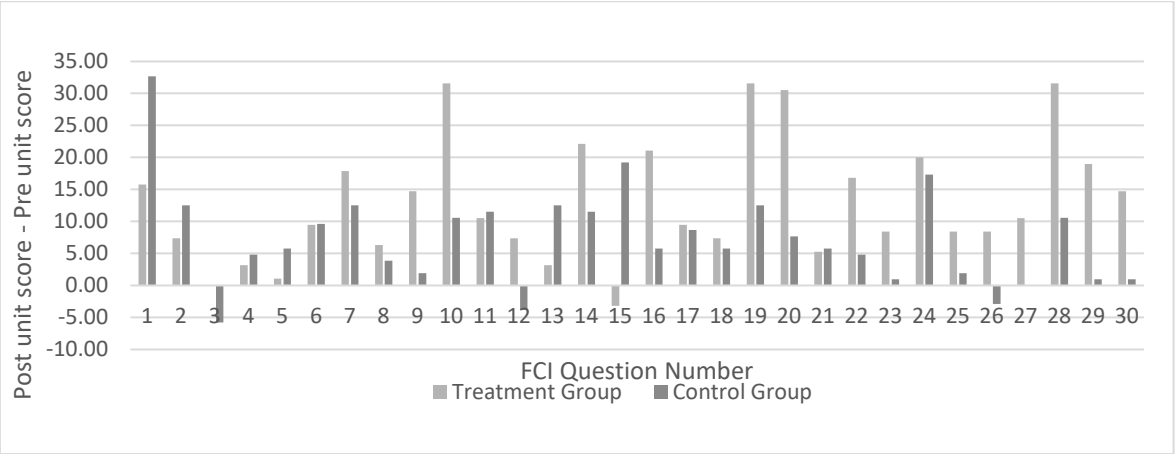


Figure 4.7 *Comparison of Percentage Change of Correct Responses per Question in FCI*

In the development of the FCI, the Newtonian concept was divided into five dimensions and each question was attributed to one of these dimensions (see Table 2.2) (Hestenes et al., 1995). Hence, it is possible to analyze the FCI results of the study based on these dimensions to determine if the intervention correlates with changes to specific dimensions of the Newtonian concept. Results grouped by dimension are shown in Table 4.10 as the percentage of correct responses in pre- and post-unit FCI assessment. The difference between pre- and post-unit was calculated and is illustrated in Figure 4.5.

Table 4.18 *A Comparison of Responses Grouped by FCI Dimension*

Dimension		Percentage Correct Control	Treatment
First Law	Pre-unit	23.61	26.55
	Post-unit	32.16	40.35
	Difference	8.55	13.80
Kinds of Forces	Pre-unit	25.00	26.40
	Post-unit	32.10	36.52
	Difference	7.10	10.12
Kinematics	Pre-unit	33.10	33.98
	Post-unit	38.60	52.78
	Difference	5.49	18.80
Second Law	Pre-unit	26.73	28.42
	Post-unit	29.23	39.37
	Difference	2.50	10.95
Superposition Principle	Pre-unit	17.31	19.37
	Post-unit	22.88	29.89
	Difference	5.58	10.53
Third Law	Pre-unit	20.43	24.21
	Post-unit	31.01	40.26
	Difference	10.58	16.05

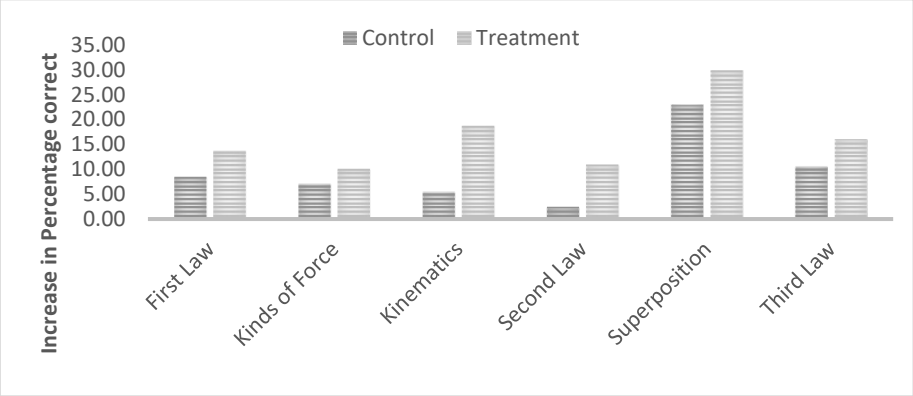


Figure 4.8 *Change in Percentage Correct Grouped by FCI Dimension*

Inspection of the dimension grouped data shows an increase in correct responses across all dimensions for both treatment and control groups, with the treatment group having a larger increase in all dimensions. The most substantial increase for both groups is in the superposition principle dimension; all questions in this dimension are shown as probing a range of dimensions (Hestenes et al. 1992a), and hence it may be that improvement in this dimension is related to improvement in a range of other dimensions. The largest difference in improvement is in the Kinematics dimension,

where the treatment group improved from 34% to 53%, compared with an increase from 33% to 39% for the control group.

In general, the analysis of individual question and dimension results shows that there is a positive correlation between the treatment group and improved response in the post-unit FCI for individual questions and dimensions.

4.6.1.2 Variation of Student Responses between Pre- and Post-Unit FCI Test

On analysis of pre-unit and post-unit student responses, it is apparent that some students responded correctly in the pre-unit test and incorrectly in the post-unit test. This R–W transition may indicate a lack of stability in the student's application or adoption of the Newtonian concept. Table 4.11 shows a comparison of the stability of students' correct pre-unit responses for the control and treatment group. Also included is a comparison of students' adoption of correct responses in the post-unit test. When comparing group data for students who changed from a correct to an incorrect response, a negative value for the difference indicates that responses were more stable in the treatment group compared with the control group. The treatment group was more stable for 27 of the 30 questions: one question (17) showed no difference between the two groups, and two questions (1 and 15) showed a higher stability for the control group. This data is illustrated in Figure 4.6, which shows the percentage of students who gave the correct response in the pre-unit test who gave an incorrect response in the post-unit test, a R–W transition, for both treatment and control groups.

When comparing students who changed from an incorrect to correct response, a positive value for the difference indicates that more students in the treatment group moved from incorrect to correct. This is the case for 22 of the 30 questions. For the remainder of the questions (1, 2, 3, 5, 11, 13, 15 and 18) the control group made more improvement than the treatment group. These results align with the findings discussed in Section 4.6.1.1.

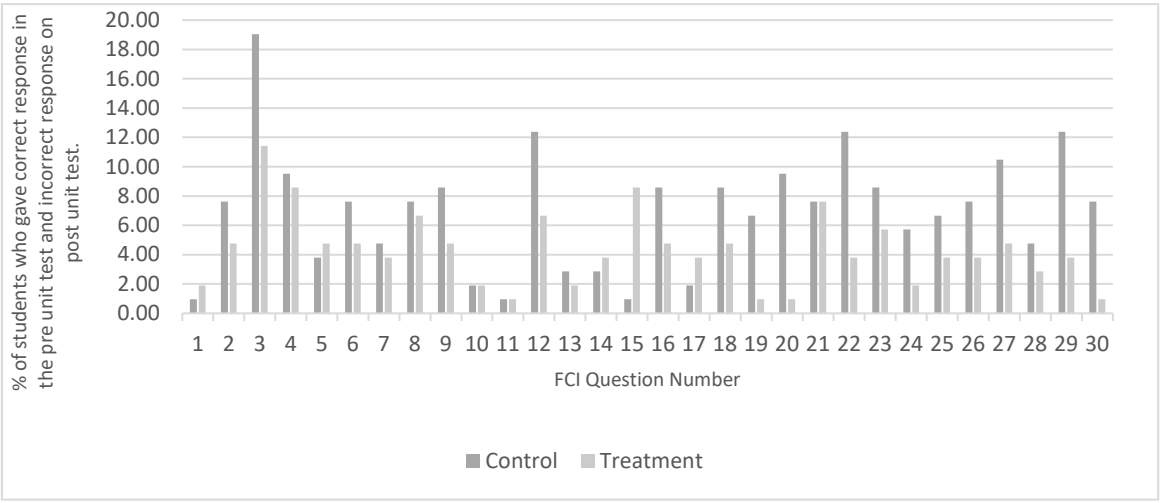


Figure 4.9 *A Comparison of the Stability of Correct Responses in the Pre-Unit Test*

Table 4.19 *A Comparison of Change in FCI Response Pre- and Post-Unit*

FCI Question Number	% Students who gave correct response in the pre-test and incorrect response in the post-test (R–W transition)			% Students who gave incorrect response in the pre-test and correct response in the post-test (W–R transition)		
	Control	Treatment	Difference (T-C)	Control	Treatment	Difference (T-C)
1	4.76	6.45	1.69	41.67	25.00	-16.67
2	38.10	25.00	-13.10	25.00	16.00	-9.00
3	39.22	26.09	-13.13	25.93	24.49	-1.44
4	25.00	17.31	-7.69	23.08	27.91	4.83
5	66.67	41.67	-25.00	10.10	7.23	-2.87
6	17.39	10.00	-7.39	30.51	31.11	0.60
7	11.63	10.00	-1.63	29.03	36.36	7.33
8	17.78	16.67	-1.11	20.00	24.53	4.53
9	33.33	19.23	-14.10	14.10	27.54	13.43
10	25.00	20.00	-5.00	13.40	36.47	23.07
11	50.00	16.67	-33.33	12.62	12.36	-0.26
12	17.33	10.94	-6.40	30.00	45.16	15.16
13	60.00	33.33	-26.67	16.00	5.62	-10.38
14	18.75	16.00	-2.75	16.85	35.71	18.86
15	16.67	60.00	43.33	21.21	7.50	-13.71
16	40.91	27.78	-13.13	18.07	31.17	13.10
17	40.00	40.00	0.00	11.00	15.29	4.29
18	75.00	62.50	-12.50	16.13	13.79	-2.34
19	28.00	3.70	-24.30	25.00	45.59	20.59
20	25.00	4.00	-21.00	27.69	41.43	13.74
21	32.00	20.51	-11.49	17.50	23.21	5.71
22	40.63	20.00	-20.63	24.66	26.67	2.01

23	30.00	21.43	-8.57	13.33	20.90	7.56
24	20.00	6.06	-13.94	32.00	33.87	1.87
25	63.64	50.00	-13.64	9.57	13.79	4.22
26	88.89	50.00	-38.89	5.21	12.64	7.44
27	25.58	14.71	-10.88	17.74	24.59	6.85
28	62.50	27.27	-35.23	16.49	38.10	21.60
29	18.31	6.67	-11.64	41.18	60.00	18.82
30	80.00	25.00	-55.00	9.47	16.48	7.01

An inspection of the data in Table 4.11 shows that a high percentage (up to 89%) of students who answered correctly in the pre-unit test answered incorrectly in the post-unit test across a range of questions for both treatment and control groups. Whilst the use of percentages allows a comparison between the two groups, it is important to review the actual number of students to whom the percentages are referring, because, if a small number of students obtained the correct response in the pre-unit test, a large percentage swing could be obtained with a small number of students responding incorrectly in the post-unit test. For example, in the control group Question 26 shows 89% of students who answered correctly in the pre-test answered incorrectly in the post-unit test—however, as the initial correct response was only achieved by nine students, this means that eight students demonstrated instability in their concept of Newton's Second Law. Similarly, a large percentage (50%) of students in the treatment group who responded correctly to Question 26 responded incorrectly in the post-unit test—however, this relates to four students demonstrating instability in their understanding.

In summary, these results show a positive correlation between the treatment and both the stability of correct responses between pre- and post-unit FCI assessment and the move from incorrect to correct response. However, these results also suggest some issues with the use of the FCI, as some students gave correct responses in the pre-unit test but subsequently answered the same question incorrectly in the post-unit test. This suggests that scores may have been elevated by guessing or the application of incomplete concept models to achieve correct responses with incorrect reasoning.

4.6.2 Student Attitudes

RQ 2a: Are pre-unit TOSRA scores for the treatment and control groups statistically significantly different?

It was found that the difference in pre-unit TOSRA scores for the treatment and control groups was not statistically significant for the any of the scales used: I, A and E. The control group had average pre-unit TOSRA scale scores of I = 27.86, *SD* = 3.42, A = 28.76, *SD* = 3.62, E = 28.09, *SD* = 3.74 (10 items per scale, each item scored out of 4). This compared with the treatment group average scale scores of I = 28.84, *SD* = 4.40, A = 29.85 *SD* = 4.21, E = 27.94, *SD* = 4.10 (10 items per scale, each item scored out of 4).

For TOSRA Scale I, a two-sample independent *t* test for unequal variance, with a 95% confidence interval, showed no significant difference in the control and treatment pre-unit scale scores, with $t(178) = 1.73$, $p = .08$. This gave *t* below the critical value of 1.97 and $p > 0.05$, indicating that the mean difference of 0.95 was not significant. These results are shown in Table 4.12.

Table 4.20 *Scale I Pre-Unit Control and Treatment Group t-Test: Two-Sample Assuming Unequal Variance*

Characteristic	Control	Treatment
Mean	27.86	28.84
Variance	11.83	19.34
Observations	104	95
Hypothesized Mean Difference	0	
Df	178	
t Stat	-1.73	
P(T<=t) two-tail	0.08	
t Critical two-tail	1.97	

For TOSRA Scale A, a two-sample independent *t* test for unequal variance, with a 95% confidence interval, showed no significant difference in the control and treatment pre-unit scale scores, with $t(186) = 1.93$, $p = .06$. This gave *t* below the critical value of 1.97 and $p > 0.05$, indicating that the mean difference of 1.08 was not significant. These results are shown in Table 4.13.

Table 4.21 *Scale A Pre-Unit Control and Treatment Group t-Test: Two-Sample Assuming Unequal Variance*

Characteristics	Control Group	Treatment Group
Mean	28.77	29.85
Variance	13.21	17.91
Observations	104	95
Hypothesized Mean Difference	0	
Df	186	
t Stat	-1.93	
P(T<=t) two-tail	0.06	
t Critical two-tail	1.97	

For TOSRA Scale E, a two-sample independent *t* test for unequal variance, with a 95% confidence interval, showed no significant difference in the control and treatment pre-unit scale scores, with $t(191) = 0.25$, $p = .80$. This gave *t* below the critical value of 1.97 and $p > 0.05$, indicating that the mean difference of 0.07 was not significant. These results are shown in Table 4.14.

Table 4.22 *Scale E Pre-Unit Control and Treatment Group t-Test: Two-Sample Assuming Unequal Variance*

Characteristics	Control	Treatment
Mean	28.09	27.95
Variance	14.10	17.01
Observations	104	95
Hypothesized Mean Difference	0	
Df	191	
t Stat	0.25	
P(T<=t) two-tail	0.80	
t Critical two-tail	1.97	

This implies that there is no significant difference between the pre-unit TOSRA Scales I, A and E score of the two groups, control and treatment. This data shows that the treatment and control groups had comparable pre-unit attitudes towards Scientific Inquiry, Adoption of Scientific Attitudes, and Enjoyment of Science Lessons.

RQ 2b: Are there statistically significant differences between the pre- and post-unit TOSRA scores for the control group?

It was found that there was a statistically significant difference in the pre- and post-unit scores for TOSRA Scales I, A and E for the control group. The control group had average pre-unit TOSRA scale scores of I = 27.86, *SD* = 3.42, A = 28.76, *SD* = 3.62, E = 28.09, *SD* = 3.74 (10 items per scale, each item scored out of 4). This compared with the post unit average scale scores of I = 28.52, *SD* = 4.16, A = 29.37, *SD* = 3.81, E = 28.75, *SD* = 4.07 (10 items per scale, each item scored out of 4).

For TOSRA Scale I, a paired two-sample *t* test for means, with a 95% confidence interval, showed a significant difference in the control pre- and post-unit scale scores, with $t(103) = 2.08$, $p = .04$. This gave *t* above the critical value of 1.98 and $p < 0.05$, indicating that the mean difference of 0.66 was significant. This difference gives an effect size $d = 0.16$. This is considered a small effect. The correlation between pre- and post-unit scores was measured using Pearson's *r*. A value of 0.71 indicates a strong relationship, and this implies that students who scored higher on the pre-unit scale scored higher on the post-unit scale. These results are shown in Table 4.15.

Table 4.23 *Control Group Pre- and Post-Unit TOSRA Scale I t-Test: Paired Two-Sample for Means*

Characteristic	Pre-Unit Scale I	Post-Unit Scale I
Mean	27.87	28.47
Variance	11.83	17.40
Observations	104	104
Pearson Correlation	0.71	
Hypothesized Mean Difference	0	
Df	103	
t Stat	-2.08	
P(T<=t) two-tail	0.040	
t Critical two-tail	1.98	
Effect size d	0.16	

For TOSRA Scale A, a paired two-sample *t* test for means, with a 95% confidence interval, showed no significant difference in the control pre- and post-unit scale scores, with $t(103) = 2.49$ and $p = .01$. This gave *t* above the critical value of 1.98 and $p < 0.05$, indicating that the mean difference of 0.61 was significant. This difference gives an effect size $d = 0.16$. This is considered a small effect. The correlation between pre- and post-unit scores was measured using Pearson's *r*. A value of 0.71 indicates a strong relationship, and this implies that students who scored higher on the pre-unit scale scored higher on the post-unit scale. These results are shown in Table 4.16.

Table 4.24 *Control Group Pre- and Post-unit TOSRA Scale A t-Test: Paired Two-Sample for Means*

Characteristic	Pre-Unit Scale A	Post-Unit Scale A
Mean	28.77	29.37
Variance	13.21	14.68
Observations	104	104
Pearson Correlation	0.79	
Hypothesized Mean Difference	0	
Df	103	
t Stat	-2.49	
P(T<=t) two-tail	0.01	
t Critical two-tail	1.98	
Effect size d	0.16	

For TOSRA Scale E, a paired two-sample *t* test for means, with a 95% confidence interval, showed significant difference in the control pre- and post-unit scale scores, with $t(103) = 2.82$ and $p = < 0.01$. This gave *t* above the critical value of 1.98 and $p < 0.05$, indicating that the mean difference of 0.67 was significant. This difference gives an effect size of $d = 0.16$. This is considered a small effect size. The correlation between pre- and post-unit scores was measured using Pearson's *r*. A value of 0.79 indicates a strong relationship, and this implies that students who scored higher on the pre-unit scale scored higher on the post-unit scale. These results are shown in Table 4.17.

Table 4.25 *Control Group Pre- and Post-Unit TOSRA Scale E t-Test: Paired Two-Sample for Means*

Characteristic	Pre-Unit Scale E	Post-Unit Scale E
Mean	28.09	28.73
Variance	14.10	16.82
Observations	104	104
Pearson Correlation	0.83	
Hypothesized Mean Difference	0	
Df	103	
t Stat	-2.82	
P(T<=t) two-tail	0.01	
t Critical two-tail	1.98	
Effect size d	0.16	

This implies that there is a significant difference between the pre- and post-unit TOSRA Scales I, A and E score for the control group. This data shows that the control group had a more positive attitude towards Scientific Inquiry, Adoption of Scientific Attitudes, and Enjoyment of Science Lessons in the post-unit response, albeit with a small effect size of $d = 0.16$ in each scale, which indicates the pre- and post-test means differ by 0.16 *SD*. The data also shows a strong correlation between pre- and post-unit scores for all scales.

RQ 2c: Are there statistically significant differences between the pre- and post-unit TOSRA scores for the treatment group?

It was found that there was not a statistically significant difference in the mean pre- and post-unit TOSRA Scale A scores for the treatment group. However, there was a statistically significant difference in the pre- and post-unit scores for TOSRA Scales I and E for the treatment group. The treatment group had average pre-unit scale scores of I = 28.84, *SD* = 4.40, A = 29.85 *SD* = 4.21, E = 27.94, *SD* = 4.10 (10 items per scale,

each item scored out of 4). This compared with the post unit average scale scores of: I = 29.44, *SD* = 4.36, A = 30.26 *SD* = 3.93, E = 28.94, *SD* = 3.87 (10 items per scale, each item scored out of 4).

For TOSRA Scale I, a paired two-sample *t* test for means, with a 95% confidence interval, showed a significant difference in the control pre- and post-unit scale scores for the treatment group, with $t(94) = 3.02$ and $p < 0.01$. This gave *t* above the critical value of 1.99 and $p < 0.05$, indicating that the mean difference of 0.58 was significant. This difference gave an effect size $d = 0.13$, which is considered a small effect. The correlation between pre- and post-unit scores was measured using Pearson's *r*. A value of 0.91 indicates a very strong relationship, and this implies that students who scored higher on the pre-unit scale scored higher on the post-unit scale. These results are shown in Table 4.18.

Table 4.26 *Treatment Group Pre- and Post-Unit TOSRA Scale I t-Test: Paired Two-Sample for Means*

Characteristic	Pre-Unit Scale I	Post-Unit Scale I
Mean	28.84	29.42
Variance	19.35	19.57
Observations	95	95
Pearson Correlation	0.91	
Hypothesized Mean Difference	0	
Df	94	
t Stat	-3.02	
P(T<=t) two-tail	<0.01	
t Critical two-tail	1.99	
Effect size d	0.13	

For TOSRA Scale A, a paired two-sample *t* test for means, with a 95% confidence interval, showed no significant difference in the control pre- and post-unit

scale scores for the treatment group, with $t(94) = 1.67$, $p = .10$. This gave t below the critical value of 1.99 and $p > 0.05$, indicating that the mean difference of 0.41 was not significant. The correlation between pre- and post-unit scores was measured using Pearson's r. A value of 0.83 indicates a very strong relationship, and this implies that students who scored higher on the pre-unit scale scored higher on the post-unit scale. These results are shown in Table 4.19.

Table 4.27 *Treatment Group Pre- and Post-Unit TOSRA Scale A t-Test: Paired Two-Sample for Means*

Characteristic	Pre-Unit Scale A	Post-Unit Scale A
Mean	29.85	30.26
Variance	17.91	15.62
Observations	95	95
Pearson Correlation	0.83	
Hypothesized Mean Difference	0	
Df	94	
t Stat	-1.67	
P(T<=t) two-tail	0.10	
t Critical two-tail	1.99	

For TOSRA Scale E, a paired two-sample t test for means, with a 95% confidence interval, showed significant difference in the treatment pre- and post-unit scale scores, with $t(94) = 5.5$ and $p = <0.01$. This gave t above the critical value of 1.98 and $p < 0.05$, indicating that the mean difference of 0.99 was significant. This difference gives an effect size of $d = 0.23$, which indicates the pre- and post-test means differ by 0.23 *SD*. This is considered a small effect size. The correlation between pre- and post-unit scores was measured using Pearson's r. A value of 0.91 indicates a very strong relationship, and this implies that students who scored higher on the pre-unit scale scored higher on the post-unit scale.

Table 4.28 *Treatment Group Pre- and Post-Unit TOSRA Scale E t-Test: Paired Two-Sample for Means*

Characteristic	Pre-Unit Scale E	Post-Unit Scale E
Mean	27.947	28.937
Variance	17.008	15.124
Observations	95	95
Pearson Correlation	0.91	
Hypothesized Mean Difference	0	
Df	94	
t Stat	-5.500	
P(T<=t) two-tail	<0.01	
t Critical two-tail	1.99	
Effect size d	0.23	

This implies that there is no significant difference between the pre- and post-unit TOSRA Scale A score treatment group. However, there is a significant difference in the pre- and post-unit TOSRA Scales I and E scores of the treatment group. This data shows that the treatment group has comparable pre- and post-unit attitudes towards Adoption of Scientific Attitudes, but that the treatment group had a more positive attitude towards the Scientific Inquiry and Enjoyment of Science Lessons in the post-unit response, albeit with small effect sizes of d = 0.13 and 0.23 respectively.

RQ 2d: Are post-unit TOSRA scores for the treatment and control groups statistically significantly different?

It was found that the difference in post-unit TOSRA scores for treatment and control groups was not statistically significant for the Scales I, A and E. The control group had average post-unit TOSRA scale scores of I = 28.52, *SD* = 4.16, A = 29.37, *SD* = 3.81, E = 28.75, *SD* = 4.07 (10 items per scale, each item scored out of 4). This compared with the treatment group post-unit average scale scores of I = 29.44, *SD* =

4.36, A = 30.26 *SD* = 3.93, E = 28.94, *SD* = 3.87 (10 items per scale, each item scored out of 4).

For TOSRA Scale I, a two-sample independent *t* test for unequal variance, with a 95% confidence interval, showed no significant difference in the control and treatment post-unit scale scores, with $t(193) = 1.46$ and $p = .15$. This gave *t* below the critical value of 1.97 and $p > 0.05$, indicating that the mean difference of 0.89 was not significant. This difference has reduced from the pre-unit value of 0.95. These results are shown in Table 4.21.

Table 4.29 *Control and Treatment Group Post-Unit Scale I t-Test: Two-Sample Assuming Unequal Variance*

Characteristics	Control	Treatment
Mean	28.53	29.42
Variance	17.46	19.57
Observations	104	95
Hypothesized Mean Difference	0	
Df	193	
t Stat	-1.46	
P(T<=t) two-tail	0.15	
t Critical two-tail	1.97	

For TOSRA Scale A, a two-sample independent *t* test for unequal variance, with a 95% confidence interval, showed no significant difference in the control and treatment post-unit scale scores, with $t(194) = 1.62$ and $p = .11$. This gave *t* below the critical value of 1.97 and $p > 0.05$, indicating that the mean difference of 0.90 was not significant. This difference is reduced from the pre-unit value of 1.08. These results are shown in Table 4.22.

Table 4.30 *Control and Treatment Group Post-Unit Scale A t-Test: Two-Sample Assuming Unequal Variance*

Characteristics	Control	Treatment
Mean	29.37	30.26
Variance	14.68	15.62
Observations	104	95
Hypothesized Mean Difference	0	
Df	194	
t Stat	-1.62	
P(T<=t) two-tail	0.11	
t Critical two-tail	1.97	

For TOSRA Scale E, a two-sample independent *t* test for unequal variance, with a 95% confidence interval, showed no significant difference in the control and treatment post-unit scale scores, with $t(197) = 0.36$ and $p = .71$. This gave *t* below the critical value of 1.97 and $p > 0.05$, indicating that the mean difference of 0.21 was not significant. This difference is increased from the pre-unit value of 0.07. These results are shown in Table 4.23.

Table 4.31 *Control and Treatment Group Post-Unit Scale E t-Test: Two-Sample Assuming Unequal Variance*

Characteristics	Control	Treatment
Mean	28.73	28.94
Variance	16.82	15.12
Observations	104	95
Hypothesized Mean Difference	0	
Df	197	
t Stat	-0.36	
P(T<=t) two-tail	0.72	
t Critical two-tail	1.97	

This implies that there is no significant difference between the post-unit TOSRA Scales I, A and E score of the two groups, control and treatment. The difference in control and treatment group Scales I and A scores was reduced compared with the pre-unit values, whilst the Scale E difference was increased. This data shows that both the treatment and control groups have comparable post-unit attitudes towards Scientific Inquiry, Adoption of Scientific Attitudes and Enjoyment of Science, as measured by the TOSRA.

4.6.2.1 TOSRA Internal Consistency

The reliability and internal consistency of the TOSRA has been reported by a number of studies (Ali et al., 2013; Fraser, 1998). In the original research, Fraser (1982) reported Cronbach Alpha values ranging between 0.66 and 0.93. For the scales and year level (10) used in this study, Fraser reported Scale I $\alpha = 0.86$, Scale A $\alpha = 0.67$, and Scale E $\alpha = 0.93$.

Cronbach Alpha values were calculated for each scale, for both the control group and treatment group's pre- and post-unit response, using individual student responses as the unit of analysis. Scale alpha scores are shown in Table 4.24.

Table 4.32 *Internal Consistency of TOSRA Scales for Control and Treatment Groups.*

		Cronbach Alpha Score	
	Scale	Control	Treatment
I	Pre	0.65	0.79
	Post	0.76	0.79
A	Pre	0.70	0.77
	Post	0.70	0.76
E	Pre	0.70	0.75
	Post	0.72	0.76

It can be seen that most scales demonstrate acceptable internal consistency, with only the pre-unit control group Scale I scale showing an alpha score below 0.7. This

improved to 0.76 in the post-unit TOSRA. These results show that the TOSRA form used in this study, a four-point Likert scale, maintains acceptable internal consistency.

4.6.3 Associations between FCI and TOSRA Data

Correlations between the FCI and TOSRA data were investigated to determine a response to the third research question:

RQ 3: Are there associations between FCI scores and student attitudes towards Science as measured by the TOSRA scales?

It is apparent that there may be an interplay between a student's attitude towards science and their achievement in the subject. To determine the presence, or lack thereof, of this association, a two-tailed significance Pearson correlation was performed with the TOSRA scale scores and the FCI score. A weak but significant correlation was found between all pre-unit attitudinal scales and students' pre-unit FCI scores for both control and treatment groups: control group Scale I $r = 0.23$ $p = 0.008$, Scale E $r = 0.19$ $p = 0.001$, Scale A $r = 0.18$ $p = 0.001$; treatment group Scale I $r = 0.19$ $p = 0.001$, Scale E $r = 0.20$ $p = 0.001$, Scale A $r = 0.25$ $p = 0.001$, with r values between 0.20 and 0.35 indicating a slight correlation (L. Cohen et al., 2007), whilst p values < 0.05 indicate a statistical significance in the correlation (Muijs, 2013). This implies that there is a slight correlation between students who scored higher on the attitudinal scales and those who scored higher on the pre-unit FCI. These results are shown in Table 4.25.

Table 4.33 *Correlation between Pre-Unit TOSRA Scale and Pre-Unit FCI Score*

TOSRA Scale	Control group		Treatment group	
	Pearson's *r*	*p*	Pearson's *r*	*p*
I	0.23	0.008	0.19	0.001
A	0.19	0.001	0.20	0.001
E	0.18	0.001	0.25	0.001

An analysis of post-unit FCI correlation with pre-unit TOSRA scores shows a weak but significant correlation for both the control and treatment groups for all scales except the treatment group E scale (control Scale I $r = 0.26$ $p = <0.001$, Scale A $r = 0.23$

p = < 0.001, Scale E r = 0.22 p = <0.001. Treatment Scale I r = 0.18 p = <0.001, Scale A r = 0.19 p = <0.001). This implies students who had more positive pre-unit attitudes towards science tended to have higher post-unit FCI scores. The strongest correlation was found with the treatment groups pre-unit Scale E and post-unit FCI score (r = 0.43 p = < 0.001), with r values between 0.35 and 0.5 indicating a moderate correlation. Students in this group who scored higher on the Enjoyment of Science scale scored higher in the post-unit FCI. A summary of these values is provided in Table 4.26.

Table 4.34 *Correlation between Pre-Unit TOSRA Scale and Post-Unit FCI Score*

TOSRA Scale	Control group		Treatment group	
	Pearson's r	p	Pearson's r	p
I	0.26	<0.001	0.18	<0.001
A	0.23	0.001	0.19	<0.001
E	0.22	<0.001	0.43	<0.001

A weak but significant correlation was found between all post-unit attitudinal scales and students' post-unit FCI scores for both the control and treatment groups, except treatment group Scale A (Control Scale I r = 0.28 p = 0.008, Scale A r = 0.23 p = 0.001, Scale E r = 0.23 p = 0.001. Treatment Scale I r = 0.14 p = 0.008, Scale A r = 0.16 p = 0.001). Pearson's r values < 0.3 indicate a weak correlation (L. Cohen et al., 2007) whilst p values <0.05 indicate a statistically significant correlation. This implies students who had more positive attitudes towards science post-unit tended to have higher post-unit FCI scores. The strongest correlation was found with the treatment group's post-unit Scale E and post-unit FCI score (r = 0.42 p = <0.001), with r values between 0.35 and 0.5 indicating a moderate correlation. Students in this group who scored higher on the Enjoyment of Science scale scored higher in the post-unit FCI. A summary of these values is provided in Table 4.27.

Table 4.35 *Correlation between Post-Unit TOSRA Scale and Post-Unit FCI Score*

TOSRA Scale	Control group		Treatment group	
	Pearson's *r*	*p*	Pearson's *r*	*p*
I	0.28	0.008	0.14	0.008
A	0.23	0.001	0.16	0.001
E	0.23	0.001	0.42	<0.001

4.7 Chapter Review

This chapter presented data with the aim of developing responses to the research questions of the study.

An initial analysis of the data was performed to ensure there was no significant difference in the pre-unit responses to both the FCI and TOSRA. Results showed comparability between the groups, with no significant difference between the treatment (27%) and control (26%) groups.

Results were discussed in regard to the first research question and the pre- and post-unit FCI responses. It was shown that there was a significant difference between the increase in raw score, normalized gain and effect size between the control and treatment groups. A significant difference in normalized gain between control and treatment groups of 0.08 was reported; this gives a t value of -5.5 with $p = 0.000$, which indicates the difference in the means is significant (sig level $p < 0.05$). These results are summarized in Table 4.28.

Table 4.36 *Summary of FCI Score Data*

FCI Measure	Control Group	Treatment Group
Change in average score /30	+ 7	+14
Normalized Change	0.08	0.19
Effect Size	0.81	1.47

Further analysis of individual question responses was performed, and this determined that the treatment group made greater improvement than the control group on 20 out of the 30 FCI questions. When results were grouped by dimension, both

groups showed improvement in all dimensions, with the treatment group showing the greatest improvement in all dimensions.

A comparison of pre-unit and post-unit responses revealed that a significant number of students made correct responses in the pre-unit test but changed to incorrect responses in the post-unit test. It was shown that this instability of response was most evident in the control group and it was proposed that this may relate to incomplete development of the Force concept, leading to variability in response.

The second research question was addressed by analysis of the TOSRA data. A comparison of the control and treatment group pre-unit responses showed no significant difference in the scores for all scales. These results are shown in Table 4.29. It was shown that there was a small but significant change for both the control and treatment groups across TOSRA Scales I and E when pre- and post-unit results were compared, but that the differences in Scale A scores was only significant for the control group. The differences between the post-unit scores for the control and treatment groups were not significant for any of the scales used. The correlation between pre- and post-unit scales was strong for the control group (r = 0.71 to 0.79) and very strong for the treatment group (r = 0.83 to 0.91). This indicates that students who had a high score in the pre-unit scales maintained these high scores in the post-unit scales.

Table 4.37 *Summary of TOSRA Data*

Scale	Control group score /40			Treatment group score /40		
	Pre-unit	Post-unit	Effect d	Pre-unit	Post-unit	Effect d
I	27.86	28.52	0.16	28.84	29.44	0.13
A	28.27	29.37	0.16	29.85	30.26	Not sig.
E	28.09	28.75	0.16	27.94	28.94	0.23

The third research question was addressed by a comparison of the TOSRA and FCI data and an investigation of correlation between these results. Pearson's *r* values show weak but significant correlations for both groups between all attitudinal scales and

pre-unit FCI scores, weak but significant correlations for all control group pre- and post-unit scales and post-unit FCI score, weak but significant correlations for the treatment group's pre- and post-unit Scales I and A and post-unit FCI score, moderate but significant correlations between pre- and post-unit Scale E scores and post-unit FCI score. The Pearson *r* values for each of these correlations are shown in Table 4.30.

Table 4.38 *Correlation of TOSRA and FCI Scores*

TOSRA Scale		Control Group		Treatment Group	
		Pre-	Post-	Pre-	Post-
I	Pre	0.23	0.26	0.09	0.18
	Post		0.28		0.14
A	Pre	0.19	0.23	0.20	0.19
	Post		0.23		0.16
E	Pre	0.18	0.22	0.25	0.43
	Post		0.22		0.43

Chapter 5: Discussion and Conclusion

5.1 Chapter Overview

This chapter will use the data analysis results presented in Chapter 4 to address the individual research questions of the study. This data will be compared and contrasted with the research literature discussed in Chapter 2 to place the findings in the context of other studies. A conclusion will then be drawn regarding the overarching question of the study. This will be followed by a discussion of the implications of this study's findings for the Science classroom and in curriculum development. Finally, the limitations of the study will be discussed and areas for further research highlighted.

5.2 Summary of Study

The overarching question for this study was '*Does a combination of the application of a Mastery for Learning (ML) approach with Blended Learning (BL) activities—specifically the use of Minds on Physics—affect student academic performance and attitudes towards science?*'

The primary focus of this study was to determine if the use of Blended Learning activities as correctives in a Mastery Learning approach had a significant impact on student achievement and attitudes towards science in an Australian high school Physics course. The study population consisted of Year 10 students from a high school in South East Queensland (N = 199). Students completed the Physics unit as part of their core Science course and rotated through a variety of topics throughout the year. All Physics classes were taught by the same teacher and followed the same curriculum and learning activities. Students were assigned to classes based on school conditions, and these classes were then allocated to the treatment or control group. The first group received access to the MOP program to provide mastery assessment and corrective activities

related to identified misconceptions, while the second group completed the same general learning activities but did not complete mastery assessment activities or correctives. Student achievement was measured using the FCI in a pre- and post-unit application, and student attitudes were measured pre- and post-unit using selected TOSRA scales.

The study was focused on responding to three groups of research questions. The first group relates to the effect of the MOP approach on student attainment.

RQ 1 Student Attainment

RQ 1a: Are pre-unit FCI scores for the treatment and control groups statistically significantly different?

RQ 1b:Are there statistically significant differences between the pre- and post-unit FCI scores for the control group?

RQ 1c: Are there statistically significant differences between the pre- and post-unit FCI scores for the treatment group?

RQ 1d: Are post-unit FCI scores for the treatment and control groups statistically significantly different?

The second group of research questions focused on the effect of the MOP approach on students' attitudes towards Science.

RQ 2 Student Attitudes

RQ 2a: Are pre-unit TOSRA scores for the treatment and control groups statistically significantly different?

RQ 2: Are there statistically significant differences between the pre- and post-unit TOSRA scores for the control group?

RQ 2c: Are there statistically significant differences between the pre- and post-unit TOSRA scores for the treatment group?

RQ 2d: Are post-unit TOSRA scores for the treatment and control groups statistically significantly different?

The third research question relates to any associations between student achievement and attitudes towards Science.

RQ 3 Associations

RQ 3: Are there associations between FCI scores and student attitudes towards Science as measured by the TOSRA scales?

5.3 Findings

5.3.1 Research Question 1

The first set of research questions relate to the effect of the BLMPC approach on student attainment as measured by the FCI. The BLMPC approach used curriculum aligned questioning to identify students' misconceptions and then a range of BL activities (simulations, animations, text information) as corrective activities prior to reassessment and progression. The use of Mastery progression approaches has been shown to be effective in a range of applications with average effect sizes of $d = 0.58$ across a large meta-analysis of previous research (Hattie, 2009). Of key importance in effective Mastery progression approaches is the type and quality of corrective activities (Guskey, 2010). The use of BL approaches such as simulations (Crook et al., 2014; Finkelstein et al., 2005) and MLA (Chandra & Fisher, 2009) have been shown to be effective in improving student understanding of a range of Science curriculum concepts.

RQ 1a:Are pre-unit FCI scores for the treatment and control groups statistically significantly different?

FCI scores were determined for all students involved in the study before the implementation of any learning activities. The mean FCI score for the control group was 7.6/30 with $SD = 2.4$ and for the treatment group was 8.2/30 with $SD = 2.9$. A t-test was conducted to determine if the difference in FCI scores between the two groups was

statistically significant. The results, $t(197) = 1.66$ and $p = 0.97$, indicate that the mean difference between the scores was not statistically significant. In conclusion, there is no significant difference in the FCI scores of students in the control and treatment groups, and this implies that students in each group had a comparable pre-unit understanding of the Newtonian Force Concept.

RQ 1b: Are there statistically significant differences between the pre- and post-unit FCI scores for the control group?

Pre- and post-unit FCI scores were determined and compared for students in the control group. The pre-unit FCI mean score was 7.6/30 with $SD = 2.4$ compared with a post-unit mean score of 9.8/30 with $SD = 3.3$. The mean difference of 2.2 was determined to be significant using a t-test, with $t(103) = 8.47$ and $p < 0.01$. The individual pre- and post-unit student FCI scores were used to calculate the normalized change and effect size. These give a measure of the amount of improvement in student attainment. The normalized change from pre- to post-unit for the control group was 0.08 ± 0.02 with $SD = 0.16$ and an effect size of 0.81.

The normalized change achieved by the control group is near the bottom of the range for traditional lecture classes (Von Korff et al. 2016). However, this data is from US and Canadian college students completing university Mechanics courses, and the variation in demographic, study cohort size and course type makes comparisons with the current study problematic. The effect size of 0.81, which is classed as a large effect (Hattie, 2009), shows a substantial change in students' pre- and post-unit FCI scores and hence their understanding of the Newtonian Force concept. The apparent disparity between the comparatively small normalized change and large effect size is explained due to the latter accounting for class size and the inclusion of variance in individuals' scores; hence the effect size is a more sensitive single number measure than the normalized gain (PhysPort, 2018).

In conclusion, there are statistically significant differences between the pre- and post-unit FCI scores of the control group, and this indicates an improvement in the understanding of the Newtonian Force Concept, possibly due to the learning activities of the course.

RQ 1c Are there statistically significant differences between the pre- and post-unit FCI scores for the treatment group?

Pre- and post-unit FCI scores were determined and compared for students in the treatment group. The pre-unit FCI mean score was 8.2/30 with $SD = 2.9$ compared with a post-unit mean score of 12.1/30 with $SD = 4.0$. The mean difference of 3.9 was determined to be significant using a t-test, with $t(95) = 13.70$ and $p < 0.01$. The individual pre- and post-unit student FCI scores were used to calculate the normalized change and effect size. These give a measure of the amount of improvement in student attainment. The normalized change from pre- to post-unit for the control group was 0.19 ± 0.01 with $SD = 0.13$ and an effect size of 1.47.

The normalized change achieved by the treatment group is close to the average for traditional lecture classes (Von Korff et al. 2016). However, this data is from US and Canadian College students completing university Mechanics courses, and the variation in demographic, study cohort size and course type makes comparisons with the current study problematic. The effect size of 1.47, which is classed as a large effect (Cohen 1969), shows a substantial change in students' pre- and post-unit FCI scores and hence their understanding of the Newtonian Force Concept. The apparent disparity between the normalized change and large effect size is explained due to the latter accounting for class size and the inclusion of variance in individuals' scores, leading to the effect size being a more sensitive single number measure than the normalized gain (PhysPort, 2018).

In conclusion, there are statistically significant differences between the pre- and post-unit FCI scores of the treatment group. This indicates an improvement in the understanding of the Newtonian Force Concept and may be due to the learning activities of the course and the BLMP approach.

RQ 1d Are post-unit FCI scores for the treatment and control groups statistically significantly different?

A comparison of the post-unit raw FCI scores of the treatment (12.1/30) and control groups (9.8/30) shows a statistically significant difference in the means of 2.3, with a t-test showing $t(186) = 4.48$ and $p < 0.001$. Further, there is a difference in both the normalized gain and effect size of the two groups, with the control group achieving $c = 0.08 \pm 0.02$ with $SD = 0.16$ and an effect size of 0.81, compared with the treatment group achieving $c = 0.19 \pm 0.01$ with $SD = 0.13$ and an effect size of 1.47. This gives a difference in normalized gain of 0.08 ± 0.03 and a difference in effect size of 0.66.

An analysis of individual question responses was also conducted. This determined that the treatment group showed the largest score improvement for 20/30 of the FCI questions. When results were grouped by the FCI dimensions (Hestenes et al., 1995), the treatment group showed the largest increase in all dimensions, with the most substantial difference in the kinematics dimension, a key aspect of the unit of study. The differences in score improvement between the dimensions is perhaps due to the Kinematics focus of the ACARA curriculum, whilst the Newtonian concepts developed in the course are sufficient to correctly address the other dimensions there is limited specific instruction and BLMPC activities related to circular or projectile motion. Responses were also analyzed to determine the types of variations in student pre- and post-unit item response. A number of students made correct responses in the pre-unit tests, but incorrect responses in the post unit test. When comparing the stability of correct responses in the treatment and control groups (students who maintained the

correct response in pre- and post-unit tests), it was found the treatment group was more stable in 27 of the 30 questions.

These findings reflect the research literature on the effect of mastery approaches to student achievement, with C. Kulik et al. (1990) finding a mean effect size of 0.52, and Bloom (1984) claiming mastery programs lead to an effect of one full standard deviation in the mean. The results in this study show a greater effect than those reported by Hattie (2009): 1.47 in this study compared to 0.58 in the Hattie meta-analysis of Mastery approaches. Studies of Blended Learning approaches have reported a range of impacts dependent of the type of activity used, from effect size of 0.35 for combination face-to-face and online instruction (Means et al., 2010) to effect sizes of 0.76 for ITS (VanLehn, 2011). The comparatively large effect size is perhaps due to the limited coverage of the Newtonian concept prior to the Year 10 ACARA curriculum, hence initial FCI scores were generally low.

In conclusion, the treatment group demonstrated significantly more improvement in the FCI raw score and normalized gain and effect size, demonstrated a larger improvement in all dimensions of the Newtonian Force Concept, and showed greater stability in correct responses from the pre- to post-unit test. Further, the improvements in the treatment group exceeded many of those reported in the literature for separate Mastery-based and Blended Learning approaches. It may be that this improvement is due to the combination of Mastery progression with Blended Learning corrective activities used in this study.

5.3.2 Research Question 2

The second set of research questions relate to student attitudes towards science. Students' attitudes towards a subject can have a significant impact on their engagement and attainment (Kind et al., 2007; Siegel & Ranney, 2003), and hence it is important to measure the impact of any novel approach on students' attitudes to ensure these

attitudes are not negatively impacted. Further, if any approach can lead to improvements in students' attitudes towards a subject, it may lead to an improvement in long term achievement. In the study of attitudes towards science, concern is frequently raised regarding the definition of 'attitude' and hence its measurement as a unidimensional construct (Osborne et al., 2003). It is apparent that there are many conflated concepts that relate to and determine attitude towards a subject (Kind et al., 2007), and that some of these attitudinal aspects are related to complex psycho-social parameters of the students' prior experiences in the subject. For this study, the measurement of attitudes towards science was limited to those factors that are most closely related to the curriculum context of the ACARA Science course. Mastery Learning approaches have been shown to have a positive influence on student affect towards a range of subjects (C. Kulik et al., 1990), although the correlation and breadth of study is weaker than with studies of attainment.

For this study, attitudes towards science were defined as being those attitudes measured by the TOSRA scales of: Adoption of Scientific Attitude (A), Enjoyment of Science Lessons (E), and Attitude to Scientific Inquiry (I). Scale I relates to a student's acceptance of the scientific investigative method of learning; this aligns with the inquiry model of learning used in both the treatment and control learning activities. Scale A relates to a student's acceptance of the willingness to change opinions or ideas based on evidence, a key parameter in overcoming misconceptions. Scale E aims to measure a student's enjoyment of Science lessons; this is an important parameter, as enjoyment may have an impact on engagement in learning activities.

The internal consistency of the TOSRA scales A, E and I were determined for the cohorts in the study. Internal consistency is a measure of the reliability of the test items; in the case of the TOSRA it refers to the consistency of responses within a scale, and hence provides a measure of the reliability of the scale in measuring a particular

attitude. Cronbach alpha values were calculated for both the control and treatment groups for pre- and post-unit TOSRA responses. Acceptable internal consistency was found for each scale for both groups pre- and post-unit. Control group alpha values were; Scale I pre–unit $\alpha = 0.65$ post-unit $\alpha = 0.76$, Scale A pre–unit $\alpha = 0.70$ post-unit $\alpha = 0.70$, Scale E pre–unit $\alpha = 0.70$ post-unit $\alpha = 0.72$. Treatment group alpha values were Scale I pre–unit $\alpha = 0.79$ post-unit $\alpha = 0.79$, Scale A pre–unit $\alpha = 0.77$ post-unit $\alpha = 0.76$, Scale E pre–unit $\alpha = 0.75$ post-unit $\alpha = 0.76$. In comparison with the values reported by when using a five-point Likert scale, the Scale I and E scores indicate lower consistency in this study (Fraser [1982] reported Scale I $\alpha = 0.86$ and Scale E $\alpha = 0.93$) and the Scale E score indicate higher internal consistency in this study (Fraser [1982] reported Scale A $\alpha = 0.67$).

In conclusion, the TOSRA scales A, I and E in the four-point Likert scale format used in this study show acceptable internal consistency, and hence each scale can be considered a reliable measure of a common attitude or opinion.

RQ 2a Are pre-unit TOSRA scores for the treatment and control groups statistically significantly different?

TOSRA scale scores were determined for all students involved in the study before the implementation of any learning activities. The control group had average pre-unit TOSRA scale scores of: I = 27.86, *SD* = 3.42, A = 28.76, *SD* = 3.62, E = 28.09, *SD* = 3.74 (10 items per scale, each item scored out of 4). This compared with the treatment group average scale scores of: I = 28.84, *SD* = 4.40, A = 29.85 *SD* = 4.21, E = 27.94, *SD* = 4.10 (10 items per scale, each item scored out of 4). T-tests were performed to determine the statistical significance of any differences in TOSRA scale scores; these showed no significant difference in any of the pre-unit scales scores between the control and treatment group, with Scale I $t(178) = 1.73$ $p = .08$, Scale A $t(186) = 1.93$ p=0.06, and Scale E $t(191) = 0.25$ $p = .80$.

In conclusion, the treatment and control groups had comparable pre-unit attitudes towards Scientific Inquiry, Adoption of Scientific Attitudes, and Enjoyment of Science Lessons. This implies that students in both the treatment and control groups were equally likely to accept the scientific investigative process as a way of obtaining information, be willing to adjust their ideas when presented with new information, and enjoy science lessons with the associated impact on engagement.

RQ 2b Are there statistically significant differences between the pre- and post-unit TOSRA scores for the control group?

Pre- and post-unit TOSRA scale scores were determined and compared for students in the control group. The control group had average pre-unit TOSRA scale scores of: I = 27.86, *SD* = 3.42, A = 28.76, *SD* = 3.62, E = 28.09, *SD* = 3.74 (10 items per scale, each item scored out of 4). This compared with the post unit average scale scores of: I = 28.52, *SD* = 4.16, A = 29.37, *SD* = 3.81, E = 28.75, *SD* = 4.07 (10 items per scale, each item scored out of 4). T-tests were performed to determine the statistical significance of any differences in pre- and post-unit TOSRA scale scores. Significant differences were found in TOSRA scales I, A and E.

For Scale I, $t(103) = 2.08$, $p = .04$, showing that the difference of 0.66 was significant but with a small effect size $d = 0.16$.

For Scale A, $t(103) = 2.49$, $p = .01$, showing the difference of 0.61 was significant but with a small effect size of $d = 0.16$.

For Scale E, $t(103) = 2.82$, $p = <.01$, showing the difference of 0.67 was significant but with a small effect size $d = 0.16$.

In conclusion, the TOSRA scale A, I and E scores showed a significant but small improvement between the pre- and post-unit response for the control group. This indicates that students' attitudes towards science improved slightly across all the

TOSRA scales used. This implies that the learning activities of the course may have had a positive impact on students' attitudes as measured by the three TOSRA scales.

RQ 2c Are there statistically significant differences between the pre- and post-unit TOSRA scores for the treatment group?

Pre- and post-unit TOSRA scale scores were determined and compared for students in the treatment group. The treatment group had average pre-unit scale scores of I = 28.84, *SD* = 4.40, A = 29.85 *SD* = 4.21, E = 27.94, *SD* = 4.10 (10 items per scale, each item scored out of 4). This compared with the post unit average scale scores of: I = 29.44, *SD* = 4.36, A = 30.26 *SD* = 3.93, E = 28.94, *SD* = 3.87 (10 items per scale, each item scored out of 4). It was found that there was not a statistically significant difference in the mean pre- and post-unit TOSRA Scale A scores for the treatment group. However, there was a statistically significant difference in the pre- and post-unit scores for TOSRA Scales I & E for the treatment group.

For Scale I, $t(94) = 3.02$, $p = <.01$, showing that the difference of 0.58 was significant but with a small effect size $d = 0.13$.

For Scale A, $t(94) = 1.67$, $p = .10$, showing the difference of 0.41 was not significant.

For Scale E, $t(94) = 5.5$, $p = <.01$, showing the difference of 0.99 was significant but with a small effect size $d = 0.23$.

In conclusion, the TOSRA Scales I and E scores showed a significant but small improvement, but Scale A showed no significant difference, between the pre- and post-unit response for the treatment group. This indicates that students' attitudes towards Enjoyment of Science Lessons (E) and Scientific Inquiry (I) improved, but their Adoption of Scientific Attitude (A) remained constant. This implies that the learning activities of the course and the BLMP approach may have had a positive impact on

students' attitudes towards using the scientific inquiry method as a way of learning new information, and their enjoyment of Science learning activities.

RQ 2d Are post-unit TOSRA scores for the treatment and control groups statistically significantly different?

A comparison of post-unit TOSRA scores between treatment and control groups shows the differences in Scales I, A and E scores are not significant. The control group had average post-unit TOSRA scale scores of: I = 28.52, *SD* = 4.16, A = 29.37, *SD* = 3.81, E = 28.75, *SD* = 4.07 (10 items per scale, each item scored out of 4). This compared with the treatment group post-unit average scale scores of: I = 29.44, *SD* = 4.36, A = 30.26 *SD* = 3.93, E = 28.94, *SD* = 3.87 (10 items per scale, each item scored out of 4).

T-tests were conducted to determine if the difference in scale scores was significant between the control and treatment groups' post-unit scores, and no significant differences were found.

For Scale I, $t(193) = 1.46$ $p = .15$, showing the difference of 0.92 is not statistically significant.

For Scale A, $t(194) = 1.62$ $p = .11$, showing the difference of 0.89 is not statistically significant.

For Scale E, $t(197) = 0.36$ $p = .71$, showing the difference of 0.19 is not statistically significant.

In conclusion, the treatment and control groups had comparable post-unit attitudes towards Scientific Inquiry, Adoption of Scientific Attitudes, and Enjoyment of Science Lessons. This implies that, in this study, neither the treatment nor control activities had a significant effect on students' attitudes towards Science.

5.3.3 Research Question 3

The final research question aimed to investigate any relationships between student attainment and attitudes towards Science. This was achieved by comparing pre-unit FCI and TOSRA scale scores, post-unit FCI scores and pre-unit TOSRA scores, and post-unit FCI and TOSRA scale scores. A review of the literature indicates that studies in this area are limited, with no studies found comparing these two metrics. Studies relating the attitudes of students towards Physics and their conceptual understanding have shown a positive correlation and scope for these attitudes to change, by varying amounts, when students are exposed to various learning experiences (Perkins et al., 2006).

RQ 3 Are there associations between FCI scores and student attitudes towards Science as measured by TOSRA scales?

Associations between TOSRA and FCI scores were investigated using a Pearson correlation. A small but significant correlation was found between all pre-unit attitudinal scales and students' pre-unit FCI scores (control group Scale I $r = 0.23$ $p = 0.008$, Scale E $r = 0.18$ $p = 0.001$, Scale A $r = 0.19$ $p = 0.001$; treatment group Scale I $r = 0.19$ $p = 0.001$, Scale E $r = 0.25$ $p = 0.001$, Scale A $r = 0.20$ $p = 0.001$). This implies that students who scored higher on the pre-unit attitudinal scales tended to score higher on the pre-unit FCI.

Post-unit FCI scores showed a small but significant correlation with pre-unit attitudinal scales for both control and treatment groups (control group Scale I $r = 0.26$ $p = <0.001$, Scale E $r = 0.22$ $p = <0.001$, Scale A $r = 0.23$ $p = 0.001$; treatment group Scale I $r = 0.18$ $p = <0.001$, Scale E $r = 0.43$ $p = <0.001$, Scale A $r = 0.19$ $p = <0.001$). This implies that students who scored higher on the pre-unit attitudinal scales tended to score higher on the post-unit FCI.

Post-unit FCI scores showed a significant correlation with post-unit attitudinal scales for both control and treatment groups (control group Scale I $r = 0.28$ $p = 0.008$,

Scale E $r = 0.23$ $p = 0.001$, Scale A $r = 0.23$ $p = 0.001$; treatment group Scale I $r = 0.14$ $p = 0.008$, Scale E $r = 0.42$ $p = <0.001$, Scale A $r = 0.16$ $p = 0.001$). This implies that students who scored higher on the post-unit attitudinal scales tended to score higher on the post-unit FCI.

In conclusion, there appear to be weak but positive associations between student attitudes towards science, as measured by the TOSRA scales I, E and A, and their attainment as measured by the FCI. This implies that students with more positive attitudes towards science demonstrated greater understanding of the Newtonian Force Concept. These findings align with those of other studies, such as those by Kind et al. (2007), Schommer (1994), and Siegel and Ranney (2003), who showed that student attitudes may affect student persistence, and hence performance. This may explain why students in the control group had the highest correlation between Enjoyment of Science lessons and post unit FCI score, as persistence may be a determining factor in the success of Mastery-based approaches.

5.3.4 Principle Research Question

The principal research question was '*Does a combination of the application of a Mastery for Learning (ML) approach with Blended Learning (BL) activities—specifically the use of Minds on Physics—affect student academic performance and attitudes towards science?'*

This question has a number of facets that were investigated using the range of research questions discussed above. The findings related to this principal research question were that there was a positive and significant improvement in student academic performance, as measured by the FCI, when the MOP platform is used to combine a ML approach with BL activities. While both control (d = 0.81) and treatment groups (d = 1.47) made improvements, the treatment group made greater gains in all the metrics investigated. Measurements of attitudes towards Science were made using

selected TOSRA scales. Whilst there was an improvement in all scales between pre- and post-unit analysis, there was no significant difference between the control and treatment groups. These results indicate that the use of the MOP intervention did not have an effect on the students' attitudes towards science.

5.4 Implications

This study investigated the use of Blended Learning activities as correctives in a Mastery Progression approach to learning aspects of the Newtonian Force Concept. It showed that the use of the MOP platform in a BLMPC led to improvements in understanding of the Newtonian Force Concept as measured by the FCI, but had no statistically significant effect on students' attitudes towards science. These findings indicate that the use of BL activities as correctives is an effective way of improving students' understanding of the Newtonian Force Concept when compared with the control approach adopted in this study.

Concerns regarding the application of ML approaches often relate to the amount of work and effort required in producing a range of effective corrective activities (Block & Anderson, 1975). This study shows that commercially available resources can be used to implement a ML approach in a Science classroom, and that these can be integrated into a regular teaching and learning framework with significant positive benefits.

Student attitudes towards a subject can be an important parameter in determining success and enjoyment, and this study has shown that learning activities can lead to improvements in students' attitudes towards Science. Further, it has shown that there is a correlation between students' attitudes towards Science and their attainment in the course. It may therefore be beneficial to develop more positive attitudes towards scientific inquiry, a greater enjoyment of science lessons, and more acceptance of scientific approaches, prior to a complex unit being taught.

5.4.1 Recommendations

It is recommended that teachers adopt a Mastery Learning approach to the Year 10 Physics curriculum and integrate the use of Blended Learning activities as correctives in the Mastery cycle. Care should be taken to ensure that Mastery Learning assessments are clearly matched to specific ACARA based learning goals and that the Blended Learning corrective activities provide a range of different learning approaches.

Further it is recommended that teachers adopt this approach to other areas of the curriculum and assess their impact on student achievement.

5.5 Limitations of the Study

The focus of this study was the effectiveness of the use of the MOP platform, to identify student misconceptions and provide corrective activities in a BL format, on student achievement and attitudes towards Science. The scope of the subject was limited to the Year 10 ACARA Forces and Motion content, which aligns with aspects of the Newtonian Force Concept.

Measurement of achievement was limited to students' responses to the FCI. As this is a multiple-choice concept inventory, there is little opportunity for students to explain their understanding in a detailed manner. Further, there are aspects of the Newtonian Force Concept that are beyond the scope of the ACARA course, and hence some of the FCI questions require students to apply their understanding in a manner that has not been explicitly taught during the course (for example, motion in two dimensions). Findings from this study may not be applicable to other topics within other Physics and wider Science curriculums.

Students' attitudes towards science were measured using selected TOSRA scales, and hence the findings are limited to the impact of the approach on adoption of scientific attitude, enjoyment of Science lessons, and attitude to scientific enquiry.

Conclusions regarding wider attitudes towards Science should not be drawn from this study.

The quality of corrective activities in ML is of obvious importance. The findings of this study are limited to the use of the MOP platform to deliver these corrective activities, and it may be that the positive impact of the BLMP approach in this study would not be replicated with other corrective activities.

Additional limitations exist due to a number of cohort features. The cohort was limited to a single year level in a single Australian high school (N = 199). It may be that features of this cohort affected the outcomes of the study. For example, it may be that the effectiveness of the approach is reliant on the cohort's high level of familiarity with ICT and associated learning activities, or that the expertise of the teacher had a significant effect on the class dynamic and attainment.

Limitations to the application of the findings may also be present due to assumptions made in the study which may affect the internal validity of the process. It was assumed that all students who participated in the study, in both the control and treatment groups, completed the learning activities during the course to the best of their abilities. The internal validity of the study could also have been compromised by students not completing the pre- and post-unit assessments accurately and honestly, or by control students obtaining access to the MOP platform before the completion of the study. There was no evidence of these issues occurring during the study.

5.6 Further Research

The results of this study show positive benefits of the use of a BLMP on student attainment in an ACARA Year 10 Physics context. Further research should focus on expanding this approach to a wider cohort, such as the QCAA Senior Physics syllabus. This study has not determined the causal factor in these improvements, and further research is required to determine if the personalized nature of the corrective activities,

or the repeated questioning methodology, or some other combination of factors, led to the improvement. It may also be that the relationships determined are only due to the MOP platform, and therefore other forms of BLMP approach should be investigated to determine if the findings can be more generally extrapolated to a variety of platforms. This could be further enhanced by moving outside of the Physics curriculum to the wider range of Science subjects.

The use of the FCI has been shown to be beneficial in this context. Although originally designed for use in higher level courses, it appears to provide a measure of student understanding of the Newtonian Force Concept. However, further research in the form of a factor analysis for this type of cohort is required to increase the understanding of the conceptual constructs that are actually being measured. A further focus on the right-to-wrong transitions may also prove beneficial in explaining the characteristics the FCI is measuring in this cohort.

The results collected for this cohort could also be used to further analyze the validity of the FCI as a measurement of students' understanding of the Newtonian Force concept. A factor analysis of the results using the EW5 and SSG5 factors may be beneficial in this circumstance.

Small but significant correlations were determined between attainment and attitudes that students held toward Science. An increase in qualitative data (such as more detailed questioning or focused interviews) may prove beneficial in determining the link between these two student parameters.

5.7 Chapter Review

This chapter used the data analysis from the study to address the research questions and place the findings in the context of the wider literature surrounding Blended and Mastery Learning, and the use of the FCI and TOSRA.

It was shown that the use of the MOP platform as a BLMPC approach to learning the ACARA Year 10 Physics topics of Force and Motion had a positive impact on student attainment, with an increase in FCI effect size of 0.66 over the control group. It was also shown that the approach led to a small but significant improvement in some aspects of students' attitudes towards Science as measured by the TOSRA Scales I and E.

Implications of these findings were discussed, with a recommendation for the use of this approach in the context of the study cohort. Finally, the limitations of the study were highlighted, in terms of the cohort characteristics, threats to the study validity, and the application of the findings to a wider population. These limitations provided the basis for the suggestions for further research.

References

ACARA. (2020). Content description: The Australian curriculum v8.1. Retrieved from http://www.australiancurriculum.edu.au/science/curriculum/f-10?layout=1#level10

Adams, W. K., Finkelstein, N. D., Reid, S., Dubson, M., Podolefsky, N., Wieman, C. E., & LeMaster, R. (2004, August). *Research-based design features of web-based simulations.* Paper presented at the AAPT Summer Meeting, Sacramento, CA. Retrieved from https://phet.colorado.edu/publications/Simulation%20Design%20AAPT%2004.pdf

Akpan, J. P. (2002). Which comes first: Computer simulation of dissection or a traditional laboratory practical method of dissection. *Electronic Journal of Science Education, 6*(4). Retrieved from http://wolfweb.unr.edu/homepage/crowther/ejse/akpan2.pdf

Ali, M., Mohsin, M., & Iqbal, M. (2013). The discriminant validity for Urdu version of Test of Science-Related Attitudes (TOSRA). *International Journal of Humanities and Social Science, 3*(2), 29-39.

Anderson, L. (1976). An empirical investigation of individual differences in time to learn. *Journal of Educational Psychology, 68*(2), 226–233. http://dx.doi.org/10.1037/0022-0663.68.2.226

Angell, C., Kind, P. M., Henriksen, E. K., & Guttersrud, Ø. (2008). An empirical-mathematical modelling approach to upper secondary physics. *Physics Education, 43*(3). https://doi.org/10.1088/0031-9120/43/3/001

Arlin, M. (1984). Time, equality, and mastery learning. *Review of Educational Research, 54*(1). https://doi.org/10.3102/00346543054001065

Aschbacher, P. R., Li, E., & Roth, E. J. (2009). Is science me? High school students' identities, participation and aspirations in science, engineering, and medicine. *Journal of Research in Science Teaching, 47, 564-582*. https://doi.org/10.1002/tea.20353

Bayraktar, S. (2001). A meta-analysis of the effectiveness of computer-assisted instruction in science education. *Journal of Research on Technology in Education, 34*(2), 173–188. https://doi.org/10.1080/15391523.2001.10782344

Black, P., & Wiliam, D. (2010). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 90*(1), 81–90. https://doi.org/10.1177%2F003172171009200119

Block, J. (1972). Student learning and the setting of mastery performance standards. *Educational Horizons, 50*(4), 183–191. Retrieved from http://www.jstor.org/stable/42925690

Block, J. (1977). Individualized instruction: A mastery learning perspective. *Educational Leadership, 34*(5), 337–341.

Block, J., & Anderson, L. (1975). *Mastery learning in classroom instruction*. Macmillan.

Block, J., & Burns, R. (1976). Mastery learning. *Review of Research in Education, 4*(1), 3–49. https://doi.org/10.3102/0091732X004001003

Bloom, B. (1968). Learning for mastery. *Evaluation Comment, 1*(2), 1–12.

Bloom, B. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher, 13*(6), 4–16. Retrieved from http://www.jstor.org/stable/1175554

Bloom, B. (1987). A response to Slavin's mastery learning reconsidered. *Review of Educational Research, 57*(4), 507–508. https://doi.org/10.3102/00346543057004507

Bloom, B., Hastings, J. T., & Madaus, G. F. (1971). *Handbook on formative and summative evaluation of student learning*. McGraw-Hill.

Bransford, J., Brown, A., & Cocking, R. (2000). *How people learn: Brain, mind, experience, and school*. National Academy Press.

Carroll, J. (1989). The Carroll model: A 25-year retrospective and prospective view. *Educational Researcher, 18*(1), 26–31. Retrieved from http://www.jstor.org/stable/1176007

Cavanagh, R. F., & Romanoski, J. T. (2007). Rating scale instruments and measurement. *Learning Environments Research, 9*(3), 273–289. https://doi.org/10.1007/s10984-006-9011-y

Chandra, V., & Fisher, D. (2005, November). *The application of the results of learning environments research to an innovative teacher-designed website*. Paper presented at the AARE Annual Conference, Parramatta. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.129.8999&rep=rep1&type=pdf

Chandra, V., & Fisher, D. L. (2009). Students' perceptions of a blended web-based learning environment. *Learning Environments Research, 12*(1), 31–44. https://doi.org/10.1007/s10984-008-9051-6

Chandra, V., & Watters, J. J. (2012). Re-thinking physics teaching with web-based learning. *Computers & Education, 58*(1), 631–640. https://doi.org/10.1016/j.compedu.2011.09.010

Chauhan, S. (2017). A meta-analysis of the impact of technology on learning effectiveness of elementary students. *Computers & Education, 105*, 14–30. https://doi.org/10.1016/j.compedu.2016.11.005

Chi, M. T., Siler, S.A., Jeong, H., Yamauchi, T.,Hausmann, R.G. . (2001). Learning from human tutoring. *Cognitive Science, 25*, 471–533. https://doi.org/10.1207/s15516709cog2504_1

Churchill, R., Ferguson, P., Godinho, S., Keddie, A., Letts, W., Mackay, J., McGill, M., Moss, J., Nagel, M., Nicholson, P., & Vick, M. (2013). *Teaching: Making a difference*. John Wiley and Sons.

Clark, M. C., & Sharf, B. F. (2007). The dark side of truth(s): Ethical dilemmas in researching the personal. *Qualitative Inquiry, 13*(3), 399–416. https://doi.org/10.1177%2F1077800406297662

Coe, R., Waring, M., Hedges, L. V., & Arthur, J. (2017). *Research methods and methodologies*. Sage.

Cohen, J. (1969) *Statistical power analysis for the behavioral sciences*. Academic Press.

Cohen, A., & Wollak, J. (2006). Test administration, security, scoring, and reporting. In R. Brennan (Ed.), *Test administration, scoring and reporting* (pp. 355–386). Praeger.

Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education* (6th ed.). Routledge.

Coletta, V. P., & Phillips, J. A. (2005). Interpreting FCI scores: Normalized gain, preinstruction scores, and scientific reasoning ability. *American Journal of Physics, 73*(12), 1172–1182. Retrieved from https://dx.doi.org/10.1119/1.2117109

Conole, G., de Laat, M., Dillon, T., & Darby, J. (2008). 'Disruptive technologies', 'pedagogical innovation': What's new? Findings from an in-depth study of students' use and perception of technology. *Computers & Education, 50*(2), 511–524. https://doi.org/10.1016/j.compedu.2007.09.009

Creswell, J. (2012). Choosing a mixed methods design. In *Educational Research*. Pearson.

Crook, S., Sharma, M. D., & Wilson, R. (2014). An evaluation of the impact of 1:1 laptops on student attainment in senior high school sciences. *International Journal of Science Education, 37*(2), 272–293. https://doi.org/10.1080/09500693.2014.982229

Crook, S., Sharma, M. D., Wilson, R., & Muller, D. A. (2013). Seeing eye-to-eye on ICT: Science student and teacher perceptions of laptop use across 14 Australian schools. *Australasian Journal of Educational Technology, 29*(1). https://doi.org/10.14742/ajet.72

Cuban, L. (2001). *Oversold and underused: Computers in the classroom*. Harvard University Press.

Damavandi, M. E., & Kashani, Z. S. (2010). Effect of mastery learning method on performance, attitude of the weak students in chemistry. *Procedia – Social and Behavioral Sciences, 5*, 1574–1579. https://doi.org/10.1016/j.sbspro.2010.07.327

Darling-Hammond, L. (2004). Standards, accountability, and school reform. *Teachers College Record, 106*(6), 1047–1085.

Davis, D., & Sorrell, J. (1995). Mastery learning in public schools. *Educational Psychology Interactive*. Retrieved from http://www.edpsycinteractive.org/files/mastlear.html

DeVore, S., Stewart, J., & Stewart, G. (2016). Examining the effects of testwiseness in conceptual physics evaluations. *Physical Review Physics Education Research, 12*(2), 020138. https://doi.org/10.1103/PhysRevPhysEducRes.12.020138

DeWeese, S. V., & Randolph, J. J. (2011, February). *Effective use of correctives in mastery learning*. Paper presented at the Association of Teacher Educators National Conference, Orlando, FL. Retrieved from https://eric.ed.gov/?id=ED523991

Dillashaw, F. G., & Okey, J. R. (1983). Effects of a modified mastery learning strategy on achievement, attitudes, and on-task behavior of high school chemistry students. *Journal of Research in Science Teaching, 20*(3), 203–211. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/tea.3660200304

Discipline Based Science Education Research Center. (2018). Assessments – Biological sciences. Retrieved from http://dbserc.pitt.edu/Assessment/Assessments-Biological-Sciences

Driscoll, M. (2002). Blended learning: Let's get beyond the hype. *E-learning*. Retrieved from http://elearnigmag.com/ltimagazine.

Dweck, C. (2007). *Mindset: The new psychology of success*. Random House.

Dweck, C. (2013). *Self theories: Their role in motivation, personality, and development*. Psychology Press.

Eaton, P., & Willoughby, S. D. (2018). Confirmatory factor analysis applied to the Force Concept Inventory. *Physical Review Physics Education Research, 14*(1). https://doi.org/10.1103/PhysRevPhysEducRes.14.010124

Education Council. (2018). *Optimising STEM Industry-School Partnerships: Inspiring Australia's Next Generation Final Report*. Retrieved from http://www.educationcouncil.edu.au/site/DefaultSite/filesystem/documents/Reports%20and%20publications/Publications/Optimising%20STEM%20Industry-School%20Partnerships%20-%20Final%20Report.pdf

Eyre, H. L. (2007). Keller's personalized system of instruction: Was it a fleeting fancy or is there a revival on the horizon? *The Behavior Analyst Today, 8*(3), 317–324. http://dx.doi.org/10.1037/h0100623

Finkelstein, N. D., Adams, W. K., Keller, C. J., Kohl, P. B., Perkins, K. K., Podolefsky, N. S., Reid, S., & LeMaster, R. (2005). When learning about the real world is better done virtually: A study of substituting computer simulations for laboratory equipment. *Physical Review Special Topics – Physics Education Research, 1*(1). https://doi.org/10.1103/PhysRevSTPER.1.010103

Francis, P., Figl, C., & Savage, C. (2009, January). *Mastery learning in a large first year physics class.* Paper presented at Motivating Science Undergraduates: Ideas and Interventions Conference. Sydney, Australia. Retrieved from: https://openjournals.library.sydney.edu.au/index.php/IISME/article/view/6218

Fraser, B. J. (1982). *TOSRA Test of Science Related Attitudes handbook*. Australian Council for Educational Research.

Fraser, B. J. (1998). Classroom environment instruments: Development, validity, and applications. *Learning Environments Research, 1*, 7–33.

Fraser, B. J., & Butts, W. L. (1982). Relationship between perceived levels of classroom individualization and science-related attitudes. *Journal of Research in Science Teaching, 19*(2), 143–154. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/tea.3660190206

Fraser, B. J., & Kahle, J. B. (2007). Classroom, home and peer environment influences on student outcomes in science and mathematics: An analysis of systemic reform data. *International Journal of Science Education, 29*(15), 1891–1909. https://doi.org/10.1080/09500690601167178

Gardner, P. L. (1975). Attitudes to science: A review. *Studies in Science Education, 2*(1), 1–41. https://doi.org/10.1080/03057267508559818

Goss, P., Hunter, J., Romanes, D., & Parsonage, H. (2015). *Targeted teaching: How better use of data can improve student learning*. Grattan Institute. Retrieved from http://www.grattan.edu.au/

Graesser, A., McNamara, D., & Graesser, A. (2010). Self-regulated learning in learning environments with pedagogical agents that interact in natural language. *Educational Psychologist – EDUC PSYCHOL, 45*. https://doi.org/10.1080/00461520.2010.515933

Graham, C. (2005). Blended learning systems: Definitions, current trends, and future directions. In C. J. Bonk & C. R. Graham (Eds.), *Handbook of blended learning: global perspectives, local designs (*pp. 3–20)*).* Pfeiffer Publishing.

Grant, L., & Spencer, R. (2003). The personalized system of instruction: Review and applications to distance education. *International Review of Research in Open and Distance Learning, 4*(2). https://doi.org/10.19173/irrodl.v4i2.152

Green, N. (2016). What to look for in a personalized learning plan. Retrieved from http://www.dreambox.com/blog/personalizedlearningplan#sthash.ubJ00yA3.dpuf

Guskey, T. (1980). Mastery learning: Applying the theory. *Theory Into Practice, 19*(2), 104–111. https://doi.org/10.1080/00405848009542882

Guskey, T. (1986). Implementing mastery learning. *NASSP Bulletin*, 70, 125–126. https://doi.org/10.1177%2F019263658607049033

Guskey, T. (1997). *Implementing mastery learning*. Wadsworth.

Guskey, T. (2007). Closing achievement gaps: Revisiting Benjamin S. Bloom's "Learning for mastery". *Journal of Advanced Academics, 19*(1), 8–31.

Guskey, T. (2010). Lessons of mastery learning. *Educational Leadership, 68*(2), 52–57. Retrieved from https://uknowledge.uky.edu/edp_facpub/14

Güzer, B., & Caner, H. (2014). The past, present and future of blended learning: An in depth analysis of literature. *Procedia – Social and Behavioral Sciences, 116*, 4596–4603. https://doi.org/10.1016/j.sbspro.2014.01.992

Hake, R. R. (1998). 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. https://doi.org/10.1119/1.18809

Hambleton, R. K. (2004). Theory, methods, and practices in testing for the 21st century. *Psicothema, 16*(4), 696–701.

Hanushek, E. (1997). Assesing the effects of school resources on student performance: An update. *Educational Evaluation and Policy Analysis, 19*(2), 141–164.

Harackiewicz, J. M., Barron, K. E., Tauer, J. M., Carter, S. M., & Elliot, A. J. (2000). Short-term and long-term consequences of achievement goals: Predicting interest and performance over time. *Journal of Educational Psychology, 92*, 316–330.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

Heller, P., & Huffman, D. (1995). Interpreting the force concept inventory: A reply to Hestenes and Halloun. *The Physics Teacher, 33*(8), 503–503. https://doi.org/10.1119/1.2344279

Henderson, C. (2002). Common concerns about the force concept inventory. *The Physics Teacher, 40*(9), 542–547. https://doi.org/10.1119/1.1534822

Hestenes, D., & Halloun, I. (1995). Interpreting the force concept inventory: A response to March 1995 critique by Huffman and Heller. *The Physics Teacher, 33*(8), 502. https://doi.org/10.1119/1.2344278

Hestenes, D., Wells, M., & Swackhamer, G. (1992a). Force concept inventory. *The Physics Teacher, 30*(3), 141–158. https://doi.org/10.1119/1.2343497

Hestenes, D., Wells, M., & Swackhamer, G. (1992b). A taxonomy of misconceptions. *The Physics Teacher, 30*, 141–158.

Hestenes, D., Wells, M., Swackhamer, G., Halloun, I., Hake, R. R., & Mosca, E. (1995). Revised force concept inventory. *The Physics Teacher*. Retrieved from http://modeling.asu.edu/R&E/Research.html

Hopkins, G. (2015). How can teachers develop students' motivation – and success? Retrieved from http://www.educationworld.com/a_issues/chat/chat010.shtml

Huffman, D., & Heller, P. (1995). What does the force concept inventory actually measure? *The Physics Teacher, 33*(3), 138–143. https://doi.org/10.1119/1.2344171

Inui, T. S. (2015). The charismatic journey of mastery learning. *Academic Medicine, 90*(11), 1442–1444. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/26375264

Jackson, J., Dukerich, L., & Hestenes, D. (2008). Modeling instruction: An effective model for science education. *Science Education, 17*(1) 10–17. Retrieved from https://www.researchgate.net/publication/265530286_Modeling_Instruction_An_Effective_Model_for_Science_Education

Kaplan, R. M., & Saccuzzo, D. P. (1997). *Psychological testing: Principles, applications, and issues* (4th ed.). Thomson Brooks/Cole Publishing Co.

Keller, F. S. (1968). "Good-bye teacher...". *Journal of Applied Behaviour Analysis, 1*, 79–89. http://dx.doi.org/10.1901/jaba.1968.1-79

Kind, P., Jones, K., & Barmby, P. (2007). Developing attitudes towards science measures. *International Journal of Science Education, 29*(7), 871–893. https://doi.org/10.1080/09500690600909091

Klopfer, L. (1971). Evaluation of learning in science. In B. Bloom, J. Hastings, & G. Madaus (Eds.), *Handbook on formative and summative evaluation of student learning*. (pp. 559–642) Mcgraw Hill.

Kulik, C., Kulik, J., & Bangert-Drowns, R. (1990). Effectivness of mastery learning programs: A meta analysis. *Review of Educational Research, 60*(2), 265–299.

Kulik, J., Bangert, R., & Williams, G. (1983). Effects of computer-based teaching on secondary school students. *Journal of Educational Psychology, 75*(1), 19–26. https://psycnet.apa.org/doi/10.1037/0022-0663.75.1.19

Kulik, J., C Kulik, & Smith, B. (1976). Research on the personalized system of instruction. *Innovations in Education & Training International*, *13*(1), 23–30. https://doi.org/10.1080/1355800760130104

Kulik, J., Kulik, C., & Carmichael, K. (1974). The Keller plan in science teaching. *Science*, *83*, 379–383.

Lasry, N., Rosenfield, S., Dedic, H., Dahan, A., & Reshef, O. (2011). The puzzling reliability of the force concept inventory. *American Journal of Physics*, *79*(9), 909–912. https://doi.org/10.1119/1.3602073

Levin, T. (1979). Instruction which enables students to develop higher mental processes. *Evaluation in Education*, *3*(3), 173–220. https://doi.org/10.1016/0191-765X(79)90006-5

Madsen, A., McKagan, S., & Sayre, E. C. (2017). Best practices for administering concept inventories. *The Physics Teacher*, *55*(9). https://doi.org/10.1119/1.5011826

Madu, N. E. (2010). *Associations between teachers' interpersonal behaviour, classroom learning environment and students' outcomes* [Unpublished doctoral thesis]. Curtin University.

Marx, J. D., & Cummings, K. (2007). Normalized change. *American Journal of Physics*, *75*(1), 87–91. https://dx.doi.org/10.1119/1.2372468

Masters, G. (2016, January 1). The 'long tail' of underachievement. *Teacher Magazine*. Retrieved from https://www.teachermagazine.com.au/columnists/geoff-masters/the-long-tail-of-underachievement

Matters, G., & Masters, G. N. (2014). *Redesigning the secondary–tertiary interface: Queensland Review of Senior Assessment and Tertiary Entrance. Extract: Recommendations and major features of proposed design*. Retrieved from https://www.acer.org/queensland-review

Mayer, R. E. (2009). Cognitive theory of multimedia learning. In *Multi-media learning*. Cambridge University Press.

Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist*, *38*(1), 43–52. https://doi.org/10.1207/S15326985EP3801_6

McRae, P. (2015). Myth: Blended learning is the next ed-tech revolution. *ATA Magazine*, *95*(4). Retrieved from http://www.teachers.ab.ca/Publications/ATA%20Magazine/Volume%2095%202014-15/Number-4/Pages/Myth-Phil-McRae.aspx

Means, B., Toyama, Y., Murphy, R., Bakia, M., & Jones, K. (2010). *Evaluation of evidence-based practices in online learning: A meta-analysis and review of online learning studies*. Retrieved from www.ed.gov/about/offices/list/opepd/ppss/reports.html

Melbourne Graduate School of Education. (2018). Mastery learning, Australian research summary. Retrieved from http://www.evidenceforlearning.org.au/the-toolkit/australasian-research-summaries/mastery-learning/

Miles, K. S. (2010). Mastery learning and academic achievement. Retrieved from http://search.proquest.com/docview/193327442

Miller, K., Lasry, N., Reshef, O., Dowd, J., Araujo, I., & Mazur, E. (2010, July). *Losing it: The influence of losses on individuals' normalized gains.* Paper presented at the American Institute of Physics PER Conference, Portland, Oregon. Retrieved from: https://www.per-central.org/items/detail.cfm?ID=10460

Morgan, K. (2011). *Mastery learning in the science classroom*. National Science Teachers Association.

Morris, G. A., Harshman, N., Branum-Martin, L., Mazur, E., Mzoughi, T., & Baker, S. D. (2012). An item response curves analysis of the Force Concept Inventory. *American Journal of Physics*, *80*(9), 825–831. https://doi.org/10.1119/1.4731618

Motamedi, V., & Sumrall, W. J. (2000). Mastery learning and contemporary issues in education. *Action in Teacher Education*, *22*(1), 32–42. https://doi.org/10.1080/01626620.2000.10462991

Muijs, D. (2013). *Doing quantitative research in education with SPSS*. SAGE Publications.

Muller, D. A., Eklund, J., & Sharma, M. D. (2005, November). *The future of multimedia learning: Essential issues for research*. Paper presented at the Australian Association for Research in Education, Parramatta, Australia. Retrieved from https://www.aare.edu.au/data/publications/2005/mul05178.pdf

Ngozi, A., & Chinedum, M. (2012). The effect of using mastery learning approach on academic achievement of senior secondary school II physics students. *Educational Technology*, *51*. 10735–10737.

Niemiec, R. P., & Walberg, H. J. (1985). Computers and achievement in the elementary schools. *Journal of Educational Computing Research*, *1*(4), 435–440. Retrieved from http://journals.sagepub.com/doi/abs/10.2190/7TP6-GYVE-1V8F-RGV7

Oliver, M., & Trigwell, K. (2005). Can 'blended learning' be redeemed? *E-Learning and Digital Media*, *2*(1), 17–26. https://doi.org/10.2304%2Felea.2005.2.1.17

Osborne, J., Simon, S., & Collins, S. (2003). Attitudes towards science: A review of the literature and its implications. *International Journal of Science Education*, *25*(9), 1049–1079. https://doi.org/10.1080/0950069032000032199

Patchan, M. M., Schunn, C. D., Sieg, W., & McLaughlin, D. (2016). The effect of blended instruction on accelerated learning. *Technology, Pedagogy and Education*, *25*(3), 269–286. https://doi.org/10.1080/1475939X.2015.1013977

Perkins, K. K., Gratny, M. M., Adams, W. K., Finkelstein, N. D., & Wieman, C. E. (2006). Towards characterizing the relationship between students' interest in and their beliefs about physics. *AIP Conference Proceedings*, *818*(1), 137–140. Retrieved from https://aip.scitation.org/doi/abs/10.1063/1.2177042

PhysPort. (2018). Validated assesment instruments. Retrieved from https://www.physport.org/assessments/

Queensland Tertiary Admission Centre. (2016). Applications and entry requirements. Retrieved from http://www.qtac.edu.au/applications/entry-requirements

Rosenberg, M., & Hovland, C. (1960). Cognitive, affective and behavioral components of attitudes: Attitude, organization and change. In M. Rosenberg (Ed.), *Attitude organization and change : an analysis of consistency among attitude component*. (pp. 1–14) Yale University Press.

Said, M. N. H. M., & Zainal, R. (2017). A review of impacts and challenges of flipped-mastery classroom. *Advanced Science Letters*, *23*(8), 7763–7766. https://doi.org/10.1166/asl.2017.9571

Savinainen, A., & Scott, P. (2002). The force concept inventory: A tool for monitoring student learning. *Physics Education*, *37*(1). https://doi.org/10.1088/0031-9120/37/1/306

Schommer, M. (1994). Synthesizing epistemological belief research: Tentative understandings and provocative confusions. *Educational Psychology Review*, *6*(4), 293–319. https://doi.org/10.1007/BF02213418

Scott, T. F., Schumayer, D., & Gray, A. R. (2012). Exploratory factor analysis of a force concept inventory data set. *Physical Review Special Topics – Physics Education Research*, *8*(2). https://doi.org/10.1103/PhysRevSTPER.8.020105

Shafie, N., Shahdan, T. N. T., & Liew, M. S. (2010). Mastery Learning Assessment Model (MLAM) in teaching and learning mathematics. *Procedia – Social and Behavioral Sciences, 8*, 294–298. https://doi.org/10.1016/j.sbspro.2010.12.040

Siegel, M. A., & Ranney, M. A. (2003). Developing the changes in attitude about the relevance of science (CARS) questionnaire and assessing two high school science classes. *Journal of Research in Science Teaching, 40*(8), 757–775. https://doi.org/10.1002/tea.10110

Slavin, R. E. (1987). Mastery learning reconsidered. *Review of Educational Research, 57*(2), 175–213.http://dx.doi.org/10.3102/00346543057002175

Smetana, L. K., & Bell, R. L. (2012). Computer simulations to support science instruction and learning: A critical review of the literature. *International Journal of Science Education, 34*(9), 1337–1370. https://doi.org/10.1080/09500693.2011.605182

Smith, J. I., & Tanner, K. (2010). The problem of revealing how students think: Concept inventories and beyond. *CBE: Life Sciences Education*, *9*(1), 1–5. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/20194800

Songer, N. B. (2007). Digital resources versus cognitive tools: A discussion of learning science with technology. In S. Abell & N. Lederman (Eds.), *Handbook of research on science education*. (pp 471–490) Lawrence Erlbaum Associates Publishers.

Sparks, S. (2016). Blended learning research yields limited results. *Education Week*. Retrieved from http://www.edweek.org/ew/articles/2015/04/15/blended-learning-research-yields-limited-results.html

Stein, J., & Graham, C. R. (2014). *Essentials for blended learning: A standards-based guide*. Taylor & Francis.

Stockwell, B. R., Stockwell, M. S., Cennamo, M., & Jiang, E. (2015). Blended learning improves science education. *Cell*, *162*(5), 933–936. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/26317458

Thompson, S., DeBrotolli, L., & Underwood, C. (2017). *PISA 2015: A first look at Australia's result*. Retrieved from https://research.acer.edu.au/ozpisa/22

Throndsen, I., & Turmo, A. (2013). Primary mathematics teachers' goal orientations and student achievement. *Instructional Science: An International Journal of the Learning Sciences, 41*(2). https://doi.org/10.1007/s11251-012-9229-2

University Admissions Centre. (2016). UAC undergraduate admission requirements. Retrieved from http://www.uac.edu.au/undergraduate/admission/index.shtml

University of Colorado Boulder. (2019). PhET research. Retrieved from https://phet.colorado.edu/en/research

Van der Linden, W. J., & Hambleton, R. K. (2013). *Handbook of modern item response theory*. Springer Science and Business Media.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist, 46*(4), 197–221. https://doi.org/10.1080/00461520.2011.611369

Viness, S., Colquitt, G., Pritchard, T., & Johnson, C. (2017). Using the personalized system of instruction to differentiate instruction in fitness. *The Physical Educator, 74*(3), 518–550. http://dx.doi.org/10.18666/tpe-2017-v74-i3-7420

Von Korff, J., Archibeque, B., Gomez, K. A., Heckendorf, T., McKagan, S. B., Sayre, E. C., Shenk, E. W., Shepherd, C., & Sorell, L. (2016). Secondary analysis of teaching methods in introductory physics: A 50 k-student study. *American Journal of Physics, 84*(12), 969–974. https://dx.doi.org/10.1119/1.4964354

Wallace, C., & Bailey, J. (2010). Do concept inventories actually measure anything? *Astronomy Education Review, 9*(1). http://dx.doi.org/10.3847/AER2010024

Wambugu, P. W., & Changeiywo, J. M. (2008). Effects of Mastery learning approach on secondary school students' physics achievement. *Eurasia Journal of Mathematics, Science & Technology Education, 4*(3), 293–302.

Wang, J., & Bao, L. (2010). Analyzing force concept inventory with item response theory. *American Journal of Physics, 78*(10), 1064–1070. https://doi.org/10.1119/1.3443565

Wieman, C., Adams, W., & Perkins, K. (2008). PhET: Simulations that enhance learning. *Science, 322*(5902), 682–683. Retrieved from http://science.sciencemag.org/content/sci/322/5902/682.full.pdf

Wolters, C. A. (2004). Advancing achievement goal theory: Using goal structures and goal orientations to predict students' motivation, cognition, and achievement. *Journal of Educational Psychology, 96*, 236–250.

Yasuda, J., Mae, N., Hull, M. M., & Taniguchi, M. (2019). Analyzing the measurement error from false positives in the force concept inventory. *Journal of Physics: Conference Series, 1287*. http://dx.doi.org/10.1088/1742-6596/1287/1/012033

Zhang, J., Chen, Q., Sun, Y., & Reid, D. J. (2004). Triple scheme of learning support design for scientific discovery learning based on computer simulation: Experimental research. *Journal of Computer Assisted Learning, 20*, 269–282.

Zimmerman, B. J., & Dibenedetto, M. K. (2008). Mastery learning and assessment: Implications for students and teachers in an era of high-stakes testing. *Psychology in the Schools, 45*(3), 206–216. https://doi.org/10.1002/pits.20291

Appendix 1: Sample Questions from the FCI

1. Two metal balls are the same size but one weighs twice as much as the other. The balls are dropped from the roof of a single story building at the same instant of time. The time it takes the balls to reach the ground below will be:
   (A) about half as long for the heavier ball as for the lighter one.
   (B) about half as long for the lighter ball as for the heavier one.
   (C) about the same for both balls.
   (D) considerably less for the heavier ball, but not necessarily half as long.
   (E) considerably less for the lighter ball, but not necessarily half as long.

2. The two metal balls of the previous problem roll off a horizontal table with the same speed. In this situation:
   (A) both balls hit the floor at approximately the same horizontal distance from the base of the table.
   (B) the heavier ball hits the floor at about half the horizontal distance from the base of the table than does the lighter ball.
   (C) the lighter ball hits the floor at about half the horizontal distance from the base of the table than does the heavier ball.
   (D) the heavier ball hits the floor considerably closer to the base of the table than the lighter ball, but not necessarily at half the horizontal distance.
   (E) the lighter ball hits the floor considerably closer to the base of the table than the heavier ball, but not necessarily at half the horizontal distance.

3. A stone dropped from the roof of a single story building to the surface of the earth:
   (A) reaches a maximum speed quite soon after release and then falls at a constant speed thereafter.
   (B) speeds up as it falls because the gravitational attraction gets considerably stronger as the stone gets closer to the earth.
   (C) speeds up because of an almost constant force of gravity acting upon it.
   (D) falls because of the natural tendency of all objects to rest on the surface of the earth.
   (E) falls because of the combined effects of the force of gravity pushing it downward and the force of the air pushing it downward.

Appendix 2: Taxonomy of Naïve concepts probed in the FCI

Table II. A Taxonomy of Naïve Conceptions Probed by the Inventory (August, 1995).

	Inventory Item
0. Kinematics	
K1. Position-velocity undiscriminated	19B,C,D
K2. Velocity-acceleration undiscriminated	19A; 20B,C
K3. Nonvectorial velocity composition	9C
K4. Ego-centered reference frame	14A,B
1. Impetus	
I1. Impetus supplied by "hit"	5C,D,E; 11B,C; 27D; 30B,D,E
I2. Loss/recovery of original impetus	7D; 8C,E; 21A; 23A,D
I3. Impetus dissipation	12C,D; 13A,B,C; 14E; 23D; 24C,E; 27B
I4. Gradual/delayed impetus build-up	8D; 10B,D; 21D; 23E; 26C; 27E
I5. Circular impetus	5C,D,E; 6A; 7A,D; 18C,D
2. Active Forces	
AF1. Only active agents exert forces	15D; 16D; 17E; 18A; 28B; 30A
AF2. Motion implies active force	5C,D,E; 27A
AF3. No motion implies no force	29E
AF4. Velocity proportional to applied force	22A; 26A
AF5. Acceleration implies increasing force	3B
AF6. Force causes acceleration to terminal velocity	3A; 22D; 26D
AF7. Active force wears out	22C,E
3. Action/Reaction Pairs	
AR1. Greater mass implies greater force	4A,D; 15B; 16B; 28D
AR2. Most active agent produces greatest force	15C; 16C; 28D
4. Concatenation of Influences	
CI1. Largest force determines motion	17A,D; 25E
CI2. Force compromise determines motion	6D; 7C; 12A; 14C; 21C
CI3. Last force to act determines motion	8A; 9B; 21B; 23C
5. Other Influences on Motion	
CF. Centrifugal force	5E; 6C,D,E; 7C,D,E; 18E
Ob. Obstacles exert no force	4C; 5A; 11A,B; 15E; 16E; 18A; 29A
Resistance	
R1. Mass makes things stop	27A,B
R2. Motion when force overcomes resistance	25A,B,D; 26B
R3. Resistance opposes force/impetus	26B
Gravity	
G1. Air pressure-assisted gravity	3E; 11A; 17D; 29C; 29D
G2. Gravity intrinsic to mass	3D; 11E; 13E; 29C
G3. Heavier objects fall faster	1A; 2B,D
G4. Gravity increases as objects fall	3B; 13B
G5. Gravity acts after impetus wears down	12D; 13B; 14E

Appendix 3: Test of Science Related Attitudes

The TOSRA is designed to measure 7 science related attitudes (Fraser 1982), for the purposes of this study the attitudes of students in the following areas will be measured:

Attitude to Scientific Inquiry

Adoption of Scientific Attitudes

Enjoyment in Science lessons

Attitude to Scientific Inquiry (Scale I)	Adoption of Scientific Attitudes (Scale A)	Enjoyment in Science lessons (Scale E)
1 (+)	11 (+)	21 (+)
2 (-)	12 (-)	22 (-)
3 (+)	13 (+)	23 (+)
4 (-)	14 (-)	24 (-)
5 (+)	15 (+)	25 (+)
6 (-)	16 (-)	26 (-)
7 (+)	17 (+)	27 (+)
8 (-)	18 (-)	28 (-)
9 (+)	19 (+)	29 (+)
10 (-)	20 (-)	30 (-)

Table A.3 Scale allocation and scoring for each item.

The responses are marked on a 5 point Likert scale from Strongly agree to Strongly disagree.

For positive items (+), responses SA, A, N, D, SD are scored 5,4,3,2,1 respectively.

For negative items (-), responses SA, A, N, D, SD are scored 1,2,3,4,5 respectively.

Omitted or invalid responses are scored 3.

(Adapted from Fraser, 1981)

Note: This page is not for the use of students

Test of Science Related Attitudes

Directions

This test contains a number of statements about science. You will be asked what you yourself think about the statements. There are no right or wrong answers. Your opinion is what is wanted.

All answers should be given on the separate Answer Sheet, please do not write on this booklet. USE A PENCIL

For each statement draw a circle around.

SA if you STRONGLY AGREE with the statement;
A if you AGREE with the statement;
D if you disagree with the statement;
SD if you strongly disagree with the statement;

If you change your mind about an answer erase your original answer and circle your final choice.

Although some statements are fairly similar to other statements, you are asked to indicate you opinion about all statements

Section 1 Scientific Inquiry
These questions are about your attitude to scientific inquiry.

1. I would prefer to find out why something happens by doing an experiment than being told.
2. Doing experiments is not as good as finding out information from teachers.
3. I would prefer to do experiments than to read about them.
4. I would rather agree with other people than do an experiment to find out for myself.
5. I would prefer to do my own experiments than to find out information from a teacher.
6. I would rather find out about things by asking an expert than by doing experiments.
7. I would rather solve a problem by doing an experiment than being told the answer.
8. It is better to ask the teacher the answer than to find out by doing experiments.
9. I would prefer to do an experiment on a topic than read about it in a science magazine.
10. It is better to be told scientific facts than to find them out from experiments.

Section 2 Scientific Attitudes
These questions are about your adoption of scientific attitudes

11. I enjoy reading about things that disagree with my previous ideas.
12. I dislike repeating experiments to check that I get the same result
13. I am curious about the world we live in.
14. Finding out about new things is unimportant.
15. I like to listen to people whose opinions are different from mine.
16. I find it boring to hear about new ideas.
17. In science experiments, I like to use new methods that I have not used before.
18. I am unwilling to change my ideas when evidence shows the ideas are poor.
19. In science experiments, I report unexpected results as well as expected ones.
20. I dislike listening to other people's opinions.

Section 3 Enjoyment
These questions are about your enjoyment of science.

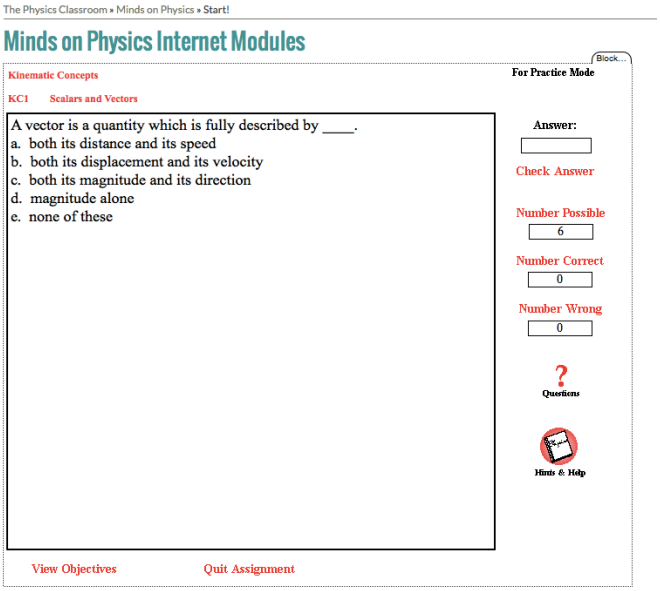
21. Science lessons are fun.
22. I dislike science lessons.
23. School should have more Science lessons each week.
24. Science lessons bore me.
25. Science is one of the most interesting school subjects.
26. Science lessons are a waste of time
27. I really enjoy going to science lessons.
28. The material covered in science lessons is uninteresting.
29. I look forward to science lessons.
30. I would enjoy school more if there were no science lessons.

Appendix 4: Minds on Physics Questions and Correctives

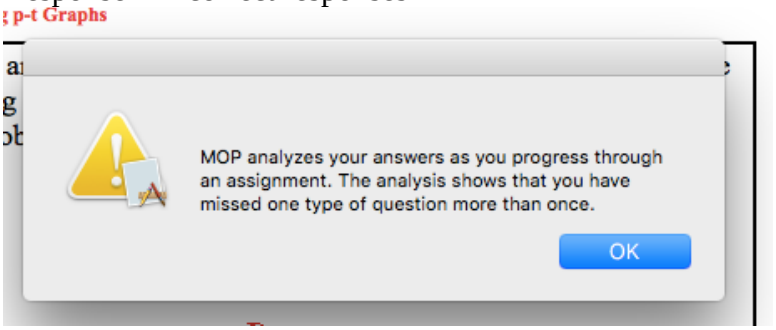
Example Objectives

- **Assignment 1:**
  - The student should understand the distinction between a vector and a scalar.
  - The student should be able to identify basic quantities which are vectors and scalars.

Example Question Page



Response if incorrect responses



Example Correctives

Text form

The Q: A scalar is a quantity which is fully described by ___.

Define Help **Definition of a Scalar:**
A **scalar** is a quantity which is fully described by magnitude alone.

Hot Link ➡ What is the difference between a scalar and a vector quantity?

Close Window

The Q: A vector is a quantity which is fully described by ___.

Define Help **Definition of a Vector:**
A **vector** is a quantity which is fully described by both a magnitude and a direction.

Hot Link ➡ What is the difference between a scalar and a vector quantity?

Scalars and Vectors

Introduction
Scalars and Vectors
Distance and Displacement
Speed and Velocity
Acceleration

Physics is a mathematical science. The underlying concepts and principles have a mathematical basis. Throughout the course of our study of physics, we will encounter a variety of concepts that have a mathematical basis associated with them. While our emphasis will often be upon the conceptual nature of physics, we will give considerable and persistent attention to its mathematical aspect.

The motion of objects can be described by words. Even a person without a background in physics has a collection of words that can be used to describe moving objects. Words and phrases such as *going fast, stopped, slowing down, speeding up,* and *turning* provide a sufficient vocabulary for describing the motion of objects. In physics, we use these words and many more. We will be expanding upon this vocabulary list with words such as *distance, displacement, speed, velocity,* and *acceleration.* As we will soon see, these words are associated with mathematical quantities that have strict definitions. The mathematical quantities that are used to describe the motion of objects can be divided into two categories. The quantity is either a vector or a scalar. These two categories can be distinguished from one another by their distinct definitions:

- **Scalars** are quantities that are fully described by a magnitude (or numerical value) alone.
- **Vectors** are quantities that are fully described by both a magnitude and a direction.

The remainder of this lesson will focus on several examples of vector and scalar quantities (distance, displacement, speed, velocity, and acceleration). As you proceed through the lesson, give careful attention to the vector and scalar nature of each quantity. As we proceed through other units at The Physics Classroom Tutorial and become introduced to new mathematical quantities, the discussion will often begin by identifying the new quantity as being either a vector or a scalar.

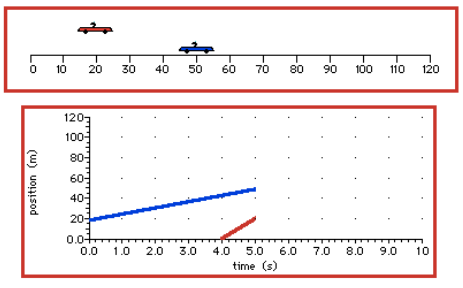
Experimental Form

One Dimensional Kinematics

Lab Descriptions (html) | Auxiliary Items | Scoring Rubrics

Title of Lab	Lab Description
Speedometer Lab	html
Speedometer Cubed Lab	html
Diagramming Motion Lab	html
Position-Time Graphs Lab	html
Interpreting the Slope Lab	html
Velocity-Time Graphs Lab	html
Match That Graph Lab	html
Two-Stage Rocket Lab	html
Free Fall Lab	html
Dune Buggy Challenge Lab	html

Interactive Simulation Form



Appendix 5: Consent Letter

A study of the Efficacy of "Minds on Physics" on High School Physics Student Achievement and Attitude

Dear Students,

As you are aware XXXXXXXX College has invested significant resources in the use of ICT in Teaching and Learning. As part of this program a number of different activities are being used across the school, one of these is the use of various online resources within the Science department. To see if these activities are successful you will be asked to complete some activities during lesson time. The results of these will be used to decide which resources are the most useful to aide student learning. The results of these studies will also be used as part of a research project being conducted by Professor Rob Cavanagh and Mr Sam Roberson in conjunction with Curtin Universities Science and Maths Education Center. All student results will be anonymous within this study and students may elect not to have their data shared with the study.

Project Details
The project is aimed at researching effective ways to use ICT within the Year 10 Physics curriculum. This will aide teachers in better targeting teaching activities to meet student needs and improve the learning outcomes for all students.

What do you have to do?
Complete some questionnaires and quizzes
Complete the learning activities set by your teacher

Is this extra work?
No, we are just collecting information about the work you already do.

Are there any risks associated with participation?
As the project involves activities central to the everyday activity of schools, there are no perceived risks in participation.

What are the benefits associated with participation?
Students will receive individualised reports on their understanding of topics and offered revision materials targeted at these misconceptions.

What will happen to the information provided?
All information provided will be kept confidential. Student data will be anonymous in the study.

What if I don't want to do it?

- Discuss it with your teacher or Mr Roberson.

Whom should I contact if I have any questions?
Curtin University Human Research Ethics Committee (HREC) has approved this study (HRE2017-0265). Should you wish to discuss the study with someone not directly involved, in particular, any matters concerning the conduct of the study or your rights as a participant, or you wish to make a confidential complaint, you may contact the Ethics Officer on (08) 9266 9223 or the Manager, Research Integrity on (08) 9266 7093 or email hrec@curtin.edu.au. or by post at HREC GPO Box U1987, Perth WA 6845
If you have any questions or you would like more information, please feel free to call or email me at any time. (Ph: 07 5474 0022; email sam.roberson@postgrad.curtin.edu.au or sroberson@bne.catholic.edu.au)

If you agree to your results being included in this study please complete the attached form and return it to your Science teacher. Please retain this information letter for your records.

Kind regards,

Sam Roberson
Leader of Learning & Assistant Curriculum Development Leader - Science
XXXXXXX College

Rob Cavanagh PhD
Chair of Pacific Rim Objective Measurement Society
Professor of well-being metrics
School of Education
Curtin University
Kent St
BENTLEY 6102

Tel 61 08 9266 2162
Fax 61 9266 2547

Please return this completed form to:

Sam Roberson
Assistant Curriculum Development Leader. - Science

I (student name)_____hereby give permission for my information and responses to Concept Inventories and Surveys to be used in a study to assess the effectiveness of various learning activities.

I (Parent/Guardian)_____hereby give permission for the above students information and responses to Concept Inventories and Surveys to be used in a study to assess the effectiveness of various learning activities.