2012

Alignment of intended learning outcomes, curriculum and assessment in a middle school science program

Reid J. Smith
Edith Cowan University

Edith Cowan University

Copyright Warning

USE OF THESIS

The Use of Thesis statement is not included in this version of the thesis.

Alignment of Intended Learning Outcomes, Curriculum and Assessment in a Middle School Science Program

Reid Smith

B.Sci (Biomed.), Grad. Dip Ed. (Sec.)

Student Number: 10048934

Submitted in partial fulfilment of the requirements of the Degree of Master of Education in the Faculty of Arts and Education, Edith Cowan University

Supervisors: Assoc. Prof. Graeme Lock

Prof. Mark Hackling

ABSTRACT

This study focused on the intended learning outcomes, curriculum and assessment in the science curriculum offered at a regional independent Middle School in the state of Victoria, Australia. In-school assessment has indicated that the current science curriculum of this Middle School may not develop students' skills in scientific literacy as effectively as intended. One hypothesis to explain this deficit is that there is a misalignment of intended outcomes, curriculum materials and assessment. This study aimed to determine the extent to which the intended curriculum and assessment in this Victorian middle years' science program is aligned to its stated goals and objectives and to design, implement and evaluate a model for assessing the degree of alignment of intended outcomes, curriculum and assessment.

Participants in the study were asked to analyse curriculum materials and assessment tasks from two different science courses at the case study school. These curriculum materials and assessments were scored against a series of instruments adapted from curriculum evaluation models used in previous research. The reviewers scored the material to determine the degree of alignment between the intended outcomes, curriculum materials and assessment tasks. The data provided an insight into both the degree of alignment of the curriculum as well as the features of strongly aligned curriculum materials. The effectiveness of the evaluation model was determined by analysis of the scoring data and semi-structured interviews with the participants.

The current investigation established that the case study Middle School science program had some degree of alignment, but there were a number of materials and tasks which were not adequately aligned. The features of the curriculum materials and assessment tasks generally matched those identified in the literature, and provided the basis for potential reform to increase the degree of alignment in intended curriculum and assessment in science courses designed to address scientific literacy.

The study also demonstrated that the model of curriculum evaluation was effective in establishing the alignment of curriculum materials and assessment with intended goals, particularly when enacted by teachers and administrators within the school context who had been trained. The curriculum analysis can highlight areas of the science curriculum which are not aligned and hence focus curriculum reform efforts.

DECLARATION

I certify that this thesis does not, to the best of my knowledge and belief:

(i) incorporate without acknowledgment any material previously submitted for a degree or diploma in any institution of higher degree or diploma in any institution of higher education;

(ii) contain any material previously published or written by another person except where due reference is made in the text of this thesis; or

(iii) contain any defamatory material.

(iv) contain any data that has not been collected in a manner consistent with ethics approval.

Signature Date 4/8/12

ACKNOWLEDGMENT

The completion of this study was made possible with the support and encouragement of a number of people.

First, I thank my wife Sally and daughters Susannah and Emily for their love and support across the time of this research study. Also, I give special thanks to Robert and Judi.

Also, my sincere thanks are extended to my supervisors, Associate Professor Graeme Lock and Professor Mark Hackling. Their expertise in both the subject matter and the intricacies of the thesis, genuine support, encouragement and commitment to what has been a significant journey has been much appreciated.

I also give thanks to the participants of this study, who gave significant amounts of time and energy to this project.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE: OVERVIEW OF THE STUDY

Introduction

Chapter one introduces the reader to the purpose and context of the research project. The first section discusses the background of the study, in effect setting the scene for the reader. It discusses the purposes of middle years science curricula, and describes the curriculum currently used in the case study school. The next section outlines the problem this research is designed to address. Section three deals with the significance of this research, justifying its importance to the field of science education, while the fourth section defines the specific research questions that this study has attempted to answer. The final section provides an outline of the thesis.

Background

For many years, middle years science curricula focussed particularly on the development of scientific knowledge (learning of key theories and facts of science) in preference to scientific skills (such as use of scientific equipment, development of an experimental method, interpretation of experiment results). Apart from some very specific programs, these curricula valued the memorisation of information with some requisite understanding of scientific phenomena (Carey, Evans, Honda, Jay & Unger, 1989; Chinn & Malhotra, 2002; National Research Council (NRC), 1996; Zimmerman, 2000). For example, the Curriculum Standards Framework used in Victorian schools until 2006, the CSF II, contained outcomes, which focussed on knowledge of science rather than scientific literacies (Victorian Curriculum Assessment Authority, 2000). This contrasts with the Australian Academy of Science's stance on Scientific Literacy (Hackling & Prain, 2005), which emphasises the importance of scientific literacy in being able to engage with and solve problems within real world contexts.

A number of studies have recognised that the key goal of a middle years science program should be to increase students' scientific literacy (Goodrum, Hackling & Rennie, 2001; Millar & Osborne, 1998; National Research Council (NRC), 1996). This is reflected in the Australian Science Curriculum produced as the new national curriculum framework for science education (Australian Curriculum, Assessment and Reporting Authority (ACARA, 2011). This statement is designed to guide the creation

of science curriculum in each of Australia's states and territories, and it acknowledges the need for developing the inquiry skills that are at the heart of scientific literacy. The national curriculum, along with the aforementioned middle years research reports, clearly show a need to adjust the content of science curricula to reflect this goal of developing students' scientific literacy.

The development of the Victorian Essential Learning Standards (VELS) by the Victorian Curriculum and Assessment Authority (VCAA) in 2005 is consistent with the national curriculum framework. This VELS curricula, introduced into both public and private education sectors, now features a skills-based approach which requires educators to change both instructional style and assessment methods in order to most effectively develop the specified skills.

This study examines a middle years (students aged 10 – 15) science curriculum created and implemented in a regional, independent K-12 school. The curriculum was developed by the school's science teachers who had experience in both teaching scientific concepts and skills as well as curriculum design, in conjunction with external consultants Margaret Forster of the Australian Council of Education Research (ACER) and Stephan Millett from the Wesley College Middle School in Western Australia. The curriculum has been in existence since 2002 and is remarkably similar to the VELS program considering it predates the state curriculum by three years.

The middle years curriculum in the case study school followed the Victorian state school curricula (CSF, CSF II) closely during the 1990s. Later, the case study school chose to develop a new course based on the teaching and assessment of skills rather than a heavy emphasis on content knowledge. Thus, the middle years program is broken into eight key learning areas (Thinking, Literacy, Mathematics, Global Learning, Languages other than English (LOTE), Health and Physical Education, Visual Arts and Performing Arts), each with its own set of essential skills and understandings.

The school's middle years science curriculum (known as Thinking Science) has a specific set of Essential Learning Outcomes (ELOs) against which the students are assessed over their time in the Middle School (listed in Appendix A). The ELOs are the skills judged by the academic staff of the case study school to be essential to develop

students' scientific literacy as they approach their non-compulsory studies and life post-schooling.

The purpose of this middle years science curriculum is to develop the inquiry skills that contribute to the development of students' scientific literacy. It was intended that the traditional science topic areas, such as atomic theory, schemes of classification and the behaviour of light would provide conceptual contexts for the teaching of science inquiry skills used in the collection, analysis and communication of evidence. To emphasise the importance of scientific literacy, formal assessment is made primarily of science inquiry skills. The science concepts are used both to provide a context for the teaching of inquiry skills, and are also embedded in the assessment used to assess student achievement. The scientific literacy skills of each student are tracked using a continuum (also known as a progress map, as shown in Appendix B).

The students are assessed according to the goals of the program. Online reports and formal feedback relate only to the ELOs, as they are the only outcomes formally assessed by this curriculum. Although conceptual knowledge is addressed, developed and assessed, formal reporting only occurs for the process outcomes. The students are assessed on these ELOs by use of a school-developed progress map. The performance of students in each of the ELOs is monitored and developed throughout their time in the Middle School.

Student progress in the case study school is monitored by Heads of Middle School using both internal and national standards testing such as the International Competitions and Assessments for Schools (ICAS) program provided by the University of New South Wales. This testing allows the school to triangulate the data provided by the internal assessment, which is important for the verification of quality of instruction and perceived progress of students (Boudett, City & Murnane, 2005). The ICAS test focuses on the domains of Measuring and Observing, Interpreting Data, Predicting/Concluding from Data, Investigating and Reasoning/Problem Solving. Each of these test domains map across aspects taught in the science program. Although the ICAS testing is a single event that uses multiple choice questions to test understanding, and only addresses seven of the 14 Essential Learning Outcomes, it is the best external measure the school currently has available to validate its internal assessments.

Appendix C indicates which of the ELOs are addressed by the ICAS test and which are not.

Each year, the senior leadership team use results of previous years to estimate the level of performance expected by students on the Science ICAS testing. The ICAS results provide three key pieces of data. The first is a raw test score, based on the number of items correctly answered by each student. The second is a percentile ranking for each student, comparing the student's raw score to the results of students in the same year level state-wide. The last piece of data is a standardised score with a maximum rating of 100, against which the student is tracked over time.

Problem

Given the specific focus and curriculum time devoted to developing students' scientific literacy, it was anticipated by the science staff at the case study school that the Years 5 to 9 cohorts would achieve two benchmarks:

1. The students would progress at a rate three standardised points greater than the average state progression.
2. The students in each year level would average three raw score marks above the state test mean score in the ICAS testing.

However, results have shown that the students science inquiry skills are not progressing as quickly as was anticipated, with the cohort mean lying on or just above the state mean, which is well below the expected three mark differential. Secondly, students attending the case study Middle School are progressing at a rate only slightly greater (8.2 points) than the rate of a student in the state-wide cohort over a year (7.6 points) in questions relating directly to science inquiry skills, which again is less than that expected at the case study school, given rates of improvement in other learning areas on similar assessments (International Competitions and Assessment for Schools (ICAS) Report, 2008).

The Thinking Science curriculum occupies a single block in a six block timetable, each of which has 230 minutes per week. This means Thinking Science has between three and four 70 minute lessons per week, as a rotating timetable exists on alternate

Mondays. When considering the time given to the development and assessment of these skills, however, the marginal difference in ICAS score progression is not as great as the program was expected to produce. The other area of concern is the significant difference in student performances from one class to the next. At this stage, there is a concern that the class to which a student is assigned significantly limits the learning that they are able to achieve in a year. This raises concerns about whether the current curriculum is achieving its intended goals of improving students' scientific literacy.

One hypothesis for this lack of student improvement is that the taught curriculum and assessment currently used to address scientific literacy are misaligned with the intended goals of the Thinking Science curriculum.

Rationale and Significance

A program, which intends to directly teach a particular skill set, but has curriculum materials and assessment that do not match this goal will have limited effectiveness. Some assessment tasks have already been identified by subject matter experts as poor indicators of student performance. It is possible that these materials could be negatively impacting on student progress. It is important to ensure that the curriculum, assessment and instruction in the science program are aligned, as the research literature indicates that constructive alignment enhances learning outcomes (Biggs, 1996).

The introduction of the Australian Curriculum, with its ties to school resourcing, will mean that a large number of schools and departments will undergo a period of curriculum realignment. The ability of a school, particularly those in the independent sector, to be able to determine the degree of alignment of their curriculum to both the Australian Curriculum and the associated National Assessment Program: Literacy and Numeracy (NAPLAN) becomes an important factor in their ability to both attain funding and to improve school performance as reported on the MySchool website (http://www.myschool.edu.au). Given the complexity of re-aligning curriculum, a framework for alignment which can be used by a school's teachers and administrators would prove useful.

This research project will make a contribution to knowledge in science education in a number of areas. The data collected from the proposed study should provide insights

into how well the assessment and curriculum aligns with the stated goals of the curriculum. In a local sense, it should allow realignment of the implemented middle years' science curriculum at the case study school. Consequently, the Researcher will be able to identify methods by which curriculum and assessment could be strengthened in order to achieve its stated goals. By ensuring that the curriculum and assessment are properly aligned with the goals of the program, the program itself should provide better outcomes for the students.

This investigation will also contribute to knowledge in the field of constructive alignment of middle years' science curriculum, as it aims to develop an approach for assessing the alignment of intended outcomes, curriculum and assessment. At this point, although a number of models for evaluating alignment have been proposed, few of them have been reviewed for effectiveness. In particular, this research aims to develop and evaluate a model for assessing the alignment the intended outcomes, assessment and curriculum, as applied by subject experts within a working school environment.

Purpose and Research Questions

The purpose of this research was to develop, implement and evaluate a method for evaluating the alignment of intended outcomes, curriculum materials and assessment in a Middle School science program.

Specifically the research project focuses on two questions:

1) To what extent are the intended outcomes, curriculum and assessment in this Middle School science curriculum constructively aligned?

2) How effective is the curriculum evaluation model developed and implemented in this study for evaluating the alignment of intended outcomes, curriculum materials and assessment?

Outline of Thesis

Chapter two presents a review of the literature that relates to the aims and objectives of the study. The review first considers the purpose of science education, the nature of science curricula and assessment. The chapter then describes a historical perspective of the development of science curricula, as well as an analysis of alignment in science curricula. Next, the importance of alignment of intended outcomes, curriculum and assessment in secondary schools is emphasised, as well as a discussion of the common models of alignment. This discussion is used to generate a conceptual framework for this study synthesised from the work of Webb (1997), Chinn and Malhotra (2002), and Kesidou and Roseman (2002), which was used to guide both the evaluation and subsequent revision of the case study assessment and curriculum materials.

Chapter three discusses the methodology used in this research, including its design, procedure and instruments, analysis of data, and limitations. The next chapter presents the data collected whilst considering whether the Middle Years science program is constructively aligned. Chapter five considers the alignment methodology itself using statistical methods and interview data from the reviewers. Chapter six discusses and analyses in detail the findings of the current study in the context of the research literature. Finally, chapter seven, highlights a series of recommendations which emerged from the research findings, and provides a conclusion to the study.

Summary

This chapter has established the context in which the research will occur. The first section provided background information identifying the importance of the curriculum design and implementation, and the possible misalignment of objectives in the case study school. Section two identified the research problem. The third section discussed the uniqueness of this study, identifying a lack of research in the area of curriculum alignment tools, particularly when dealing with science literacy programs studied in situ. Section four outlined the rationale and significance of this research and section five outlined the purpose and the two broad research questions that this study attempted to answer.

The next chapter includes a review of the relevant literature which defines the concepts of curriculum and assessment, and describes the importance of aligning curriculum, assessment and instruction in an education program, the structure of which was discussed in the outline of the thesis.

CHAPTER TWO: REVIEW OF THE LITERATURE

Introduction

A comprehensive review of the literature relevant to the research is presented in this chapter. The first section discusses the purpose of science education, defining scientific literacy and highlighting the importance of fundamental scientific literacies and epistemological beliefs in science. This section also discusses pedagogical approaches to science education. The second section defines curriculum, then discusses the design of science curriculum, both intended and implemented. The third section describes current assessment practices in science. Section four considers the importance of curriculum alignment, particularly in the area of science education. A general definition of alignment and a brief description of both the backwards and constructive curriculum design process follows. Section six reviews the variety of approaches used to analyse curricula, considering a wide range of different models. The seventh section presents the conceptual framework around which the research conducted in this project was based by considering the role of assessment, curriculum materials and intended outcomes in student learning, and what methods could be used to develop alignment. The framework also considers how backwards design and constructive alignment fit into the development of a coherent curriculum framework. The final section provides a conclusion and briefly summarises the key issues discussed in the literature review.

Purpose of Science Education

In recent years, a number of reports (Goodrum, Hackling & Rennie, 2001; Carey et al., 1989; Chinn & Malhotra, 2002; NRC, 1996) have identified the most important aspects of compulsory science education in the middle years of schooling (ages 10 – 15). Traditionally, the science curriculum has offered a series of modules: for example Light, Earth and Space, set within specific science disciplines (DeBoer, 1991; Gallagher, 1991; Hodson, 1998). At times, traditional courses attempt to develop understanding of scientific methods, such as developing an awareness of a fair experiment, which involves a focus on the control of multiple variables. Contrary to the content of these traditional syllabi and curriculum frameworks, recent studies (Goodrum et al., 2001; NRC, 1996; Carey et al., 1989; Chinn & Malhotra, 2002) have shown that the primary

purpose of science education in the compulsory years should be to develop scientifically literate citizens. This has been recognised with the introduction of an inquiry strand in the Australian Curriculum (ACARA, 2011).

Defining Scientific Literacy

There is much variation in the definition of scientific literacy in the literature. Roberts (2007) classifies the various conceptions of scientific literacy along a dimension with Vision I and Vision II as the poles of the dimension. Vision I conceptions look inwards at the workings of science itself, the processes of science as well as the laws and principles which are derived from its study. Vision I would include the knowledge of scientific method, how to control variables and when to confirm or refute a hypothesis. Vision II ideas, however, tend to look outwards from science; the effects science has on community discourse and decision making on socio-scientific issues. A good example of Vision II scientific literacy is the ability to use appropriate scientific information in the debate on climate change. Most definitions of scientific literacy presented in the literature lie on a continuum between these two visions of scientific literacy.

The United States' National Research Council (1996) defines scientific literacy as:

> Scientific literacy means that a person can ask, find, or determine answers to questions derived from curiosity about everyday experiences. It means that a person has the ability to describe, explain, and predict natural phenomena. Scientific literacy entails being able to read with understanding articles about science in the popular press and to engage in social conversation about the validity of the conclusions. Scientific literacy implies that a person can identify scientific issues underlying national and local decisions and express positions that are scientifically and technologically informed. A literate citizen should be able to evaluate the quality of scientific information on the basis of its source and the methods used to generate it. Scientific literacy also implies the capacity to pose and evaluate arguments based on evidence and to apply conclusions from such arguments appropriately.
>
> (NRC, 1996, p. 22)

A similar view of scientific literacy is presented by Goodrum et al. (2001) in a review of the status of teaching and learning in Australian schools. These authors indicate that a scientific literate person should be able to:

- be interested in, and understand the world around them;
- engage in discourses of and about science;
- be sceptical and questioning of claims made by others about scientific matters;
- be able to identify questions, investigate and draw evidence-based conclusions; and
- make informed decisions about the environment and their own health and well-being.

(Goodrum et al., 2001, p. 7)

A group which presents a Vision II view of scientific literacy is the Organisation for Economic Co-operation and Development (OECD) who, in their Programme for International Student Assessment (PISA) study, define scientific literacy as:

> an individual's scientific knowledge and use of that knowledge to identify questions, to acquire new knowledge, to explain scientific phenomena, and to draw evidence-based conclusions about science-related issues, understanding the characteristic features of science as a form of human knowledge and inquiry, awareness of how science and technology shape our material, intellectual and cultural environments, and willingness to engage in science-related issues, and with the issues of science, as a reflective citizen.
>
> (OECD, 2006, p. 12)

Scientific literacy, in the context of this research project, describes the ability to comprehend and communicate scientific information, as well as pose questions, observe, analyse and develop evidence-based conclusions from scientific investigations. It is essentially the competencies required for active participation in scientific investigation. For the most part it is a Vision I definition, but also embraces elements of a Vision II scientific literacy program. This is particularly evident in the Chemistry unit at Year 9 in the case study school, where the students spend a significant amount of time testing hypotheses about a series of 'drugs' being released onto the market (practising the Vision I science process skills) and then reflecting on the impact that their 'research' would have on the company and consumers (Vision II).

The definition of scientific literacy presented by this Middle School program involves not only the ability to use the literacies of science to communicate scientific ideas and information, but also the ability to use scientific concepts and principles to make sense of the world around them. According to Hackling and Prain (2005), scientific literacy is important because it "encompasses a range of learning outcomes that enable individuals

to navigate their way through life, rather than focusing solely on preparing them for future studies of science in the non-compulsory years" (p.17). A scientifically literate person has a positive disposition to engage with scientific issues and uses conceptual understandings, science processes and literacies of science to solve problems within real-world contexts (Hackling & Prain, 2008) as seen in Figure 1.

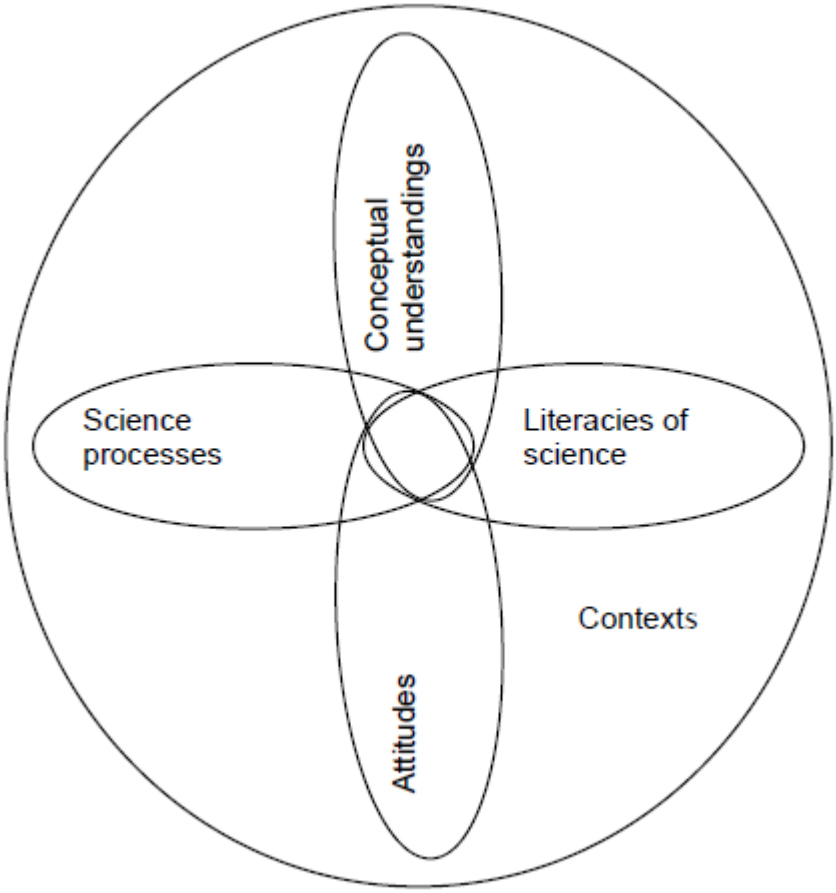

Figure 1: Scientific Literacy – A Multidimensional Construct (from Hackling & Prain, 2008, p. 7)

Hackling and Prain (2008) argue that to communicate scientific ideas and evidence requires mastery of scientific specific literacies and representational forms. Science has its own social language – a range of communication styles and techniques which are peculiar to science (Mortimer & Scott, 2003). The ability to communicate observations and insights in conventional ways is an important part of a science program which has a focus on scientific literacy.

The importance of the dialogic nature of coming to an understanding of science is emphasised by Mortimer and Scott (2003), and they explain that the laws and theories of science

> are developed within the science community and have been, and continue to be, subject to social validation….Science can thus be seen as a product of the scientific community, a distinctive way of talking and thinking about the natural world, which must be consistent with the happenings and phenomena of that world.
>
> (Mortimer & Scott, 2003, p.12 – 13)

For a child to engage in the learning of science and use of science in the everyday world they must build a specialist vocabulary to express their ideas. The greater the gap between the everyday perception of an event and the science views of that event, the greater the demand for the specialist vocabulary and representational forms. Mortimer and Scott (2003) contend that even the methods of arguing in a science context are necessary for the proper learning of science.

Hackling and Prain (2008) demonstrate that the literacies of science are not independent of the other aspects of scientific literacy. For a science investigation to be conducted appropriately, literacies of science need to be used to represent data generated from the experimentation. Data patterns and relationships are identified using the science processes and then reported using the science literacies (Hackling & Prain, 2008). These literacies and processes of science are the inquiry skills of the Australian Curriculum.

The ELOs of the middle years science curriculum in this case study are used to assess students' progress towards scientific literacy (Vision I and II). Appendix D contains a list of the ELOs for this science curriculum, with an indication of whether they are aligned to the Vision I or Vision II models proposed by Roberts (2007).

Process skills and epistemological beliefs

The process skills and epistemological beliefs, as defined by Carey, et al. (1989), are important aspects of scientific literacy. Process skills include observation, measurement and designing fair experiments. Many curricula offer the opportunity for students to develop their process skills through a range of exercises and use of scientific methods. Carey et al. (1989) also emphasise the need to develop in students an understanding of the nature and goals of science which are a valued facet of scientific literacy (Carey et al., 1989; Chinn & Malhotra, 2002); for example, an understanding of fair

experimentation does not necessarily automatically include the understanding of the purpose of experimentation.

Kuhn and Phelps (1982) demonstrated that students in the middle years can often have difficulty in understanding the importance of experimentation. The students struggled to determine the difference between theory and evidence. Students tend to see evidence existing only as an example of the theory, rather than understanding that the evidence is independent of the theory. Student understanding of the theoretical basis of science can be improved by instruction (Carey et al., 1989). Therefore, a middle years science curriculum based on scientific literacy, should include processes involved in authentic science investigations and developing a broad understanding of the nature of science (Lederman, 2006).

Pedagogical Approaches to Developing Scientific Literacy

The development of scientific literacy differs from the traditional model of science education. Aikenhead (2006) argues that traditional science teaching focuses mainly on the transmission of canonical disciplinary ideas, and, despite efforts to reform science teaching over the last part of the 20th century, there has been resistance to change due to the enculturation of science teachers by their own science schooling. Tytler (2007) believes that the traditional science model, which serves to preserve the status of scientific knowledge for the elite, needs to change so that all students have access to, and enthusiasm for, the concepts and literacies of science. In his article, Tytler argues that the literacies of science and student interest are best developed when an inquiry and discursive based (social constructivist) method is utilised. Cavagnetto (2010) also argues that argument-based interventions are a key facet of teaching scientific literacy.

Biggs (1996) describes the constructivist approach as occurring when meaning is not imposed or transmitted by the teacher, but rather it is created through the students' learning activities and assessment. According to this view, a student upon whom meaning is imposed will tend to learn the supplied information without any depth of understanding (surface learning), and hence will be unable to integrate this knowledge with their previous knowledge and understanding. An example of surface learning would be the ability of a student to define the Law of Reflection without being able to apply this concept to a practical situation.

In the investigation of the practices of outstanding science educators Tytler, Waldrip and Griffiths (2004) developed a set of principles that recognised effective teaching and learning of science in the Science in Schools (SiS) project. Of the eight components in the SiS, almost all of them link directly to the constructivist perspective described by Biggs (1999). As a conclusion to the study, Tytler and colleagues describe the best practice by science teachers as that which considers the learner as "an active sense-maker who engages with phenomena and ideas in order to construct knowledge" (p. 187).

The inquiry based approach is an integral component of a learning environment in which the learner acts as a sense maker. In inquiry learning, students undertake investigations in which they have the opportunity to practise the full range of science inquiry skills including: formulating research questions or hypotheses, designing experiments, collecting and interpreting scientific observations, and developing conclusions to communicating their findings.

The Inter-Academies Panel report (2006) on Inquiry Based Science Education (IBSE) indicates that "through engaging in the processes of scientific inquiry, students acquire scientific literacy, meaning a general understanding of: the important ideas of science, the nature of scientific investigation and the evaluation and interpretation of evidence." (p. 11) The Panel report indicates that the constructivist view of making meaning supports the claim that IBSE can lead to improvement in scientific literacy.

According to the Inter-Academies Panel report, IBSE programs have two key characteristics:

1. Students develop concepts that enable them to use critical and logical reasoning to make sense of the scientific aspects of the world around them.
2. Students embark on this learning through their own activity, guided and led by teachers who use a range of techniques to explore concepts within the students' own work.

Tytler (2007) also contends that the use of an inquiry approach to teaching and learning has a positive effect on students' attitudes to science which is described by Hackling &

Prain, 2008, as a key component in scientific literacy. It is anticipated that the use of the IBSE approach helps engage students in science and reduces the number of students moving away from secondary school science. The importance of inquiry has recently been recognised by the Australian Curriculum, Assessment and Reporting Authority, with the inclusion of an inquiry strand in the Australian Curriculum (ACARA, 2011).

Driver, Asoko, Leach, Mortimer and Scott (1994) argue that science knowledge is socially constructed and validated. Simply encountering scientific phenomena, or making empirical observations, does not itself enable students to develop scientific ideas and theories. They argue that the development of scientific ideas and principles involves constructing a shared language among a group of people, and that the development of this shared understanding occurs through both personal and social processes. This view of scientific learning is a social constructivist view. In the social constructivist model, students make sense of shared experiences with science observations and phenomena, and then use prior knowledge, past experience and discussion with their peers to construct meaning.

The role of the teacher in a social constructivist paradigm

The role of the teacher in a social constructivist paradigm is to present the students with opportunities to encounter science phenomena and to scaffold their learning, and is vastly different to traditional or empiricist views. Driver et al. (1994) believe that the role of the teacher:

> …..has two important components. The first is to introduce new ideas or cultural tools where necessary and to provide the support and guidance for students to make sense of these for themselves. The other is to listen and diagnose the ways in which the instructional activities are being interpreted to inform further action. (p. 8)

This involves a fundamental shift in the way the teacher is perceived in the classroom. As Shuell (1986) asserts:

> If students are to learn desired outcomes in a reasonably effective manner, then the teacher's fundamental task is to get students to engage in learning activities that are likely to result in their achieving these outcomes……It is helpful to remember that what the student does is actually more important in determining what is learned than what the teacher does. (p. 429)

Essentially, the teacher acts as an interventionist. The teacher works to facilitate group work, argumentation, dialogue and debate. The teacher does not merely present information to the students; rather, s/he guides the students and helps students develop the scientific literacies required at key moments in the investigative process. This notion of the teacher as 'coach' is similar to that presented by Bransford, Brown and Cocking (2000), where the teacher "provides feedback for ways of optimizing performance" (p. 177). At the conclusion of a cycle of activity, teachers encourage the students to engage in reflection to evaluate their own, as well as others' scientific literacy. It is through this cycle of investigation, intervention, evaluation and reflection that scientific literacy is best developed (Inter-Academies Panel, 2006).

Defining Curriculum

The term 'curriculum' is widely used and is referred to in a number of ways by different Researchers. They range from a view that curriculum materials are a list of course content and associated teaching aids (Richmond, 1971; Kesidou & Rosemann, 2002) through to views, as held by this Researcher, that a curriculum moves "beyond mapping out the topics and materials, it specifies the activities, assignments and assessments to be used in achieving its goal" (Wiggins & McTighe, 2001, p. 3). This definition of curriculum is similar to those of Marsh (1996), Print (1993) and Ross (2001). Curriculum materials in this research project will refer to standard physical materials used to frame, plan and implement instruction, but does include assessment pieces used to formally measure student progress.

ACARA, which released the Australian Curriculum in a draft form in March 2010, comments:

> The national curriculum will detail what teachers are expected to teach and students are expected to learn for each year of schooling. The curriculum will describe the knowledge, skills and understanding that students will be expected to develop for each learning area across the years of schooling. This description of curriculum content will result in a curriculum sequence that will represent what is known about the progression of learning in that area.
>
> (ACARA, 2009, p. 4)

The definition presented by ACARA (2009) refers to a curriculum framework which will be used by schools to develop their curriculum materials. However, the curriculum, as it is implemented in the classroom, can often differ from that which was intended in curriculum framework documents.

Intended and implemented curriculum

Other definitions of curriculum, such as those presented by Grundy (1987) and Cornbleth (1990), include the actual delivery of the curriculum materials. They differentiate between the 'intended curriculum' (represented by the curriculum goals, materials and assessments) and the 'implemented curriculum' (the actual teaching and learning occurring in each classroom). A curriculum may have the goal of teaching the importance of controlling variables in an experiment, and have a range of curriculum materials (worksheets, experiments etc.) to support the progress towards this goal (collectively the intended curriculum), but the pace, lesson structure, instruction and classroom climate (the implemented curriculum) can influence how the material is taught.

This distinction between intended and implemented curriculum is the focus of this study. Although a set of curriculum materials can be closely aligned with the ultimate goals of the curriculum, the effectiveness of the curriculum in achieving these goals is largely dependent on its actual mode of implementation in the classroom. In fact, the implemented curriculum can sometimes vary greatly from the intended curriculum (Cornbleth, 1990; Grundy, 1987).

Science Curriculum

The extent to which a science course develops scientific literacy is dependent on several factors. First, the curriculum goals must explicitly state that scientific literacy forms a highly valued portion of the course; second, the curriculum itself must provide opportunities to learn the various aspects that comprise scientific literacy; third, teachers must support students to practise and apply these skills within appropriate learning activities; and fourth, assessments must provide opportunities to measure student progress in developing scientific literacy.

Assessment

Another important aspect of the curriculum is the assessment associated with it. Dochy and McDowell (1997) describe assessment as a tool to determine the rate of progress of a student against both individually negotiated goals and previous performances. This relates well to the definition presented by Wiggins et al. (2001) that assessment involves "the determining of the extent to which the curricular goals are being and have been achieved" (p. 3) i.e. summative assessment. Assessments can also be both diagnostic and formative, and are used to inform teaching and learning. In fact, Hattie (2003) argues that the assessment data is most important when we:

> Move away from considering achievement data as saying something about the student, and start considering achievement data as saying something about their teaching. If students do not know something, or cannot process the information, this should be clues for teacher action, particularly teaching in a different way. (p. 2)

This view of assessment, as being an indication of how teaching must be changed in response to the student data, is supported by Black and Wiliam (1998a), who consider formative assessment to involve four elements:

1. establishing a standard or expected level of performance
2. gathering information on a student's current performance
3. developing a process to compare the two performance levels
4. adjusting teaching to alter, or rather close, that gap (p.4)

Essentially, Hattie (2003), and Black and Wiliam (1998a) argue that formative assessment can, and should, be used as a source of feedback to improve both teaching and learning. In this vein, it is important to note that a single assessment can fulfil a number of purposes. It is possible for an assessment tool to be both summative and formative. For the purposes of this case study of alignment, the analysis addresses the extent to which summative assessments are aligned with goals and learning tasks.

Importance of Curriculum Alignment

Alignment of curriculum can be defined in a number of ways. Tyler (1949) indicates that alignment occurs when the curriculum offered across the grades builds and supports what has already been learnt in earlier years. The current research takes this curriculum alignment a step further by defining it as occurring when "expectations and assessments are in agreement and serve in conjunction with one another to guide the system towards students learning what they are supposed to know" (Webb, 1997, p. 3).

Biggs (1999) emphasises the importance of alignment of assessment with the course objectives. He agrees with Ramsden (1992), who says that assessment is the curriculum as far as the students are concerned. To some extent, the student will learn what is being assessed as much as what is in the curriculum. Biggs (1999) asserts that assessment should be designed in such a way that "if students focus on the assessment, they will be learning what the objectives say they should be learning" (p. 68).

This view of the integral place of assessment in the curriculum alignment is supported by La Marca, Redfield, Winter and Despriet (2000), who contend that the alignment process must consider the assessment of student learning to be the key indicator of alignment. According to La Marca et al., alignment is

> the degree to which assessments yield results that provide accurate information about student performance regarding academic content standards at the desired level of detail to meet the purpose of the alignment system…in a manner that clearly conveys student proficiency as it relates to the content standards. (p. 24)

According to Biggs (1996), alignment of desired outcomes to the selected learning activities and the associated assessment is recognised as a crucial element of good teaching. He emphasises the need for this alignment in curriculum design, which he labels constructive alignment. Biggs (1996) holds a constructivist view, believing that meaning is not imposed or transmitted by the teacher, but rather it is created by the students' learning activities and assessment. Biggs asserts that for students to be meaningfully engaged and bridging a learning gap, the curriculum needs to be focussed on what the students are able to do.

From this focus, the curriculum is then designed so that the desired outcomes, teaching/learning activities and assessment align. For example, if the desired outcome of a period of teaching time is the ability to apply the Law of Reflection to a practical situation, then the curriculum can be tailored to expose the students to the conceptual ideas required for a student to reach that goal.

This constructive alignment in curriculum development incorporates a design process, where first of all the outcomes of the course are identified in terms of what the students should be able to do at the conclusion of the program. This is usually expressed as a series of verb statements about what the student will be able to do as a result of the curriculum. Then, the gap between what the students understand or can do before they undertake the course and what they are expected to be able to do as a result of the course is identified. Once the learning to be undertaken has been identified, the curriculum is designed in such a way to allow students to confront their prior understandings and make adjustments to their skill set based on carefully designed activities. It is important to note that Biggs believes that the students should be *engaged* in the learning activities, implying that the activities need to hold student interest and provide cognitive challenge. The progress of a student through the curriculum is then tracked by using assessment tasks which are strongly aligned with the intended outcomes of the course, providing the teacher with information necessary to adjust the experience in the classroom to better allow the student to reach the intended goals.

Essentially, the constructive design process focuses not on what teachers do, but instead on what outcomes the student will achieve. Using the above example, a constructively designed course will focus on the development of student understandings and skills in the pursuit of the ability to apply the Law of Reflection. Conversely, a course not designed constructively may simply describe a series of activities a teacher can utilise in the teaching of the Light topic. This allows the teacher to move away from a coverage-focused instructional model, where the purpose of the teacher is to deliver a set number of pages from a textbook in a certain time, and be a more responsive tutor or coach for the students. By keeping the focus of the learning in the classroom in keeping with the intended goals of the program, then both teacher and learner are more focused on what needs to occur in the classroom in order for the goals to be achieved.

Biggs (1996) presents constructive alignment in a form which seems complementary to the backwards design process presented by Wiggins and McTighe (2001). Like Biggs, Wiggins and McTighe recognised the increasing prevalence of coverage teaching – teaching in which the aim is simply to get through a certain amount of material in a certain amount of time, with little emphasis on whether a student has actually learnt anything by the time the course is completed. They refer to this style of teaching as "Teach, test and hope for the best" (p. 5). In recognising the limitations of this style of teaching, Wiggins and McTighe developed a style of curriculum design called backwards design. Each step of the backwards design process involves a focusing question:

> What is worthy and requiring of understanding?
> What is evidence of understanding?
> What learning experiences and teaching promote understanding, interest and excellence?
>
> (Wiggins & McTighe, 2001, p.36)

Backwards design and constructive alignment

The starting point in both backwards design and constructive alignment is what the learner should be able to do/know/demonstrate at the conclusion of the course. This approach focuses on the development of the learner, as opposed to the coverage of course content valued by some programs and criticised by others (DeBoer, 1991; Gallagher, 1991; Hodson, 1998).

To adequately map student learning over a period of time, assessment must be aligned with the curriculum (Biggs, 1999; La Marca et al., 2000; Webb, 1997). In most curricula, there is very little alignment between assessment materials and the described curriculum (Chinn & Malhotra, 2002; Germann, Haskins & Auls, 1996; Stern & Ahlgrehn, 2002; Webb, 1997). It is difficult to accurately represent a student's achievement according to the intended goals when the assessment does not align with the course goals. Webb's (1997) analysis shows that teachers are more likely to attend to the stated goals of the course if they are aware that the relevant assessments will directly feature these concepts and skills.

Without proper alignment (Biggs, 1996; Webb, 1997; Wiggins & McTighe, 2001), achieving intended outcomes will be limited because the students would not be learning that which is being assessed. Thus, for any curriculum to be considered effective, it must be analysed for proper alignment of intended outcomes, curriculum and assessment.

Approaches to the Analysis of Curriculum

Analyses of curriculum materials, which represent the intended curriculum, have been conducted in a number of different ways. Beane (1993) used broad methods of analysing content, but the analysis was limited in that it only analysed a small sample of specific curriculum content. In contrast, Kesidou and Roseman (2002) described a method by which the content and implied pedagogy of various types of curriculum materials can be analysed. Research based criteria (see Figure 2) were used to analyse a series of curriculum materials in order to determine whether the curriculum materials were likely to contribute to the attainment of state-mandated benchmarks and standards. This type of content analysis, using experienced judges to score curriculum according to specific criteria has proved quite successful. Its accuracy has been acknowledged in subsequent studies, which used the Kesidou and Roseman model to analyse course materials (Stern & Ahlgrehn, 2002). This model would be suitable in this case study for the analysis of the relationship between the curriculum materials and the goals of the subject, as it has been tested for validity in a large number of situations and provides reliable support materials. Also, the fact that it uses a relatively simple scale of 0 to 3 means that it should have good inter-rater reliability of judgements (Stern & Algrehn, 2002).

Another method of alignment uses the revised Bloom's taxonomy (Anderson, Krathwohl, Airasian, Cruikshank, Mayer, Pintrich, Raths & Wittrock, 2001). This method, presented by Anderson (2002), uses a grid called a taxonomy table whereby the goals, curriculum and assessments are tracked against the four dimensions of knowledge identified by Anderson et al. (2001) (See Figure 2). The case study course assesses only the procedural knowledge of science, thereby eliminating three of the four dimensions from the taxonomy table. Also, as measurement for each of the ELOs is made on a continuum, which includes a sliding scale of cognitive difficulty similar to the cognitive process dimensions, it makes the use of the taxonomy table less

appropriate than other methods. However, as some of the assessments presented in the case study have a greater emphasis on some levels of the taxonomy than others, the addition of the table provides a useful overview of the types of skills demonstrated by the students on different assessments.

Kesidou and Roseman (2002)	**Anderson et al. (2002)**
• Identifying and maintaining a sense of purpose • Taking into account student ideas • Engaging students with relevant phenomena • Developing and using scientific ideas • Promoting student thinking	• Factual Knowledge • Conceptual Knowledge • Procedural Knowledge • Metacognitive Knowledge

Figure 2: *Criteria for methods of scoring alignment of curriculum materials.*

Alignment of assessment

A number of studies have been undertaken to determine whether assessment is aligned with the goals of a particular curriculum. From these studies a large number of alignment methods have been developed, ranging in complexity and usefulness. Bhola, Impara and Buckendahl (2003) classify alignment methods into three broad categories: low, moderate and high complexity models. Low complexity models are simple alignment frameworks which define alignment as "the extent to which the items on a test match relevant content standards" (p. 22). Usually these types of methods use a simple Likert scale to match individual items to particular content strands. Moderate complexity models recognise that alignment is generally defined as more than just a content match, and also examine cognitive complexity such as item difficulty. Finally, the high complexity models consider cognitive complexity, congruence across years, content and a range of other factors (Bhola et al., 2003).

Seven major methodologies of assessing alignment are Project 2061 (Stern & Algrehn, 2002), the Webb analysis (Webb, 1997), Achieve (Rothman, Slattery, Vranek & Resnick, 2002), Surveys of Enacted Curriculum (Porter & Smithson, 2001), the La

Marca method (La Marca et al., 2000) and the methods developed by Germann, Haskins and Aul (1996), and Chinn and Malhotra (2002). Figure 3 presents a summary of these models.

The American Academy for the Advancement of Science developed a moderate complexity program called Project 2061, whose goal was to analyse science materials for the depth of science content and skill provision. Stern and Ahlgrehn's (2002) investigation of Project 2061's method for determining alignment analysed a range of assessment materials for their alignment and validity according to three distinct criteria: alignment to curriculum goals, testing for understanding and informing instruction. The content analysis used a large variety of criteria, particularly focusing on test-based materials such as textbook quizzes and commercial term papers, and suggested methods to improve the alignment of assessment tasks.

Webb (1997) produced a high complexity process for determining the validity of tasks, irrespective of their content. The analysis was based on five main criteria: content focus; articulation across grades and ages; equity and fairness; pedagogical implications; and system applicability. A content analysis of a range of different assessment materials was made based on this framework. Of particular interest in the study were the "high stakes" (Broadfoot, 1996) national and state testing programs instituted in the United States of America. Members of a trained national committee scored the most commonly used textbooks, assessment instruments and curriculum guides available for the Science and Mathematics standards, using the Webb analysis criteria. This study found that many assessment programs used to assess state and national standards did not reflect the emphases present in the curriculum materials or coherently reflect the curriculum goals of the American national science curriculum. In a review of the Webb analysis program, Webb (2007) and Martone and Sireci (2009) both noted the process requires significant and sustained reviewer training at the beginning of the process and identified that averaging reviewer ratings across standards and objectives might mask differences and inflate degrees of alignment. Martone and Sireci (2009) also noted that the advantages of the Webb analysis are its clear guidelines as to the acceptable standard of alignment and the provision of a measure of alignment to ultimately "illustrate the relationship between what is being asked of students, how this is being assessed, and what trade-offs are being made in the process." (p. 1342)

Stern and Ahlgrehn (2002)	Webb (1997)	La Marca et al (2000)	Achieve (2002)	SEC (2001)	Chinn and Malhotra (2002)
• alignment to curriculum goals • testing for understanding • informing instruction.	• Content focus • Articulation across grades and ages • Equity and fairness • Pedagogical implications • System applicability	• Content match • Depth match • Emphasis • Performance match • Accessibility	• Performance centrality • Cognitive demand • Level of challenge • Balance of items • Item fit analysis	• Topic coding of items, standards and instructional content • Expectations of student performance • Cognitive levels	• Generating research questions • Designing Studies: select variable(s) • Designing Studies: planning procedures • Designing Studies: controlling variables • Designing Studies: planning measures • Making Observations • Explaining Results: transforming observations • Explaining Results: finding flaws • Explaining Results: indirect reasoning • Explaining Results: generalisation • Explaining Results: types of reasoning • Developing Theories: level of theory • Developing Theories: co-ordinating results from multiple studies • Studying research reports

Figure 3: *Criteria used in six studies for scoring alignment of assessment.*

A model which draws heavily on the Webb methodology is that developed by La Marca et al. (2000). The La Marca model is designed to align assessment systems to state standards, specifically those relating to the requirements of Title I Education Act legislation (United States Department of Education, 1999). The model uses five dimensions: content match, depth match, emphasis, performance match and accessibility, which are very similar to those used by Webb. The limited range of application of this model (as it is designed to be used for very specific curricula) means it is less useful than the original Webb analysis for this particular study.

The Achieve methodology described by Rothman et al. (2002) involves a judgement of the alignment of both overall assessment tasks and individual test items. It takes a slightly different form depending on the subject area, whether English, Mathematics or Science, and differs from the Webb and Project 2061 methods by disaggregating the results of the subject matter experts reviewing the items. The high complexity Achieve protocol is applied in two stages. The first is to analyse a test item by item, comparing each item to the intended learning outcome it is designed to assess, and then considering the group of items as a whole. The approach considers the assessments in terms of the balance of test items relative to the intended outcomes, sources and levels of challenge, as well as comparisons between assessments in terms of cognitive demand. Unlike the Webb method, Achieve does not give clear criteria for when items or assessments have achieved alignment, but gives more qualitative information about the coding and the possible changes which could be made as a result of the analysis (Martone & Sireci, 2009).

Porter and Smithson (2001) developed the moderate complexity Surveys of Enacted Curriculum (SEC) method of alignment determination. There are three main alignment dimensions in the SEC methodology: content match, expectations for student performance and instructional content. Subject matter experts were used in 11 states and four districts to determine the level of alignment of standards, assessments and the focus of instruction. The major difference between the SEC methodology and other alignment methods is the ability of the SEC to determine the alignment of both the intended and the enacted curriculum. This is achieved through a short period of observation of actual teaching practice, in which the SEC is used to determine the extent to which the observed instruction matches the intended outcomes and assessments (Blank, Porter & Smithson, 2001). In their review of the SEC methods, Martone and

Sireci (2009) indicate that the method, while extremely useful in the observation of the enacted curriculum, does not provide the depth of information of either the Webb or the Achieve protocols.

Finally, two other studies, Germann, Haskins and Aul (1996), and Chinn and Malhotra (2002), examined the alignment of specific science programs and the assessments used to assess student progress using low complexity alignment models. Both studies emphasised the epistemological basis of science, using the qualitative criteria listed in Figure 3 to determine the level of alignment between the assessment and the nature of 'real world' scientific inquiry. The Germann et al. (1996) study used five criteria for content analysis, which were expanded upon by the later Chinn and Malhotra (2002) study to 14 separate features. Chinn and Malhotra define real world scientific inquiry as "the processes employed in real scientific inquiry" (p.18), as they contend that "inquiry tasks commonly used in schools evoke reasoning processes that are qualitatively different from the processes employed in real scientific inquiry" (p. 175). The criteria that the study used related to the specific steps used in the generation of a scientific investigation. The Chinn and Malhotra study examined 50 tasks and scored them on whether they contained features that were deemed necessary to be an authentic assessment to be used to enhance scientific literacy of students. The scoring elements of these programs are shown in Figure 3.

It is important to consider the fact that none of the models of assessment alignment presented above were evaluated for their effectiveness when used *in situ*; each study relied on external subject matter experts to review materials produced either commercially or from a particular district in response to mandated curriculum outcomes. As the process in this study will examine the use of an alignment framework within a school by members of academic staff, the ease of use of the criteria must come into consideration when selecting an appropriate framework.

Whatever alignment process is used, it is important that it provides a measure of how well the intended outcomes of the course are represented in the curriculum materials and the assessment program. The studies provide data which can be used to guide changes to elements of the curriculum materials and assessment to ensure that they more accurately reflect the purposes of the curriculum.

Conceptual Framework

A conceptual framework is a group of concepts that are broadly defined and systematically organized to provide a focus, a rationale, and a tool for the integration and interpretation of information (Bell, 2005). In this particular case, the conceptual framework brings together the concepts of curriculum (particularly in science), scientific literacy and theories of alignment. The conceptual framework for this study is illustrated in Figure 4.

The shaded section of Figure 4 shows the curriculum as defined in this case study. The curriculum is comprised of four discrete yet related components. The intended outcomes of the curriculum are addressed through teaching and learning activities whose effectiveness is measured by the assessment program. Curriculum materials support the implementation of all three facets of the curriculum.

Quality curriculum materials, such as examinations and worksheets, that are carefully aligned to goals and assessments are critically important for effective teaching and learning. By analysing the alignment of the documented curriculum with the intended goals of the program, an indication of the alignment of the intended curriculum can be gained.

The importance of the development of students' scientific literacy is emphasised in the literature. Although the school in this case study does not label assessed skills explicitly as scientific literacy, the curriculum's stated goals match well with the scientific literacy definitions presented by the American National Research Council (1996), Hackling et al. (2001), Hackling and Prain (2008) and the National Curriculum Board (2008). This emphasis on the development of scientific literacy informs the intended outcomes of the curriculum.

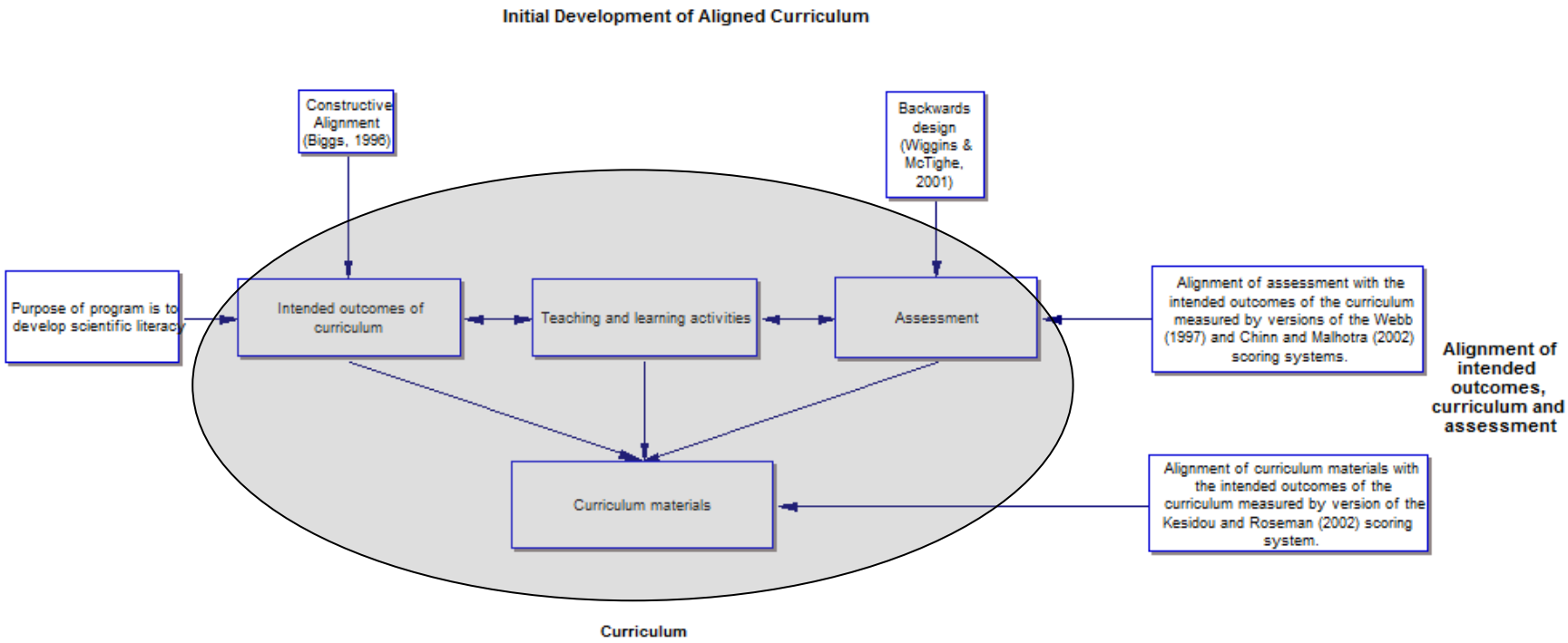

Figure 4: *Conceptual framework for alignment of middle school science curriculum*

The importance of a curriculum addressing both process skills and the relationship to authentic science inquiry for advancing student understanding was addressed in the literature (Carey et al., 1989; Chinn & Malhotra, 2002). The content analysis in this case study should therefore include criteria to analyse the relationship of the curriculum materials to both authentic science tasks and contexts of science. Several of the authors proposed methods by which a content analysis could be performed on curriculum materials. Some were too limited to be useful in the case study (Beane, 1993; Chiapetta, Sethna & Fillman, 1993; Eltinge & Roberts, 1993; Jiminez, 1994), while others contained strands and dimensions which were either inappropriate or redundant, such as the Factual, Conceptual and Metacognitive elements of the Taxonomy presented by Anderson et al. (2002). The method used by Kesidou and Roseman (2002) in Project 2061 are more appropriate and have been extensively tested in other studies and has instructional material to support their implementation. Thus, this method for content analysis seems most appropriate for the case study.

The importance of alignment of assessment with curriculum goals was emphasised by both Wiggins and McTighe (2001) and Biggs (1999). Figure 4 shows that the Biggs' constructive alignment starts with a consideration of the intended outcomes of the curriculum, whereas the backwards design process developed by Wiggins and McTighe focuses on the assessment or what will demonstrated as a competent response at the conclusion of the course. It is important that the alignment of the assessment is examined in some depth, and hence the limited scope of the analysis proposed by Stern and Ahlgrehn (2002) and Germann et al. (1996) will not provide the rigour required in this case study.

Two different methods will be used to examine the alignment of assessment in this study. The method proposed by Webb (1997), with modification, will be used to determine the level of alignment of the assessment materials to the curriculum goals. As the purpose of this study is to examine whether the documented curriculum is aligned with the intended outcomes, three criteria have been removed. These criteria relate to actual instruction, use of technology, and equity and fairness, all aspects which do not relate to the development of scientific literacy. Also, the criterion related to the sustainability of the program has been removed, as the curriculum has been in place for almost six years.

For an examination of the key features of authentic, or real world, scientific inquiry, the techniques described by Chinn and Malhotra (2002) seem to be most appropriate, as they provide a comprehensive list of the features of assessments strongly linked to authentic scientific inquiry. This framework will be cross-referenced against the cognitive process domains of the revised taxonomy table (Krathwohl et al., 2002). The taxonomy table was used to ensure that each course gives the students an opportunity to display the more complex cognitive processes.

Summary

Chapter two was divided into several distinct sections. The first section discussed the purpose of science education, differentiating between curriculum design favouring the transmission of a variety of scientific concepts in modules or topics and the development of scientific literacy. This section defined scientific literacy and highlighted the importance of processes, science literacies and epistemological beliefs in science. Finally, it discussed pedagogical approaches to science education, indicating that the social constructivist approach was most effective in developing scientific literacy. The second section briefly described current assessment practices in science.

The next section discussed the design of science curriculum, emphasising the difference between the intended and implemented curriculum, and ways in which the two could be quite different.

The importance of curriculum alignment was outlined in section three, particularly in the area of science education. A variety of complementary models, including backwards design and constructive alignment, were introduced, as well as a description of the benefits of curriculum alignment. This section identified that without alignment, student achievement of the intended curriculum outcomes will be limited, and then reported on the variety of approaches to the analysis of curriculum and the final section provided a conclusion and briefly summarises the literature findings.

The last section considered the role of assessment, curriculum materials and intended outcomes in student learning, and what methods could be used to develop alignment. From these ideas, a conceptual framework was developed to describe this study. The

framework also considers how backwards design and constructive alignment fit into the development of a coherent curriculum framework.

The next chapter discusses the methodology of this research, including its design, the instruments and materials used, ethical considerations, target population and analysis of collected data.

CHAPTER THREE: METHODOLOGY

Introduction

The research methodology is discussed in this chapter. The design and nature of the research is discussed in section one, and section two describes the context of the case and subject population. Section three describes the procedure by which the project was carried out and the data gathering tools that were employed; the basic analytical procedures are described in section four. The fifth section addresses the limitations of the research design. The ethical considerations pertinent to this research are discussed in the final section.

Research Questions

The purpose of this particular case study is to investigate two questions:

1) To what extent are the intended outcomes, curriculum and assessment in this Middle School science curriculum constructively aligned?

2) How effective is the curriculum evaluation model developed and implemented in this study for evaluating the alignment of intended outcomes, curriculum materials and assessment?

Approach

Research methodology usually falls within two broad paradigms: qualitative and quantitative approaches. Quantitative methods involve the development of a measurement system to quantify relationships in order to prove or disprove a hypothesis. In quantitative research, statistics are used in order to make sense of the data in terms of the research question. Typically, quantitative research lends itself to highly valid and highly reliable research. However, not all research questions can be suitably answered by using quantitative methods, particularly when data are non-numerical, sample size is small, or variables are difficult to isolate (Bell, 2005).

Qualitative research involves the examination and analysis of phenomena in order to discover meanings and patterns in relationships without using mathematical models. Qualitative methods include ethnographic, action research and grounded theory approaches and often involve the compilation of case studies (Bell, 2005).

The research approach in this investigation is a case study utilising mainly qualitative methods. The case study research method is an empirical inquiry approach which investigates a situation within its real-life context (Yinn, 1984). A form of qualitative descriptive research, the case study examines intensely an individual or small participant pool, drawing conclusions only about that participant or group and only in that specific context (Bell, 2005). The case study approach, utilising qualitative methods such as content analysis, is most appropriate for this study as many of the materials that will be examined are specific to the context of the case.

This case study includes a content analysis of curriculum documentation, which Krippendorf (1980) describes as "a research technique for making replicable and valid inferences from data to their context" (p. 21). The content analysis, sometimes known as a document analysis, will investigate the frequency with which particular terms and concepts appear in the curriculum materials. This analysis enables the materials of the intended curriculum to be examined for their alignment with the intended goals. Identification of the alignment of the intended curriculum, as analysed using the models of Kesidou and Roseman (2002), Webb (1997) and Chinn and Malhotra (2002) can be achieved using a document analysis approach. Finally, interviews conducted with the participants in this study were used to help determine the effectiveness of the curriculum evaluation model. It was decided that a semi-structured interview approach was the most appropriate, as there were key questions that needed to be considered to answer the research questions, yet the scope of the project meant that there may have been issues or thoughts that arose from the process that were not initially predicted by the Researcher (Bell, 2005).

This case study also utilised quantitative elements, as it used a scoring system to rate the alignment between curriculum goals and the assessment and instruction. The ultimate aim of this research project was to use the quantitative methods to give a precise and testable expression to qualitative ideas presented in the case analysis. The

complementary nature of the qualitative and quantitative methods provides opportunities for triangulation of data. Hence the study could be described as a mixed methods investigation.

Unfortunately, however, the rating given to the alignment of the curriculum with its goals is problematic, and represents a limitation of the study method employed in this study because it was based only on curriculum resources i.e. the intended curriculum. To measure the alignment of the implemented curriculum, it would be necessary to gain an insight into what actually occurs within each classroom. This could be achieved by either classroom observation or by interviewing teachers who implement the curriculum. To keep this research project manageable, the alignment analysis is restricted to just the intended curriculum.

Context of the Case

The curriculum examined in this study has been developed by a regional independent K-12 school with approximately 1350 male and female students. The MySchool ICSEA value is 1150, rating it as an advantaged school. The MySchool website entry (http://www.myschool.edu.au) for the school states:

> At [the case study school], we value learning as the key attribute of developed individuals and communities. We help students discover who they are, who they want to be and how to get there. In order for students to make optimum progress, the most important resource is the quality of teaching. [The case study school] is committed to continuous improvement in teaching practice. In order to deliver on this commitment, significant resources are allocated to both maintaining a high standard of practice and to the identification and implementation of teaching approaches proven to be the most effective, as evidenced by student-learning outcomes. The professional learning program concentrates on instruction and student outcomes, and provides opportunities for inquiry, collaboration, feedback and connections to external expertise and research. For the seventh successive year [this school] has been…ranked in the top ten schools across the state. Given our open-entry policy, this is an exceptional achievement. In 2009 the VCE results were extremely pleasing and reflect the wonderful work carried out by the students and staff throughout the year, and in the years leading up to Year 12 - 8% (10 students) achieved an ENTER over 99; 23% (32 students) over 95; 42% (55 students) over 90; 32% of study scores over 40. Of the 138 students completing Year 12, 98% of the cohort was accepted

> into tertiary institutions of their choice. Throughout a student's time at the school our focus is to maximise their competence, skills and capacity, so that, at the end of their time at the school, when they stand at the threshold of their future, they can choose their "heart's desire". This is achieved through learning about teamwork from participation in the co-curricular program, which includes extensive competitive sporting opportunities, performing arts ensembles and theatre productions, and involvement in local and overseas service activities. [The case study school] is a co-educational day and boarding school, enrolling students from Early Learning to Year 12… As a Uniting Church school, engagement with values-thinking and personal ethics is encouraged through attendance at Chapel and regular time is spent with Learning Mentors and House Teachers. However, it is by bringing rigor to the development of curriculum and the implementation of its teaching and assessment that students' future pathways are established.

The 2010 NAPLAN results showed that, in the 20 areas tested, the case study school ranks at or above the similar schools in all areas, as shown in Figure 5 below.

	Reading	Writing	Spelling	Numeracy
Year 3	**Slightly above** similar schools	**Slightly above** similar schools	**Slightly above** similar schools	**Above** similar schools
Year 5	**Above** similar schools	**Above** similar schools	**Slightly above** similar schools	**Significantly above** similar schools
Year 7	**Above** similar schools	**Slightly above** similar schools	**Slightly above** similar schools	**Above** similar schools
Year 9	**Above** similar schools	**Above** similar schools	**Above** similar schools	**Above** similar schools

Figure 5: *Comparison of the case study school to similar schools based on 2010 NAPLAN results.*

In the two previous years, 2008 and 2009, the school scored equal to or above similar schools in all forty areas, with all but four areas scoring above the results of schools with similar ICSEA values.

Procedure

This section outlines the methods used to collect, analyse and interpret the data. It also indicates the parties involved and the specific frameworks used to assess the data collected. The investigation was conducted in six phases:

Phase One: Two year levels in which a particular group of scientific concepts is taught in sequence were selected, and the intended outcomes, curriculum and assessment for these programs described.

The purpose of this study is to analyse a section of curriculum in some depth to determine the extent of alignment with the Essential Learning Outcomes it is designed to address. As it is impractical to analyse the entire five year course in depth, a selection has been made of two semester-long courses at two different year levels.

The courses selected are the Year 7 (12-13 years) and the Year 9 (14-15 years) courses, both of which use chemical concepts such as atomic structure, changes in state, chemical reactions and rates of reaction as contexts to help develop student achievement of the ELOs. These courses were selected because there is continuity not only in the goals of the curriculum, but also in the conceptual contexts that are being studied. To analyse the consistency of contextual information across year levels, described by Webb (1997) as categorical concurrence, it is necessary to have similar contexts in the two courses.

Phase Two: Participant reviewers were recruited and trained to ensure consistency in the scoring of curriculum materials and assessment.

The instructional materials were scored by three reviewers, each of whom was employed by the school in question. The reviewers were asked to participate after given an overview of the study. Each participant has significant experience in the teaching of the contextual areas over a number of years, and brings expertise to the scoring of the materials.

Although all of the reviewers have experience in teaching science skills, training in the scoring system comprised two distinct sessions. In the first, the system of scoring was

introduced, and journal articles related to the scoring system distributed in order to help the reviewers understand the basis of the system. The reviewers scored and cross-marked several carefully selected pieces of assessment and learning activities over the course of four weeks at both the case study school and the homes of the researchers. During this process, the reviewers discussed and refined their understanding of each of the criteria using both the Researcher and the related literature. Discussion of the variance in the ratings helped improve the consistency in understanding and interpretation of the scoring rubrics.

Phase Three: An alignment analysis of each of the two courses was performed.

The study used content analysis to determine whether the curriculum materials aligned with the intended outcomes of the course as described by the ELOs. The content analysis was based on an adaption of the framework presented by Kesidou and Roseman (2002).

After training, the reviewers indicated that they had a clear picture of what the intended outcomes of the course are and what alignment looked like, and also had the opportunity during the course of the review process to collaborate with each other to develop consistency in their judgements and ratings. These review sessions were recorded and documented as part of the process.

After the curriculum materials had been analysed for alignment with the ELOs, an analysis of the alignment of the assessments was conducted using the Webb (1997) framework. The purpose of this step was to determine which of the assessments truly aligned with the stated goals of the assessment, and whether they validly assessed student performance.

The assessments were further analysed for their authenticity by the same three reviewers. Each assessment was scored according to the number of features present which, according to Chinn and Malhotra (2002), are essential for the assessment task to be considered authentic. In addition, each of the assessments were checked against the cognitive process dimension identified by Anderson et al. (2001) and then utilised by Krathwohl (2002) in his taxonomy table.

Phase Four: Analyse participant ratings.

All of the rating scores were collated on a common spreadsheet. This ensured a consistent approach, and that all materials were scored on the appropriate criteria. The data were then converted into a range of graphic and statistical displays and two sets of analyses were conducted.

Firstly, each of the ratings for the assessments and curriculum materials produced by the reviewers were entered into an Excel spreadsheet, and the average of the ratings recorded. The ratings themselves gave an indication of the degree of alignment for each criterion. To determine the degree of alignment with the intended conceptual goals of the curriculum, the curriculum materials and assessments need to achieve a mean score of at least 2.0 on each of the categories scored. This value indicates an acceptable level of alignment (Kesidou & Roseman, 2002).

Any of the materials which failed to reach the mean score of 2.0 were noted, and these materials discussed in the exit interviews with each of the reviewers.

Secondly, the data were tested for inter-rater reliability. There are several methods that can be used to determine inter-rater reliability, but the method most appropriate for this particular study is Fleiss' kappa co-efficient. Fleiss' kappa expresses the extent to which the agreement between raters on a particular nominal criterion exceeds that which would be expected through pure chance (Fleiss, 1971). It is related to the Cohen's kappa measurement, but has the advantage of being able to measure the level of agreement between more than two raters, which is particularly pertinent to this study. A kappa of 0.61 indicates that the agreement of the raters is significantly different to that expected by chance, and indicates "an acceptable level of inter-rater reliability" (Fleiss, 1971, p. 277).

Phase Five: Interview participants.

Each of the reviewers participated in an exit interview that was transcribed and used as qualitative data to address the research questions. The interview consisted of a number of questions (see Figure 6) relating to the application and effectiveness of the curriculum evaluation model.

Do the instruments provide meaningful data?
Could the data provided by these instruments allow the realignment of curriculum materials, assessment and/or instruction?
Were there any criteria in any of the instruments that were unclear or extraneous?
What changes would you recommend to either the process or the instruments to improve the ease of use of the scoring instruments?
How much time has been spent, in total, scoring the curriculum materials?
Which scoring instrument was most time efficient? (i.e. Which instrument provided meaningful data within a reasonable amount of time?)
Would scoring curriculum materials using these instruments be practical in a school setting?
What changes would you recommend to either the process or the instruments to improve the reliability or quality of the data collected?

Figure 6: *Interview questions.*

Phase Six: Determine the effectiveness of the method of analysing alignment.

Two factors were taken into consideration in evaluating alignment:

1. The amount of time required to review the curriculum of a program.

   The reviewers recorded the total amount of time spent using each section of the alignment tools. This information was then used in the interviews, along with the direct questioning of the reviewers, to determine whether the alignment methods were time efficient.

2. The applicability, reliability and ease of use of each criterion.

   The reviewers were then asked to comment on each of the criteria in terms of the clarity, ease of use and applicability of each criterion during the semi-structured interviews. Reviewers were asked to comment specifically upon the criteria which showed low Fleiss' kappa co-efficients.

The last factor is particularly important. If a criterion was either poorly matched to the alignment process or had a large degree of discrepancy in the reviewer scoring, it

indicates that there is a need to either revise the criterion to make it more appropriate for the analysis (validity) or to enhance consistency of interpretation by the judges (reliability).

Assumptions

Two key assumptions underlie this study:

The training and discussion sessions conducted in the use of the scoring scaffolds promotes a strong understanding of the scoring criteria.

The application and use of assessment and course materials, as documented in the intended curriculum, is understood by participants.

Instruments

Three different instruments were used to determine the extent to which the intended outcomes (ELOs), curriculum and assessments align in the Year 7 and Year 9 courses. Each of the instruments is described below.

Alignment of curriculum materials with intended outcomes

The content analysis examined the documented curriculum materials for each year level to determine the extent of alignment of these materials with the intended outcomes. It utilised a framework similar to that presented by Kesidou and Roseman (2002). Each set of curriculum materials were reviewed against the seven criteria in Figure 7 below:

Criteria	Score					
	0 Non-existent	1 Poor/minimal detail	1.5 Fair/covered in little detail/lacking quality	2 Satisfactory/ adequate coverage	2.5 Very good/ explicit instruction	3 Excellent/ explicit, differentiated instruction
Are the ELOs of the intended curriculum addressed?						
What is the extent of curriculum materials supporting the ELOs?						
Is there an identification and maintenance of a sense of purpose towards the intended learning goals?						
Do the curriculum materials take into account student ideas on scientific literacy?						
Does the intended curriculum engage students with the ELOs?						
Does the intended curriculum develop and use scientific literacy?						
Does the intended curriculum promote student thinking about science literacy?						

Figure 7: *Scoring table for determination of alignment of curriculum materials with intended goals (Adapted from Kesidou & Roseman*, 2002).

Each of the three reviewers scored the curriculum materials on the three point scale shown in Figure 7. Using this scoring system, Kesidou and Roseman (2002) indicate that an average of at least 2.0 for each criterion is required for confirmation of satisfactory alignment of curriculum materials with intended outcomes.

Alignment of assessments with intended outcomes

The content analysis examined the assessment tasks for each year level to determine the extent of alignment with the intended outcomes of the course. It utilised the alignment framework proposed by Webb (1997). However, several of the criteria originally included in the Webb analysis (Actual Instruction, Use of Technology, Equity and Fairness, and System Applicability) were removed as they do not relate specifically to alignment of documented curriculum.

Each of the three participants scored the assessments according to the criteria outlined in Figure 8:

Criteria	Score				
	0 – 1 Insufficient	1.5 Only for the program as a whole	2 Acceptable	2.5 Only for the program as a whole	3 Full
Categorical concurrence					
Depth of knowledge consistency					
Range of knowledge tested					
Balance of representation					
Cumulative growth in content knowledge					

Figure 8: *Scoring table for determination of alignment of assessment with intended goals* (Adapted from Webb, 1997)

Again, the scoring system used a three point scale, with a brief description of the score for each criterion helping to improve reliability of the scoring. The descriptors of the criteria are provided in Appendix E. A score of 2 (adequate alignment) indicates that there is a reasonable level of agreement of assessments to the outcomes, yet there would still be room for improvement. The reviewers scored each individual assessment task and then the entire assessment program. The individual assessment tasks were scored using the three point scale as presented in Appendix E. However, when considering an assessment program in its entirety, an overall score may fall in between categories. Hence, two extra levels of differentiation were added, which are the 0.5 and 1.5 scores.

Alignment of assessment with epistemological and cognitive goals

The content analysis examined the documented assessment tasks for each year level to determine the extent of alignment of these assessments with the epistemological and cognitive goals of the course. It utilised the alignment framework proposed by Chinn and Malhotra (2002) combined with the conceptual framework presented by Krathwohl (2002). This determined the degree to which the course attempts to influence the students' beliefs of the nature and purpose of scientific inquiry. For the course to properly address this epistemology, it must feature each of the steps required in an authentic scientific inquiry.

First, the assessments were checked against the features of authentic (real world) science. For the assessments to be aligned with these goals, each of the goals must be checked at least once against the assessment for that course. Second, the science content of the assessment was assessed for its cognitive demands using the conceptual framework of Krathwohl (2002). Figure 9 on the following page was used to score the features.

		The Cognitive Process Dimension					
		1. Remember	2. Understand	3. Apply	4. Analyse	5. Evaluate	6. Create
Epistemological Goals	Generating research questions						
	Designing studies: select variable						
	Designing studies: planning procedures						
	Designing studies: controlling variables						
	Designing studies: planning measures						
	Making observations						
	Explaining results: transforming observations						
	Explaining results: finding flaws						
	Explaining results: indirect reasoning						
	Explaining results: generalisation						
	Explaining results: types of reasoning						
	Developing theories: level of theory						
	Developing theories: co-ordinating results from multiple studies						
	Studying research reports						

Figure 9: *Scoring table for the determination of alignment of assessment with epistemological goals* (Adapted from Chinn & Malhotra, 2002; Krathwohl, 2002)

Limitations of the Research Design

Three limitations have been identified in the current study. First, the issue of reviewer numbers needs to be discussed. Most research, which deals in some way with human subjects (be it qualitative or quantitative), will produce results more representative of the target population, when larger numbers of respondents are utilised. When dealing with the subject of this study, reviewers needed to be familiar enough with the science program so that they wouldn't require extra coaching, yet not so involved with the creation of the courses that they would be emotionally bound to the materials. This creates an inevitable tension; context does make a difference in research, and it would have been interesting to enlist reviewers who had no dealings with the course materials at all before their work in the study. However, with limitations in terms of time and resources, it was decided that reviewers from the case study school would be able to rate the curriculum materials and the process. Consequently, the numbers of reviewers was limited to three.

One of the central requirements in order for research findings to be considered reliable is that similar results could be expected to be replicated either in the same population at some later stage, or in other similar cohorts (Stringer & Dwyer, 2005; Weirsma & Jurs, 2004). Given that the curriculum and assessment materials produced by a school are so based in context, this may alter the effectiveness of the alignment instruments when applied to other contexts (schools).

The third potential limitation stems from the issue of validity. Burns (2000, p. 390) noted that qualitative research can suffer from validity problems, meaning that there exists the possibility that this research will not actually measure what it is supposed to measure. However, Maxwell (1992) contended that other Researchers have sought to redefine the construct of validity in terms that are more relevant to qualitative research, and have identified four different types of validity that could apply to this study:

- *Descriptive validity:* The extent to which there would be agreement between different observers, regarding the information elicited from respondents.
- *Interpretive validity:* The extent to which the descriptions of elicited information truly reflect the meaning of what respondents were trying to communicate.
- *Theoretical validity:* The extent to which the information successfully addresses the theoretical constructs the Researcher brings to the study.

- *Validity of generalisations:* This refers to the extent to which the account(s) can be extended to the rest of the target population.

Despite the best efforts to provide marking rubrics, consultation time and training, the rating of alignment of curriculum materials and assessments, as well as the effectiveness of an alignment program, is highly subjective in nature. What one individual perceives as alignment may not be seen the same way as other reviewers, or indeed the Researcher. Each interviewee was made explicitly aware of the conceptual framework of the research and was asked, as much as possible, to frame their responses within the bounds of these constructs.

Triangulation was also used to narrow the chances of invalid data being used as evidence in the subsequent findings of the research. Cresswell (2005) defines triangulation as "…the process of corroborating evidence from different individuals…types of data… or methods of data collection…in descriptions and themes in qualitative research" (p. 352). This process was used when ensuring that each reviewer was comfortable with their responses, and is similar to the process known as member checking, which strengthens validity of findings through ensuring one or more reviewers physically check the accuracy of their accounts (Cresswell, 2005).

Ethical Considerations

Prior to the commencement of the data-gathering phase, ethics approval was granted by the Edith Cowan University Human Research Ethics Committee, which is mandatory under University policy when dealing with research issues involving human subjects. Reviewers' anonymity was ensured by using only a coded number system (R1-R3). All other identifying information was removed from final transcripts. It was also ensured that participants felt no obligation to continue participating in this research, should they decide for whatever reason, to withdraw. This was made clear to each participant both in writing, via a standard consent letter that was required to be signed, and then verbally at the beginning of each interview (see Appendices F and G).

Summary

Chapter three provided an overview of the methodology used in the research. The first section discussed the nature and design of the research, indicating that the study was qualitative in nature, was couched in a case study design, and employed a scoring system and semi-structured interviews as its main data collection tools. The next section described the instruments that were used. Section three discussed perceived weaknesses of the research and identified ways that these weaknesses were minimised as far as practically possible. The main ethical considerations for this study were outlined in the final section.

The next chapter begins discussing in detail the findings of the current research by examining the degree to which the curriculum materials and assessments are aligned to the intended outcomes of the case study Middle Years' science course.

CHAPTER FOUR: FINDINGS – CONSTRUCTIVE ALIGNMENT OF THE INTENDED OUTCOMES, CURRICULUM AND ASSESSMENT IN THE MIDDLE SCHOOL SCIENCE CURRICULUM

Introduction

This chapter is divided into six sections, with the first section reiterating the aims and objectives of the present research. The second section reviews the scoring instruments implemented to review the curriculum, while the third section examines the alignment of the curriculum materials with the intended goals of the science program. The fourth section addresses the degree to which assessment is constructively aligned as indicated by the scoring data. The overall impressions of the assessment programs in Years 7 and 9 are examined in section five while the final section highlights the features of the science program that most adequately enables alignment.

The purpose of this research was to develop a curriculum evaluation model that would effectively assess the alignment of the intended outcomes, curriculum and assessment in the Middle Years science program.

Specifically the research project focused on two questions:

1) To what extent are the intended outcomes, curriculum and assessment in this Middle School science curriculum constructively aligned?

2) How effective is the curriculum evaluation model developed and implemented in this study for evaluating the alignment of intended outcomes, curriculum materials and assessment?

Three main scoring instruments were used in this study, each of which dealt with a different facet of the Middle Years science program. As discussed in Chapter Two, any education program consists of three main parts: the intended outcomes of the course; the curriculum materials designed to support attainment of these outcomes; and the assessment materials used to evaluate progress of learners towards the intended outcomes.

The scoring instruments used were designed to determine the extent of the constructive alignment (Biggs, 1996) of these elements of the program. Each of the scoring instruments has been adapted from those published by previous research. They were selected because the scoring criteria were well-elaborated with demonstrated validity, and have been tested with a range of materials previous to being used in conjunction with one another in this study.

Alignment of Intended Goals with Curriculum Materials

The scoring system used to determine the alignment of the intended goals and the curriculum materials was adapted from work by Kesidou and Rosemann (2002). Each set of curriculum materials was scored on a scale ranging from 0 to 3. Using this scoring system, Kesidou and Roseman (2002) indicate that an average of at least 2.0 for each criterion is required for confirmation of acceptable alignment of curriculum materials with the science program's intended outcomes.

After two professional learning sessions, in which the reviewers were trained on the use of the scoring system, the Year 7 and Year 9 materials were scored independently. There were opportunities for the reviewers to discuss their interpretation of the scoring criteria during the scoring process. Table 1 shows the reviewers' mean scores for each criterion, as well as the mean rating for the set of materials.

Table 1: *Alignment scores of Year 7 and Year 9 curriculum materials.*

Criteria	Score			
	Year 7		Year 9	
	Mean Score (/3)	Standard Deviation	Mean Score (/3)	Standard Deviation
Are the ELOs of the intended curriculum addressed?	2.5	0	2.3	0.24
What is the extent of curriculum materials supporting the ELOs?	1.8	0.24	1.7	0.24
Is there an identification and maintenance of a sense of purpose towards the intended learning goals?	2.5	0	2	0
Do the curriculum materials take into account student ideas on scientific literacy?	3	0	1.8	0.47
Does the intended curriculum engage students with the ELOs?	2.3	0.24	2	0.41
Does the intended curriculum develop and use scientific literacy?	2.5	0	2	0
Does the intended curriculum promote student thinking about science literacy?	3.0	0	2	0

The Year 7 curriculum materials, on the whole, show constructive alignment with the intended outcomes of the science program according to the criteria outlined by Kesidou and Roseman (2002). Individual reviewer scores are featured in Appendix G. All but one of the criteria (Criterion 2) showed a mean score greater than 2, with only the criterion investigating the extent of curriculum materials supporting the ELOs falling short of alignment. Discussions with the reviewers indicated that the curriculum materials, although generally showing a strong alignment to the intended goals, they were actually "quite limited in number" [R2]. Although teachers are required to deliver instruction designed to improve students' skills in scientific literacy, the amount and depth of material was not sufficient for the intended goals to be achieved without the construction of additional materials by the teacher. Different teachers at the case study school took responsibility for developing materials for particular sections of the course. The variation in the quality of materials from one section of the course to the next indicated that the ability of teachers to independently construct high quality and focused materials varied significantly. The reviewers recognized that this lack of adequate materials could limit the extent to which the course achieves its goals consistently

between classrooms, and without adequate resourcing the quality of the overall course may suffer in some classrooms.

One of the categories showed strong alignment: the promotion of student thinking about scientific literacy. All of the curriculum materials scored had deliberate attention paid to one or more aspects of the science literacy continua, both through their content and the formatting structure which brought attention to the ELOs on every material. The reviewers indicated that the most effective of the materials were "tightly linked to the ELOs and students would have no doubt as to what the aim of the activity was." [R1] By making the links to criteria for assessment (hence to the intended outcomes of the course) clear, students were consistently reminded about how the learning activities fit within the scientific literacy scheme.

The overall consistency of alignment in the Year 7 program is unsurprising. It is based in part on the materials produced for the Cognitive Acceleration through Science Education (CASE) program (Adey & Shayer, 2001), which have been refined over two decades to improve students' scientific literacy.

This consistency in the Year 7 materials compares favourably with the scoring of the Year 9 curriculum materials. In the latter no less than two of the criteria, the reviewers' scores indicate that the curriculum materials are not adequately aligned with intended outcomes of the course. The mean scores of the reviewers for all criteria at Year 9 were lower than the associated scores of the Year 7 materials. Reviewers noted that, although there were marginally more materials available for the teacher to access and use, they seemed less targeted to particular aspects of scientific literacy. One reviewer, [R1] commented that "…the activities in the Year 9 course seemed to consist of older, more contextually-based materials that have been shoe-horned [into] science literacy". Thus, the number of curriculum materials available to the teacher is actually less than it appears, as a significant proportion of the curriculum materials "do not actually address the development of scientific literacy" [R3]. This lack of focus of the Year 9 materials appeared frequently throughout the scoring and subsequent interviews. According to the reviewers, it seemed that the Year 9 curriculum has materials that are very much based on the transmission of the content knowledge rather than the scientific literacy, particularly when compared to the Year 7 materials.

Alignment of Assessment with Intended Goals

The alignment of assessment tasks with the intended goals was evaluated using a set of criteria and associated scoring system developed by Webb (1997). The alignment of assessments is scored on a three point scale, whose descriptors are presented in Appendix E. Again, an assessment is said to be aligned with the intended goals when each criterion has a mean score of at least 2.

Assessments used in Year 7 and Year 9 were scored individually against the criteria, and then the assessment program as a whole was scored. Tables 2 and 3 present the reviewers' ratings using the Webb framework.

Table 2: *Alignment scores for Year 7 assessments.*

Criterion	Score													
	Dog's bark		Safety task		Running race		Camping on the range		Candy Co.		Reflection booklet		Overall assessment materials	
	Mean score	Standard deviation	Mean score	Standard deviation	Mean score	Standard deviation	Mean score	Standard deviation	Mean score	Standard deviation	Mean score	Standard deviation	Mean score	Standard deviation
Categorical concurrence	2	0	2.7	0.47	2	0	1.7	0.47	3	0	1	0	2	0
Depth of knowledge consistency	2.7	0.47	3	0	3	0	2.3	0.47	2.7	0.47	2	0	2.8	0.24
Range of knowledge tested	1	0	1.3	0.47	1	0	1	0	3	0	1	0	1.7	0.24
Balance of representation	2	0	2	0	2	0	1	0	2	0	2	0	1.7	0.24
Cumulative growth in content knowledge	2.7	0.47	2.7	0.47	3	0	1.8	0.47	3	0	2	0	2.3	0.24

Table 3: *Alignment scores for Year 9 assessments.*

Criterion	Score													
	Temp prac		Conc. prac		Datsun mystery		Murder most foul		Reflection booklet		Examination		Overall assessment Materials	
	Mean score	Standard deviation	Mean score	Standard deviation	Mean score	Standard deviation	Mean score	Standard deviation	Mean score	Standard deviation	Mean score	Standard deviation	Mean score	Standard deviation
Categorical concurrence	2	0	2	0	2.3	0.47	3	0	1	0	1.7	0.47	2	0
Depth of knowledge consistency	2.8	0.47	2.8	0.47	2.7	0.47	3	0	2	0	2	0	2.5	0
Range of knowledge tested	2	0	2	0	1	0	3	0	1	0	2	0	2	0
Balance of representation	1.8	0.47	1.8	0.47	2	0	2	0	2	0	1.3	0.47	1.8	0.24
Cumulative growth in content knowledge	3	0	3	0	2	0	3	0	2	0	2.3	0.47	2.7	0.24

A considerable amount of information was generated in the determination of alignment of assessment, with the scores awarded by each individual reviewer featured in Appendices H and I. The data are unpacked by examining each individual criterion, awarding scores for the Year 7 and Year 9 programs, as well as determining the features of assessment that enable the strongest alignment.

Impressions of individual criteria

Categorical concurrence describes the degree to which the outcomes assessed on a particular assessment aligns with the curriculum materials that are associated with that assessment piece. Reviewers used the curriculum materials to determine the likely content and focus of instruction leading up to the assessment piece, and then rated them according to how well the assessment matched the curriculum documentation.

The range of mean scores for tasks in Years 7 and 9 was large. Several tasks rated only a 1 (no concurrence), while two other tasks rated below the alignment goal of 2. It must be acknowledged, however, that the two tasks that achieved the rating of one were essentially the same task performed at two different year levels. Most other areas achieved the score of 2, with several tasks being regarded as having an extremely strong link to the curriculum materials (Candy Co at Year 7 and Murder Most Foul at Year 9).

Two of the reviewers [R1 and R3] commented on the fact that the categorical concurrence score fluctuated depending on the aspect being assessed. It seems that while aspects relating to argument construction, hypothesising and data collection are frequently addressed in both the curriculum materials and the related assessment, aspects such as metacognition and ethical considerations were assessed yet had little, if any, curriculum materials associated with the instruction of these skills.

For these assessments that are not aligned on this criterion, the question must be asked why an aspect that does not seem to be taught is assessed. The developers of the science program must consider whether these aspects are indeed required portions of the course, and, if so, what instruction needs to be developed to support its development. Alternatively, simply producing curriculum materials to enable explicit instruction for assessed aspects would significantly raise the score.

Depth of knowledge consistency describes the degree to which an assessment caters for the range of cognitive ability in students. A strong score in this criterion indicates that the task has questions which elicit from the students a performance at the highest expected level of achievement. Generally, the scores for this criterion were very high, with every assessment achieving the level required for alignment. Two reviewers indicated that the open-ended nature of many of the tasks allowed the students to demonstrate a larger range of skills and understandings than the closed tasks. The lowest score was given to the Examination in Year 9, as it featured many low level questions that allowed students to achieve what appeared to be a reasonable result without demonstrating true understanding of the skills or the material. One reviewer [R2] described the Examination as being "very limited, and probably a relic from a previous course. Students didn't even need to have learnt any of the more sophisticated [concepts] in order to achieve the benchmark standard".

Range of knowledge tested describes the extent of a skill or concept that is assessed on an assessment. The scores in this area seem quite low, especially when compared to the previous category. Only one assessment in Year 7 was scored with a result above 2, and several others achieved a score of 1. Year 9 was marginally better, with all but two of the assessment pieces rating 2 or above.

All reviewers reported that the content required by the student to demonstrate their skills on several of the assessments reduced the score available. Although the tasks posed open-ended questions which seemed to supply the students with an opportunity to demonstrate a range of knowledge, the fact that the assessment often honed in on a very specific piece of content knowledge required for the demonstration of the skill influenced the reviewers to reduce the score in this criterion.

Assessment pieces that achieved a higher score on this criterion tended to ask multiple questions which, while still open, allowed the students to demonstrate their skills using a greater range of content and skill knowledge than other, smaller tasks. On closer analysis of these assessments, it seems that the efforts made to simplify tasks for younger students have actually lead to a narrowing of the focus beyond what was intended.

Balance of representation indicates the degree to which elements of the curriculum are weighted on the assessment to reflect the amount of instruction time given to these elements and the difficulty of the content. In a similar fashion to the Categorical Concurrence scoring, the reviewers used curriculum documentation to determine the scope of the instruction given in each of the aspects assessed and then related that degree of class time back to the weighting on the actual assessment.

Scores on this criterion indicate that the assessments often do not give appropriate weightings to curriculum elements, with scores ranging from 1 to a high of 2. The mean scores attributed to the balance of representation at each year level accurately reflect the comments of reviewers in the interviews. All three reviewers commented on the fact that each aspect assessed on an assessment was given equal weighting, even though the amount of time spent in class developing the skill varied greatly between aspects. One reviewer [R2] commented that "the assessments really need to be weighted differently……the amount of time spent in class clearly indicated that some aspects were more important than others, yet they were weighted the same on the task".

Cumulative growth in content knowledge indicates the degree to which assessment instruments elicit information according to how students' knowledge develops over time and how students relate these ideas. Generally the reviewers scored this category strongly. Only one of the assessment tasks was deemed not to show cumulative growth

(Camping on the Range), with all of the others being adjudged as showing alignment with the goals of developing students' science literacy. Most of the tasks were built to specifically refer to the learning that had come before the assessment, so that progress over time could be measured. The three reviewers indicated that it was encouraging to see that there was a clear progression of skill as the assessment program proceeded, although the one misaligned assessment item was "particularly divorced from the rest all of the other tasks" [R2].

Overall impressions of the assessment programs

It is interesting to view the assessment programs at Year 7 and Year 9 as a whole. The Year 7 program is aligned with the intended outcomes of the course, but has variance in the degree to which it is aligned across the criteria. The Year 9 assessments had a much greater degree of alignment than the Year 7 assessments.

As indicated by several Researchers (Broadfoot, 1996; Dochy & McDowell, 1997; Wiggins et al., 2001), each assessment piece provides only a small segment of the overall profile of a student. With the role of an individual assessment piece being to determine student achievement at a particular point in time, it is only when the entire assessment program is viewed that the alignment of the program can be properly measured.

The Year 7 assessment program seems constructively aligned with the intended outcomes of the course, but it must be noted that a significant gap appears in both the Range of Knowledge Tested and Balance of Representation criteria. Tasks in the Year 7 program consistently underperformed in these areas. This can probably be explained by the fact that, in an attempt to make the tasks shorter and more accessible by younger students, the assessment designers have narrowed the focus of the tasks, and hence inadvertently decreased the range of knowledge required. The limited range and balance of grading on these tasks means that the results from these assessments does not always accurately inform the students of their progress towards the intended goals, and so would not be considered constructively aligned. High scores were recorded in both the Depth of Knowledge Consistency and Cumulative Growth of Content Knowledge criteria.

The three reviewers indicated that they felt the removal or replacement of the weakest of the tasks, Camping on the Range, would improve the overall assessment program. The data on student learning obtained from this task is "nearly inconsequential" [R2] and "not really indicative of student learning on other tasks" [R3]. It could easily be replaced by a more informative task which is closely aligned with the intended outcomes.

The data indicate that the alignment of the Year 9 assessments is slightly better than the Year 7 assessments. The mean scores for the overall assessment program are generally higher than the minimum level of alignment, with only the Balance of Representation failing to reach that standard. Two of the reviewers felt that the variety of formats, extended length of tasks (usually expressed as openness) allowed the tasks to more adequately enable the students to demonstrate their developing skill. This result is not unexpected – the Year 9 assessment program has been taught and assessed 14 times, and the tasks adjusted each time to provide better information, particularly compared with the Year 7 course, which is earlier in its gestation.

It is interesting to note that the task that appeared in both assessment programs (the Reflection Booklet) scored exactly the same value in each year level. Despite the poor alignment scores on some criteria, reviewers indicated that this task is an integral part of the assessment program as it is the only portion of the program where the students are asked to formally report on their achievement and how they might improve on it.

Features of aligned assessment

From the data provided by the reviewers, both through the scoring and the semi-structured interviews, it is possible to identify the features of assessment tasks which are more closely aligned than others with the intended goals according to this scoring model. These features of constructively aligned tasks can guide the revision of the assessment program to further enhance its alignment. There were five broad features of an assessment and the associated program that enabled alignment.

Links to scaffolded instruction that has occurred before the assessment was undertaken. The assessments that were most aligned were carefully selected to represent the learning that had taken place in the classroom, and were administered at a time appropriate to the

learning. The less successful tasks were described as being "put in to satisfy the [reporting] timelines. It seemed like [the assessment task's] only purpose was to generate a number." [R1]

However, it is important to distinguish the difference between an assessment which is linked to instruction and an assessment which is not constructively aligned. An assessment task can be related to previous learning, both in terms of context and scientific literacy and still require students to make links and learn as they are being assessed. To adequately display their skills, students need to have the basic skills and knowledge required to engage with the task. As Broadfoot (1996) argues, if a student cannot engage with the language or the skill expectations of an assessment, and these missing skills are not what the assessment is trying to measure, then the assessment piece is invalid.

Open-ended tasks generally provided the students with more freedom to generate a response which utilised a variety of skills. Although reviewers recognised the need for deliberate practice in the lead up to the assessment, the aligned assessments featured problems which could be approached in a variety of ways, and were accessible by students at almost any point in the learning progression. This accessibility was noted by several reviewers; for a task to be successful, careful consideration needed to be given to how an underperforming student could structure their response. Two reviewers (R1 and R3) commented on the fact that the early tasks in the Year 7 program required that the students had a firm grasp of a significant amount of scientific conventions and terminology. As a consequence, teachers would "need to make sure that [the students] have been taught the science language and ideas they need to access the assessment" [R3].

Tasks involving *relating experimental ideas to contexts* also showed a greater alignment with the intended outcomes of the science program. These tasks typically took the form of an experiment related to, or extending on, theory investigated in class. Students in these tasks are required to draw on the meaning they have constructed for themselves and use it to provide a response to a question. This approach allowed greater links to the curriculum materials (categorical concurrence), allowed a range of interpretation and extrapolations (depth of knowledge consistency) and tracked growth in thinking over time (cumulative growth of knowledge). Tasks such as Candy Co and Murder

Most Foul were good examples of this relationship between experimental ideas and contexts.

Assessing *multiple aspects* on a single task was also a feature of the most aligned tasks. Although sometimes reviewers felt it was "handy to do short [tasks] which only test one aspect" [R2], multiple aspect allows the students to draw on a greater range of skills, and, in conjunction with an open task design, result in a greater range of knowledge assessed.

One of the key features identified by all reviewers was the need for *deliberate task design*. As mentioned previously in the findings, some tasks apparently consisted of a set of questions which assessed content knowledge rather than the intended outcome of the program, and then had a 'token' question or alteration made to satisfy the outcomes. The main purpose of the science course is to develop students' scientific literacy, and is measured on developmental continua (Appendix B). The most successful of the tasks had obviously been designed with the continua in mind; they required expression of a number of skills that increased in difficulty. The tasks were both not too hard that the least progressed student couldn't give a response, nor so easy that the highest performing students were not able to display the full extent of their understanding.

All of these attributes can be developed in tasks that are specifically designed for the purpose of accurately assessing against the continua. By analysing the tasks that are most aligned, it is possible to rapidly revise the assessments to enhance the alignment between the intended curriculum and the assessment program.

Alignment of the Assessment with the Epistemological and Cognitive Goals

Scoring of the alignment of assessment with the epistemological and cognitive goals of the science program was achieved by using the alignment framework proposed by Chinn and Malhotra (2002) combined with the conceptual framework described by Krathwohl (2002). In this analysis, each assessment item is mapped onto both the cognitive process dimension and the epistemological goals of the science program. Each of the assessments filled one or more of the goals and dimensions.

Each assessment in Year 7 and Year 9 was scored individually against the goals and the dimensions. The name of the assessment task is placed in the boxes corresponding to the goals and dimensions it displays. For example, the Safety Task in the Year 7 program requires the students to apply their understanding when generating a research question. So, in Table 4 below, the name of the task (Safety Task) has been transcribed into the intersection between the Apply dimension and the Generating Research Questions goal.

The process of mapping the tasks was predominately performed during a shared scoring session. Some disagreement occurred as to the nature of some of the items in several of the assessment tasks, as there were differences in opinion about the where these items fit in the Krathwohl conceptual framework. After some discussion the items were placed with agreement from each of the reviewers. Tables 4 and 5 below feature the reviewers' scores using the framework.

Table 4: *Alignment of Year 7 assessments with epistemological and cognitive goals.*

		The Cognitive Process Dimension					
		1. Remember	2. Understand	3. Apply	4. Analyse	5. Evaluate	6. Create
Epistemological Goals	Generating research questions		Safety Task Candy Co	Safety Task Candy Co	Running Race Camping on the Range	Reflection Booklet	Candy Co
	Designing Studies: select variable			Safety Task Candy Co		Reflection Booklet	Safety Task
	Designing Studies: planning procedures				Safety Task Candy Co	Candy Co Reflection Booklet	
	Designing Studies: controlling variables			Safety Task Running Race Camping on Range Candy Co		Candy Co Reflection Booklet	
	Designing Studies: planning measures			Safety Task Running Race Camping on Range Candy Co		Reflection Booklet	Safety Task Candy Co
	Making Observations					Candy Co Reflection Booklet	
	Explaining Results: transforming observations						
	Explaining Results: finding flaws				Safety Task Running Race Camping on Range Candy Co	Running Race Camping on Range Candy Co Reflection Booklet	
	Explaining Results: indirect reasoning						
	Explaining Results: generalisation			Safety Task Candy Co		Reflection Booklet	Safety Task Candy Co
	Explaining Results: types of reasoning					Running Race Camping on Range	
	Developing Theories: level of theory				Dog's Bark	Reflection Booklet	Dog's Bark Safety Task Candy Co
	Developing Theories: co-ordinating results from multiple studies						
	Studying research reports						

Table 5: *Alignment of Year 9 assessments with epistemological and cognitive goals.*

		The Cognitive Process Dimension					
		1. Remember	2. Understand	3. Apply	4. Analyse	5. Evaluate	6. Create
Epistemological Goals	Generating research questions				Examination		
	Designing Studies: select variable				Temp prac Conc prac Murder Most Foul	Conc prac Murder Most Foul	Examination
	Designing Studies: planning procedures				Temp prac Conc prac Murder Most Foul	Conc prac Murder Most Foul	Examination
	Designing Studies: controlling variables			Examination			Examination
	Designing Studies: planning measures			Examination			
	Making Observations				Temp prac Conc prac Murder Most Foul	Conc prac Murder Most Foul	
	Explaining Results: transforming observations			Temp prac Murder Most Foul	Temp prac Datsun Mystery Murder Most Foul	Datsun Mystery Murder Most Foul	Temp prac Datsun Mystery Murder Most Foul
	Explaining Results: finding flaws				Examination	Murder Most Foul Examination	
	Explaining Results: indirect reasoning				Datsun Mystery Murder Most Foul	Datsun Mystery Murder Most Foul	Datsun Mystery Murder Most Foul
	Explaining Results: generalisation		Examination	Examination	Conc prac Murder Most Foul	Conc prac Murder Most Foul	Temp prac Conc prac Datsun Mystery Murder Most Foul Examination
	Explaining Results: types of reasoning			Datsun Mystery Murder Most Foul		Datsun Mystery Murder Most Foul	Temp prac Conc prac Murder Most Foul
	Developing Theories: level of theory		Examination	Datsun Mystery Murder Most Foul Examination	Temp prac Murder Most Foul	Datsun Mystery Murder Most Foul	Temp prac Conc prac Datsun Mystery MMF Examination
	Developing Theories: co-ordinating results from multiple studies		Murder Most Foul	Murder Most Foul	Murder Most Foul		
	Studying research reports						

Both tables indicate that the assessments used in each program show particular trends in the epistemological goals and process dimensions assessed. The alignment data are best addressed by dealing with the cognitive process and epistemological goals separately, and then examining the links between the two frameworks.

Cognitive process dimensions

The assessments of both programmes focus heavily on four of the six cognitive process dimensions (Apply, Analyse, Evaluate, Create), with very little attention paid to the first two (Remember, Understand). In the Year 7 assessment program, only two tasks involve the use of the Understand dimension and none of the tasks require students to use the Remember dimension without tying it to another of the dimensions. While the Year 9 program does give more attention to the Understand dimension, it still is not addressed as comprehensively as the other dimensions.

The emphasis of the assessment program of the two year levels appears to be different. The Year 7 program (in Table 4) features application of knowledge in almost every task, and this is supported with a strong emphasis on the evaluation of their work. One of the reviewers [R2] commented that "all the kids seem to be doing in Year 7 is identifying variables, constructing methods and then evaluating their work". There is less emphasis on creating and analysing, with only the Safety Task and Candy Co providing the students with the opportunity to create their own experimental design. These tasks tend to feature more open-ended investigations, in which the students must create methods for investigation in order to test hypotheses they have constructed. It is not surprising that tasks requiring an extended and more considered response than others in the assessment program would demonstrate a stronger emphasis on the Create and Analyse dimensions than the Remember and Understand.

The Year 9 program (Table 5) features different emphases. The Analysis, Evaluate, and Create dimensions are all heavily featured throughout a number of tasks, and the distribution of assessment between the dimensions is relatively even (compared to the Year 7 program).

One reason for the increased prevalence of higher level process dimensions in the Year 9 program is the increased size and complexity of the assessment tasks. As expected the Year 7 tasks tend to be

smaller and more contained than the Year 9 assessments. Originally, the Year 7 tasks were designed in this manner to prevent the students application of effort petering out (which can sometimes occur if it is too large), and also to reduce the complexity of the ideas and models they were attempting to deal with. However, it appears that by making the tasks more manageable the designers of the assessment have "lost some of the things that make the tasks real, and make [the students] think more about their work" [R2].

The Year 9 tasks generally feature broader and more open-ended ideas and investigations requiring the students to extend some of their mental models. For example, the Concentration practical requires students to develop an understanding of the chemical measurement of concentration, link increasing concentration to increasing reaction rate and then use their mental models of particle movement to explain what they have observed. The task requires the students to design and investigation to test a hypothesis they have developed, make and analyse their observations, and then use these observations to extend their mental models of particle and collision theory. It is a good example of a task which requires the students to Create, Analyse and Evaluate during an assessment task, and, according to the reviewers, "is a better example of what [the school] is trying to develop" [R1].

Epistemological goals

The most startling differences between the assessment programs of Year 7 and Year 9 are seen in the epistemological goals of the course. The Year 7 program has a heavy emphasis on the *design* of studies; many of the tasks require the students to design a scientific investigation, including identifying variables and planning measures, but less emphasis is placed on explaining results and developing theories. It is interesting to note that the reviewers could not find a single assessment task in the program that addressed one of four of the epistemological goals: explaining results: indirect reasoning; developing theories: co-ordinating results from multiple studies; or studying research reports. The reviewers hypothesised that this could be due to the idea that "the skills are pretty difficult to teach well" [R1] to Year 7 students, particularly as they are still developing their scientific literacy.

In contrast, the Year 9 program shows a large number of tasks which require students to design a scientific investigation and explain the results. In particular, the reviewers indicated that many of the assessment tasks featured sections in which students were required to find flaws in their investigations, represent their results in a fashion which is most easily understood and then make generalisations based on the results they obtained. This was quite different to the Year 7 program, as these Year 9 assessment tasks "actually required the kids to think about how their investigation turned out, and whether their data actually had some meaning." [R3] In general, the Year 9 assessment program seemed to address more of the epistemological goals of the course during the term with a heavier emphasis on the generalisation and evaluation of the results students obtained.

Two reviewers commented on the fact that at no stage in either assessment program is a student required to study an existing research report as part of the assessment task, despite this being one of the fundamental aspects of science investigation. Although the students are often attempting to make links in their learning that involve ideas and theories that are already known to the scientific community, an emphasis on the research of others, and how almost all current research relies on previous work, would enable them to gain a greater understanding of the nature of science.

Links between epistemological goals, cognitive process dimensions and the assessment program

The reviewers' mappings, based on the assessment models in the case study science program, show that, in general, when an assessment successfully shows elements of the epistemological goals of the program, it requires the students to use four of the cognitive process dimensions in particular: Apply, Analyse, Evaluate and Create. There are very few tasks in either year level that achieve the goals without requiring the students to show elements of these four dimensions. The epistemological goals of the program are addressed most obviously when the task operates primarily in these dimensions.

The traits of assessment tasks that seem to feature most prominently in Tables 4 and 5 (and hence show greatest alignment to epistemological and cognitive goals) are those that are open-ended and student driven. Those that are smaller, closed tasks designed to elicit responses which indicate progress in particular skills did not tend to appear frequently in the tables, and addressed few of the required goals. The tasks more closely aligned to the epistemological and cognitive goals require

the students to generate a research question based on a dilemma, design an effective research strategy and then evaluate the results of their work. This mirrors the design process in academic research, with one exception. Typically real life science research has a component in which the Researchers search research reports and journals to determine the extent of the knowledge pertinent to a particular research question. As mentioned previously, the lack of any emphasis on any tasks in either of the year level assessment programs shows the students are not being exposed to a crucial step in the scientific process, and an important element of scientific literacy.

Overall Impressions of the Case Study Science Program

The ratings provided by the reviewers across the three instruments used to evaluate the case study science program give an indication of the degree to which the intended goals, curriculum materials and assessments are aligned. The goals of the program are to develop students' scientific literacy, including an understanding of how scientific research is conducted in the real world i.e. epistemological goals.

The curriculum materials appear to align well with the intended goals of the course, according to the criteria developed by Kesidou and Rosemann (2002). All but three of the criteria across the two year levels showed a mean score greater than 2. This indicates that, in general, the curriculum materials are well-aligned to the intended goals of the course, and are consistent across year levels. This consistency of format and approach enables students to identify the purpose of the materials, and how one idea and skill links to another with greater ease. The integration of the work by Adey and Shayer (1990) provides appropriate models which could be used to develop more effective curriculum materials, as the assessment pieces which were based on their work were more aligned with the intended goals of the program.

An area of weakness in the curriculum materials appears to be the number of materials available to the teacher; provision of activities and instruction directly targeted to the intended goals was lacking in both year levels, particularly in Year 9. By expanding the number and quality of these materials, the case study science program could be more effective in improving science literacy. In particular, avoiding "shoe-horned materials" [R1] and developing the resources specifically for the course

would “help make the integration of science [sic] literacy with the science contexts much more achievable for teachers” [R1].

The assessment used to measure student progress in the case study science program was also judged to be effectively aligned with the intended goals. Both the Year 7 and Year 9 programs have tasks which are far more representative of authentic scientific inquiry and promote scientific literacy than other tasks in the same program. The reviewers did believe that although the assessment tasks would give a relatively accurate indication of student progress in scientific literacy over time, improvements could be made to increase the effectiveness and accuracy of the program. In particular, greater emphasis on open-ended tasks which more strongly mirror authentic science inquiry and more thought given to the degree to which some aspects of scientific literacy are assessed compared to others would enable the tasks to be more representative of the science program’s intended goals.

The number of tasks in each year level seems adequate considering the size of each task, although it was commented that, at Year 7, “to fit in all the assessment you would need to be assessing every three to four lessons [210 – 280 minutes]……this might be too much for the young [students], especially if the tasks became longer” [R2]. Since the more effective tasks are those that are longer with greater freedom, there may be a need to reduce the number of tasks the students attempt in a term. As mentioned previously, the omission of less aligned tasks (such as Camping on the Range) would make the program far more effective as a cohesive unit.

Features of a program that most adequately enables alignment

From the data provided by the reviewers, both through the scoring and the semi-structured interviews, it is possible to identify the features of a science program which has a greater alignment in intended goals, curriculum and assessment than others. Both the curriculum materials and the assessment must match the intended outcomes of the course, since it is around the intended outcomes that the course has been constructed (Biggs, 1996), and the only purpose of the materials is to drive the development of the outcomes.

Curriculum materials which are most strongly aligned share several key features. First, they are specifically tailored to the teaching sequence. As the interventions made by a teacher in a student's development of scientific literacy are very deliberate, materials should be developed in such a way that specifically target a certain stumbling block that occurs often in the learning process. By considering carefully the nature of the intervention and the materials required to support it, curriculum developers can produce materials which are more effective in helping students develop the outcomes as presented by the curriculum. Secondly, the assessment tasks are formatted in a manner so that the intended learning from the activity or intervention are very clear to students attempting the tasks. Having a common format that indicates the aspect being worked on and the conceptual stage the material is attempting to address means students are better able to engage in the learning process by understanding and utilising the metalanguage of both science and education (Mortimer & Scott, 2003).

The features of an effective assessment program need to be considered both collectively and individually. Aligned assessment programs are strongly linked to a learning path, where the progress of learning is clearly presented to both students and staff. The most effective program had regular assessments given, with a range of different tasks.

The assessment tasks which were most strongly aligned to the intended goals of the program have five key features. First, the tasks are linked to scaffolded instruction that describes to the student the learning path that needs to occur, and provides them with the necessary skills to make the next step in their learning. Secondly, these tasks were open-ended, providing the students with more freedom to generate a response which utilised a variety of skills. It is important that the students have the freedom to generate their own ideas and concepts without having to guess what the teacher is looking for. It provides the students with an opportunity to construct meaning from what they are producing, and aligns more closely with the goals of the case study science program. Thirdly, the tasks requiring students to relate experimental ideas to contexts also showed a greater alignment with the intended outcomes of the science program. Tasks which are more closely related to authentic science inquiry seem to lend themselves better to both more effective student learning and meaning-making. The task should also assess multiple aspects, allowing for a greater range of skills to be tested. Fifth, the tasks should be deliberately designed with the continua in mind; they require application of a number of skills that increased in difficulty. The tasks need to allow both

the least progressed student to give a response and the highest performing students to display the full extent of their understanding.

Summary

Chapter four discussed the research findings relating to the constructive alignment of the intended outcomes, curriculum and assessment in the case study science curriculum. The first section reiterated the aims and objectives of the present research, while the second section reviewed the instruments used to review the curriculum. The third section examined the alignment of the curriculum materials with the intended goals of the science program, indicating that the Year 7 program, despite the limited range of curriculum materials available, had a consistently strong alignment. This differed from the Year 9 program, which, although achieving alignment according to Kesidou and Roseman's (2002) criteria overall, had two of the seven criteria which did not show adequate alignment. The reviewers commented on the fact that, at both Year 7 and Year 9, the curriculum materials were either limited in number or were not as focused on the development of scientific literacy as the related assessment tasks.

The degree to which assessment is constructively aligned as indicated by the scoring data was described in section four, and the overall impressions of the assessment programs in Years 7 and 9 were examined in section five. The data showed that the Year 7 program, with its shorter assessment length and breadth, performed relatively poorly on the Range of Knowledge Tested and Balance of Representation criteria. The Year 9 program had, on average, a much greater alignment with the course's intended goals, both in terms of scientific literacy and epistemological understanding. Almost 90% of the tasks in the case study science programs required students to work in the higher domains of the cognitive framework, with few tasks other than the examination requiring that students simply recall and relate conceptual information.

The sixth, and final, section highlighted the features of the science program that most adequately supported alignment. The data show that the assessment tasks which had the highest degree of alignment were open-ended in nature and were explicitly linked to the scaffolded instruction and the related curriculum materials. Also, they matched the epistemological goals of the program by relating directly to the elements of real-world scientific research, and were designed to directly

assess the intended goals of the course, allowing the students to demonstrate a wide range of achievement of a particular skill.

The next chapter addresses the study's findings in relation to the effectiveness of the curriculum evaluation model developed and implemented in this study.

CHAPTER FIVE: FINDINGS – EFFECTIVENESS OF THE CURRICULUM EVALUATION MODEL

Introduction

A discussion about the effectiveness of the curriculum evaluation model is presented in this chapter, which is divided into four sections. The first section explains the use of Fleiss' kappa co-efficient, while the second discusses the application of the co-efficient to the case study program. The third section provides a brief overview of the semi-structured interviews before discussing the content and implications of the interviews themselves. Finally, the degree to which the curriculum evaluation model is effective is explored in the fourth and final section.

Reliability of Ratings

One measure of the reliability of a curriculum evaluation model is the degree to which different participants are able to agree on a rating of particular materials based on a given criterion. Agreement (or similar rating) indicates that participants are able to interpret criteria appropriately and apply ratings in a similar fashion. The degree to which two or more raters have agreement in their ratings is known as inter-rater reliability. Reliability gives an indication of the confidence we can have in the consistency of ratings. If the ratings are considered reliable, then when another piece of work was to be scored by other raters, the ratings awarded would be expected to be broadly similar (Broadfoot, 2007). Note that reliability is not the same as validity; results from participants which are erroneous, yet similar, are reliable but not valid.

There are several methods used to determine inter-rater reliability, but the method most appropriate for this particular study is Fleiss' kappa co-efficient. Fleiss' kappa expresses the extent to which the agreement between raters on a particular criterion exceeds that which would be expected through pure chance (Fleiss, 1971). It is related to the Cohen's kappa measurement, but has the advantage of being able to measure the level of agreement between more than two raters, which is particularly pertinent to this study. A kappa of 0.61 indicates that the agreement of the raters is significantly different to that expected by chance, and indicates "an acceptable level of inter-rater reliability" (Fleiss, 1971, p. 277).

In this study, the kappa co-efficient calculations have been performed for the alignment of curriculum materials using the Kesidou and Roseman (2002) and Webb (1997) frameworks with the case study science program's intended goals.

Inter-rater reliability of alignment of curriculum materials with intended goals

The Fleiss' kappa co-efficients were calculated for ratings given to curriculum materials available for both the Year 7 and Year 9 course, and are presented in Table 6 below.

Table 6: *Fleiss' kappa co-efficients of reviewer ratings for alignment of curriculum materials with intended goals.*

Year Level	Fleiss' Kappa Co-efficient
7	0.69
9	0.26

The kappa co-efficients for the two year levels contrast sharply. The kappa co-efficient of 0.69 generated from the Year 7 materials indicates that the inter-rater reliability of the rating of these materials is quite high, as it sits above the acceptable level of 0.61. This rating indicates that the reviewers gave similar scores for the Year 7 set of materials, and the degree of similarity was higher than that expected of random rating allocation. However, the kappa co-efficient for the Year 9 materials is only 0.29. This value indicates that the level of agreement does not vary significantly from that expected from a random allocation of ratings, and casts some doubt on the reliability of ratings awarded by the reviewers.

This difference in kappa co-efficient could result from several factors. Firstly, the curriculum materials that are scored in the Year 7 program are significantly different in presentation and content from those of the Year 9 program. In particular, the Year 7 materials are shorter and develop a narrower range of skills than the Year 9 materials. The fact that the reviewers found a greater level of agreement for the Year 7 materials means that the criteria described by the Kesidou and Roseman (2002) framework may be more easily applied to some forms of curriculum materials

than others, particularly those that are short activities with limited scope and significant scaffolding. Secondly, the process used to train the reviewers in the use of the Kesidou and Roseman (2002) framework utilized some of the Year 7 curriculum materials assessed in this study, and hence focused on types of curriculum materials more prevalent in the Year 7 program than in Year 9. It is probable that this increase in collaborative marking on these types of materials may have resulted in a greater level of agreement when scoring them as opposed to materials which had significant differences in scope and scaffolding. Thirdly, there may have been confusion as to the meaning and interpretation of each of the criteria. Scoring only some types of materials in the training sessions may have made it difficult to determine the extent to which reviewers had a common understanding of the criteria.

Inter-rater reliability of alignment of assessment with intended goals

The Fleiss' kappa co-efficients were calculated for the assessment tasks used in both the Year 7 and Year 9 courses, which are featured in Table 7 below.

The most notable feature of the kappa co-efficients as applied to the assessment tasks is the variance in the results. Two tasks in each set of assessment materials have a co-efficient of 1.0, which indicates complete agreement (all reviewers gave the same rating to that particular assessment task). For other tasks, however, the values were as low as 0.37.

Table 7: *Fleiss' kappa co-efficients of reviewers' ratings for the alignment of assessment with intended goals.*

Year Level	Task Name	Fleiss' Kappa Co-efficient
7	Dog's Bark	0.67
	Safety Task	0.50
	Running Race	1.0
	Camping on the Range	0.48
	Candy Co	0.81
	Reflection Booklet	1.0
	Overall Assessment Program	0.37
9	Temp Prac	0.65
	Conc Prac	0.65
	Datsun Mystery	0.66
	Murder Most Foul	1.0
	Reflection Booklet	1.0
	Examination	0.43
	Overall Assessment Program	0.66

The Year 7 program had the greatest variance in kappa co-efficients. Although several of the tasks had high inter-rater reliability (Dog's Bark, Running Race, Candy Co. and the Reflection Booklet all had a kappa greater than 0.61), three tasks, as well as the overall program rating, showed a lower kappa co-efficient. With the majority of the tasks indicating that the reviewers were reliable in their scoring, it is interesting that the overall evaluation co-efficient was so low (only 0.37). This may reflect the difficulty of giving a rating for a wide variety of task types and lengths. The lack of specific instruction given to reviewers in terms of the weighting of particular tasks when rating the overall program might have contributed to the low reliability.

The inter-rater reliability of the Year 9 program was higher than that of Year 7. All but one of the tasks (Examination) had an acceptable level of inter-rater reliability, which represented a stronger level of agreement than in the Year 7 program. The most notable difference in the scoring was the co-efficient for the overall assessment program, particularly in light of the Year 7 kappa co-efficient discussed above. The Year 9 program has less variety in the types of tasks it contains; generally the tasks are open-ended and experimentally based. Therefore, when making judgments about the rating of the overall program, the reviewers found it "much, much easier to come to a decision" [R2]. It is worth noting that only two of the tasks differed from the general open-ended model favoured in Year 9: the Reflection Booklet and the examination. The reflection booklet achieved

perfect reliability, as the scope of the task is very small and the outcomes obvious. The examination, on the other hand, attempted to link recall and observation items with questions tailored more towards the scientific literacy aspects of the course. The relatively low kappa co-efficient (0.43) indicates the difficulty the reviewers had when scoring an assessment task (or program) that contained several components, which differed markedly in scope or focus.

The Fleiss' (1971) kappa co-efficients indicate two weaknesses in the curriculum evaluation procedure. First, the training program used to familiarize the reviewers with the various scoring methods was not comprehensive enough to allow them to score independently with reliability. The fact that only materials and assessments of particular types were scored in the training sessions meant that when the reviewers were faced with materials that differed from those they had practised with, there was a decrease in the reliability of the ratings awarded. Developing a training program which takes into account all of the types of materials likely to be investigated in the program, with consistent checking of the ratings awarded, could eliminate the discrepancies in the understanding of the reviewers. However, this would require a significant amount of time on the part of the reviewers, and perhaps reduce the ability of the curriculum scoring method to be implemented in schools.

Second, attempting to rate a large collection of materials as a whole made it far more difficult for reviewers to accurately decide on a rating. The data showed that the ratings awarded to the overall programs are reliable enough to provide reviewers or institutions with information which is worthwhile enough to act upon. Instead, curriculum materials and assessments could be grouped into 'like' materials and rated in these terms rather than as an entire program. This change would reduce the variables considered by the reviewers, and likely increase inter-rater reliability.

The Semi-Structured Interviews

The interviews took place at the conclusion of the rating process towards the end of September, 2009. Interviews were conducted with the three reviewers who scored the curriculum materials and assessments using the curriculum evaluation model. In all three cases, Edith Cowan University Human Research Ethics Committee guidelines were followed. All interviews were conducted face-

to-face with each respondent at the case study school. Each interview lasted approximately half an hour, with the longest taking forty-five minutes.

Each semi-structured interview consisted of a set of eight questions, but was open to exploration of related issues raised by the participant. The purpose of the interviews was to identify strengths and weaknesses in both the curriculum evaluation model and the case study science teaching program, and often the most useful responses came in the sections of the interview not directly prompted by the questions.

Verbal responses were audio-recorded and then transcribed (by the Researcher) on to a transcript summary page. This transcript summary was then viewed by the interviewee in the week after the conclusion of the interview to ensure that the transcript summary was accurate. At this point all names and identifying information were removed from the transcripts and each respondent was issued with an identification number. Printed transcripts were then given to respondents for final checking, approval and changes made if required. Only after this process had been completed was any information analysed and included in the research

In each case, interviewees showed a great deal of interest in the topic of discussion, displaying animation and obvious enthusiasm. The reviewers were extremely keen to discuss the relative merits of the curriculum evaluation model, and had obviously spent some time considering the merits of, and possible improvements to, the evaluation model.

Responses to the Interview Questions

The responses to the semi-structured interviews were grouped into themes. The themes were, in part, guided by the questions posed in the semi-structured interview. It should be noted that not all the responses attached to a particular question below resulted from a direct answer to that question; however, each of the responses included faithfully represent the intended meaning of the respondent. The major themes of discussion were: the degree to which the data were meaningful; the degree to which the instruments indicated areas for the improvement of alignment; the practicality of using this scoring system in a school system; effectiveness of the training program;

and suggested changes to the instruments and methodology to assist in the ease of use and reliability of the process.

> Do the instruments provide meaningful data? Were there any criteria in any of the instruments that were unclear or extraneous?

As the opening set of questions in the interview, these questions sparked a wide ranging discussion which encompassed several of the other semi-structured questions. All three respondents agreed that the instruments, taken as a whole, provide meaningful data. However, each of the reviewers expressed concern about elements of the data collected and the instruments used in the scoring process.

There was some concern, as discussed below, with the ability of teachers and administrators to cope with the significant data literacy demands of the process. The instruments generate a large amount of data, and the usefulness of the data is highly dependent upon the ability of the data user to understand what the data means:

> It is a large process which generates a lot of data for each course. You've got to ask yourself whether the sheer volume of information is useful...What amount of data can people actually engage with and use before they are just awash with information? [R2]

The concern seems valid – each assessment task alone generates in excess of 10 data points per reviewer, and an entire assessment program may involve a reviewer making literally hundreds of criterion referenced judgements. The fact that a large amount of information is generated from the instruments means that users of the system need to be able to make sense of the data. One reviewer [R3] commented that:

> When I sat back and looked at the data I had generated, there were a lot figures, yet I know that after scoring the materials I could categorically determine whether [curriculum materials, intended outcomes and assessments] were aligned or not. Would anyone else who looked at my scores alone be able to make the same judgement? I'm not so sure.

The other two reviewers expressed similar sentiments about the data produced. The usefulness of the data hinges upon the ability of the reader to make sense of the information. The reviewers felt the provision of too many data points, without an appropriate way of isolating the areas of importance, severely restricts the degree to which the data is meaningful.

In terms of the data generated by individual instruments, reviewers' opinions showed some consistency. All three reviewers commented that the analysis of the curriculum materials (using the Kesidou and Roseman (2002) model) was quite meaningful when taken as a curriculum program. "Using the criteria, it cut right to the heart of whether [the curriculum materials] had any relevance to the course goals." [R1] However, all three reviewers commented on the fact that scoring, although useful on the large scale, could mask problems with specific materials.

> For example, when I scored the Year 7 [materials], most of them really aligned well with the goals. But there were two sheets that really stood out for me. One was photocopied straight from a text, and I had no idea what the goal was. I couldn't figure out just how it related to the course, so I am sure the students would have had no idea! Then the one that was obviously filler about dihydrogen monoxide…..you can give a good score for the overall program, but that can hide some really poor stuff. I guess it could happen the other way around as well: bad scores for the program, but a good activity or two. [R1]

The reviewers did acknowledge, however, that the scoring of each individual curriculum material was impractical due to both the amount of data that would be generated and the amount of time required to score so many materials.

The analysis of the alignment of assessment materials with the intended goals of the program (based on the Webb (1997) model) was generally judged to be meaningful, with one criterion a notable exception. Reviewer R3 encapsulated the thoughts of both R2 and R3 when he responded that:

> Overall the criteria enabled us to produce meaningful data, except for the Categorical Concurrence. It just didn't seem all that important that absolutely everything you have studied in the class had to appear on the assessment task…I guess I had trouble determining exactly what this criteria [sic] meant, and so I doubt that my scores for [Categorical Concurrence] would be right.

The inability of the reviewers to feel confident about the data they have generated with this criterion suggests a lack of effectiveness of the training program, and perhaps some elaboration needs to occur with the scoring rubric.

In light of the Fleiss' kappa co-efficient results discussed earlier, the reviewers indicated frustration at their inability to be able to accurately score entire assessment programs. R1 expressed the problems most efficiently, "I just plucked a number out that seemed to fit with the other scores I had given. It was not at all reliable. A waste of time really." While the other reviewers did not experience the same degree of irritation with the judgements they made about the overall program, they did identify the fact that the variation and weighting of tasks made the scoring difficult. One reviewer indicated that this could be solved with a better training program, "If we had had some sort of guideline about how to score the program, we might have had a better chance. We didn't cover it in the training day, and so I felt that I was making up my own rules with that one." [R2]

The evaluation of the degree to which the assessment tasks achieved the course's epistemological goals was roundly criticized by the three reviewers. They indicated that they "just can't see how the information we get from this scoring is useful." [R3] Two of the reviewers expressed their frustration that the results were typical of what they already knew; that they could have simply "flicked through the tasks and still got an idea of which of the epistemological goals were addressed." [R3] The frustration exhibited by each of the reviewers would indicate that the information generated from the Chinn and Malhotra (2002) instrument is "not meaningful to either reviewers, teachers or administrators." [R2]

Similar concerns were expressed about the cognitive process dimension proposed by Krathwohl (2002). One of the reviewers [R2] made the point that the very goals of the program included an expectation that the students would be operating at particular levels of the cognitive dimension:

> Let's face it, I can't think of a question you can ask which has relevance to the science [sic] literacy goals of the course which would have the students just using the lower dimensions of the [cognitive process dimension]. Why would you require someone to chart all of the assessment tasks when they should be operating in the higher dimensions?

Although, when considering the case study school program, the reviewer's comment seems valid, the same spread of cognitive process results may not be observed when applied to different programs addressing scientific literacy.

> Could the data provided by these instruments allow the realignment of curriculum materials, assessment and/or instruction?

In all of the interviews, the reviewers moved to address this question as a part of the previous question's response without it being formally asked. Two of the reviewers (R1 and R2) felt that the data provided by the instruments would be useful in guiding the realignment of the curriculum materials, assessment and/or instruction. They indicated that being able to recognise the features of curriculum materials and assessment tasks that were considered aligned meant that other, less aligned tasks "could be just changed so that they were similar to the better [more aligned] tasks" [R2].

> Would scoring curriculum materials using these instruments be practical in a school setting?

Each of the participants was asked to give an estimate of the amount of time taken to score the instruments, inclusive of the time spent scoring during the training sessions. The times reported by the participants are summarised in Figure 10 below.

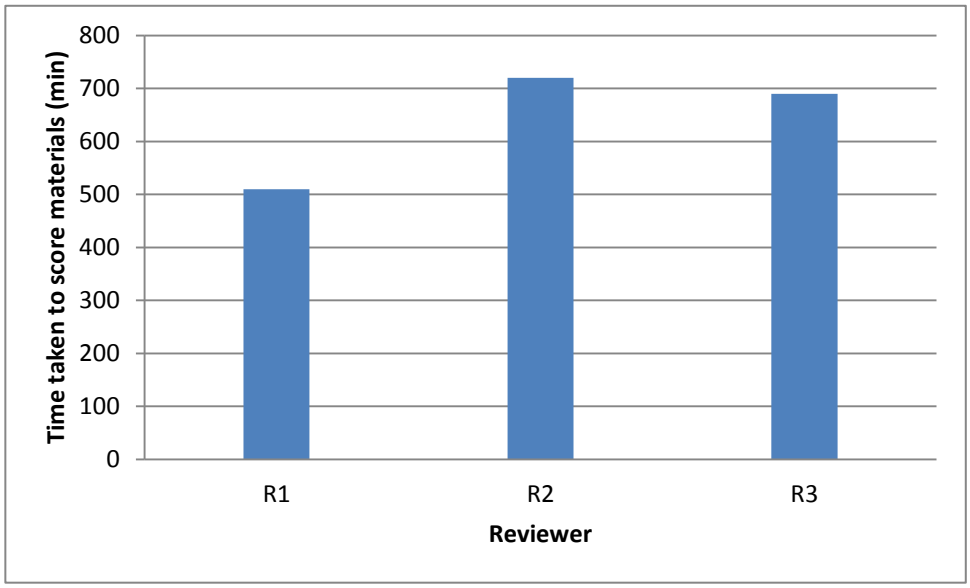

Figure 10: *Time taken for participants to score materials.*

All three reviewers indicated that the scoring took in excess of eight hours to complete, with R2 describing the process as "taking twelve hours to finish…..I had to spread it over several nights, which made it take longer. It just takes a while to remind yourself of each of the criteria, and check back through the notes made from the training session."

When asked whether the time taken was a reasonable expectation for an analysis of alignment, opinion was divided. R3 indicated that the time taken was affected by the number of sessions that the analysis was spread over:

> If you break the sections into the individual instruments, doing one [scoring of an instrument] per session, then the scoring does not take that long. I found that it was only when I either tried to do too many of the scoring sessions in a row, or had to break up the scoring of one instrument into a couple of sessions...that it seemed really difficult. Like I said before, it take time to go back over the criteria and all the standards that we agreed on. All said, though, I think that the process is not too bad – I could see schools doing this with their programs.

Both R1 and R2 indicated that the time taken to score the materials was inordinately long. Both recognised the value of the process ("it really gives you a good idea of not only what the curriculum is trying to do, but also the extent to which the designers of the material actually understand what it is they are trying to achieve" [R1]), yet indicated that it required an amount of time and effort that most teachers and administrators would not be able or willing to give. R2 captured the idea well:

> We are talking about twelve hours just to align three terms of work in one subject. I can't see an administrator or teacher being able to devote enough time to align all the courses, particularly if you expect them to do a good job. At the end of the day I am just stuffed, and I found myself reading for twenty minutes, then having to go back over it because I wasn't concentrating.

However, all reviewers indicated ways that they found during the process to more easily manage the work. R2 expressed similar opinions to R3 about the need to properly separate the scoring sessions, as "[the sessions] can be brutal if you do [the scoring] all at the same time." However, R2 believed that the biggest problem was the time taken to carefully read through all of the materials, keeping the criteria in mind:

> It's just that the amount of worksheets and assessments and experiments and notes, it is just a huge amount of work to read. And when you factor in that you have to read them and keep the standards and criteria in mind…it really drains you.

Overall, although the information that was provided was thought to be worthwhile, with the exception of the Krathwohl (2002) scaffold, discussed earlier, the process in its entirety is too unwieldy and time consuming to be practical in a school setting. However, with changes made to the number of criteria addressed as part of the alignment process, two of the reviewers felt that the alignment model was sustainable for a member of the teaching staff, given that it was conducted only annually.

How effective was the training session used to prepare for the scoring of the materials?

Although not included in the initial semi-structured interview questions, the frequency with which the training sessions was referred to, and the impact the training had on the eventual scoring of materials meant that it needed to be addressed in any consideration of the effectiveness of the program.

The training program consisted of two one hour sessions. The first was to familiarise the participants with the criteria themselves, and the second was to cross-mark a selection of the assessment tasks using the criteria. The participants appreciated the ability to communicate with other reviewers to help make a decision on some of the materials: "To come back and talk over a difficult piece was helpful, and I came away with a much better understanding of what I needed to do." [R3]

The participants indicated that the training program provided adequate guidance for some of the elements of the scoring procedure, but had some glaring omissions. R3 indicated the frustration at some elements of the training sessions:

> The sessions introduced us to the criteria, and in that sense they were okay. But, when we went back to score some other materials, we found that they didn't match pieces of work that we had practiced scoring, and I know I couldn't get a handle on where to actually score them. We really needed to see the application

> of the criteria to a greater range of tasks….in particular, the experiments I found difficult to score.

The lack of focus on the experiment materials during the training session seemed to be problematic for all the participants. The other omission from the training was the scoring of the curriculum materials in their entirety. R1 found "we hadn't made any agreement about how we should weight materials. I just plucked a number out that seemed to fit with the other scores I had given. It was not at all reliable. A waste of time really."

With an adjustment to the training schedule and focus, all participants agreed that it would be worthwhile. R2 comments:

> Keep the two sessions, and the first one in particular, with the introduction to the scoring. We just need to make sure that we have a proper understanding of what the standards are for each of the criteria. Then it would be much more effective.

Evaluation of the Model

Evaluation of the effectiveness of the analysis methodology used in this study was made using information from both the inter-rater reliability data and from the semi-structured interviews. Overall, the participants indicated that the evaluation method was successful in that it developed the type of information that would be useful for schools as they tried to align their programs, particularly in light of the focus on external testing. With some changes to the training program and the scoring instruments used, the participants believed that the program could genuinely be used in schools to determine the degree of alignment.

When considering the effectiveness of the program, two factors were taken into consideration:

1. The amount of time required to review the curriculum of a program.
2. The applicability, reliability and ease of use of each criterion.

Amount of time required to review the curriculum using the alignment methodology

The reviewers indicated that the biggest obstacle for this methodology to overcome is the amount of time taken to perform the analysis. Although the process would become faster as the participants became more experienced in the use of the criteria, spending in excess of five hours for a semester long course is prohibitive in a school setting. The time spent on the review needs to be reduced significantly to make it more manageable for teachers and administrators to use. If the program took between three and four hours to complete, then:

> it would be far more worthwhile. Obviously you couldn't [perform the scoring] that quickly unless you were an experienced teacher in that area, but I think that if you had the right people doing the [scoring] then it is certainly possible. I was doing aspects of it much faster at the end than at the beginning. [R3]

This reduction in time could be accomplished by altering several outputs of the process. The first is by experience; as the reviewers become more familiar with the criteria then the time spent reviewing materials would decrease. Secondly, the number of instruments used could be decreased, so that only those that are deemed most valid and reliable would be retained. Finally, the training program could be adjusted to make the scoring more efficient, and give stronger guidelines about how to perform the analysis. The final two conditions, alterations to the instruments and the training program, will be discussed further on page 97.

Applicability, reliability and ease of use of each criterion

For an evaluation model such as this to be successful, it is important that the scores made by the reviewers are reasonably consistent. If there is great variance in the scores achieved by reviewers, then this suggests a low reliability of that particular criterion. This lack of reliability can stem from several sources, including insufficient training of participants and lack of clarity in the wording of the criteria.

The fact that the kappa co-efficient was generally high meant that the judgements made by the reviewers were typically reliable. In particular, the Webb (1997) and Kesidou and Rosemann (2002) kappa co-efficients were quite high, with only the Year 9 materials scoring lower than would be anticipated for an aligned program. As the participants became more familiar with the

application of the criteria, the judgements themselves should become more accurate. However, the lack of consistent scoring in several areas leads to concern about both the training program and the criteria.

The training program, as discussed above, was adequate for most areas of the analysis, but had significant gaps in the understanding required for reliable judgements about materials and assessment. Reviewers found it difficult to score materials in formats with which they were unfamiliar – no real direction had been given for the scoring of entire sets of materials, or with assessment tasks which differed significantly from those used during the training session.

According to the reviewers, several changes could be implemented to improve the applicability of the training program. First, the training sessions should have materials which are deliberately selected to be scored by all participants during the session that were representative of all of the assessment materials present in a course. In particular, the participants indicated that the experimental materials needed a significant amount of time. Although there was an acknowledgement that materials will differ throughout the course, making sure that the participants had an opportunity to score a material with some similarity to the assessments on the course would make the process more effective. The Webb model would be particularly improved by this change, due to the greater complexity of its criteria.

Second, the program should:

> Include some guidance about how exactly you should spend the time. How long to [perform the analysis] in one stretch, and how best to get yourself organised. It was too easy to get lost and waste time, and [the time] could have been saved by us not having to find out by trial and error. [R2]

By taking the time to instruct the participants to complete one set of analysis in one sitting, and not try to analyse a set of curriculum materials/assessment program against all of the instruments at the same time, the time taken to score the materials could be reduced by as much as one quarter. This should improve both the reliability of the scoring performed and the total time taken to analyse a course.

The instruments used in the analysis were generally reliable according to the inter-reliability discussed earlier. The participants indicated that, providing that the training program was adequate, the scoring criteria in most areas could be confidently used to assess the alignment of the intended outcomes, curriculum materials and assessment of a science program. However, two changes were suggested to the criteria themselves.

First, the participants identified the Categorical Concurrence criterion from the Webb (1997) model as being particularly difficult to use effectively. "The wording actually makes it difficult to understand, and I sometimes had to go back through the [materials] I had marked before to find one that was similar to the one I was actually marking so that I could get a score." [R1] By rephrasing the Categorical Concurrence criterion it could be easier to identify levels of alignment, and hence improve the reliability.

The second change suggested by the participants involves the elimination of the Chinn and Malhotra (2002)/Krathwohl (2002) instrument. There are two obstacles to prevent the methodology from being easily used in schools: the time taken to make an assessment of the alignment, and the sheer volume of numerical data produced by the instruments. The elimination of the Chinn and Malhotra (2002)/Krathwohl (2002) instrument would go some way to addressing these two problems. All participants mentioned the difficulty in using the criteria associated with the Chinn and Malhotra (2002)/Krathwohl (2002) instrument, and two questioned the usefulness of the analysis itself, indicating that the information gained would not necessarily be used when the realignment of the curriculum occurred. As the instrument takes a significant amount of time to score assessments on, its elimination from the methodology would make the process more time efficient, and prevent the participants from being "drowned in data" [R3].

Summary

This chapter was divided into four sections. It began with an explanation of inter-rater reliability, while the second considered the reliability data. This analysis showed that the scoring of curriculum materials in the Year 7 program was considered reliable, yet the scoring of the Year 9 program was significantly less so. A hypothesis was put forward that this discrepancy could be due to differences in the curriculum materials, or due to identified flaws in the training provided for the reviewers.

The third section provided a brief overview of the semi-structured interviews before considering the impact of participant time and criterion effectiveness on the program. It addressed reviewer concern that the time required to score the courses (in excess of eight hours) was prohibitive for most teachers and administrators, and also doubts about the sheer volume of data produced by the instruments were raised. The data literacy demand on reviewers was a more significant factor than had been predicted. In particular, this section raised reservations about the practicality of the Krathwohl (2002) and Chinn and Malhotra (2002) instruments, given that reviewers felt that these instruments provided little useful data.

Finally, the degree to which the curriculum evaluation model is effective was explored in the fourth and final section, coming to the conclusion that, with the implementation of a more effective training program and elimination of two of the instruments, the model was useful and effective for determining the alignment of curriculum and assessment materials with intended goals.

The next chapter discusses the effectiveness of the case study science curriculum and scoring methodology by comparing the findings with established research.

CHAPTER SIX: DISCUSSION OF THE EFFECTIVENESS OF THE CURRICULUM AND AN EVALUATION OF THE SCORING MODEL

Introduction

This chapter summarises the research findings in relation to the research questions and compares them to the associated literature. The first section provides a brief overview of the research, including the study's major aims and objectives. The following section describes the alignment of the curriculum and assessments with the stated goals of the program, and the final section describes the major findings related to the evaluation of the scoring model and its implementation.

Overview

The aim of this research was to develop, implement and evaluate a method for evaluating the alignment of intended outcomes, curriculum materials and assessment in a Middle School science program.

The evaluation model was developed from the literature, and then curriculum materials, assessment instruments and intended outcomes from a Year 7 and a Year 9 program were analysed using the model. The model was implemented by three teacher-administrators at the rural case study Middle School, and then qualitative and quantitative data were used to evaluate the degree of alignment of the materials relating to the science program. The second set of data was obtained by using semi-structured interviews with the three reviewers.

Effectiveness of the Curriculum

The goals of the Middle School science program are to develop students' scientific literacy, including an understanding of how scientific research is conducted in the real world, that is, its epistemological goals. Overall, although there is a significant degree of alignment in the intended outcomes, curriculum materials and assessment tasks, there are enough instances of misalignment to partially explain the low levels of improvement of students' scientific literacy in the case study Middle School.

In general, the degree of alignment of the curriculum materials was positive. All but three of the criteria across the two year levels showed a mean score greater than 2, indicating that, in general, the curriculum materials are well-aligned with the intended goals of the course, and are consistent across year levels. It is hypothesised that this consistency in the formatting of curriculum materials allows students to more readily identify the purpose of the materials, and how one idea and skill links to another. The inclusion of the intended learning outcomes of the task with associated success criteria enables students to better able to engage in the learning process by understanding and utilising the metalanguage of both science and education (Mortimer & Scott, 2003).

The numbers of individual curriculum materials which are accurately tied to the intended learning goals was lacking in both year levels; this was clearly highlighted in both the Kesidou and Rosemann (2002) analysis and in the semi-structured interviews. The importance of a large number of curriculum materials which are strongly aligned to the intended goals of the program cannot be overstated. City, Elmore, Fiarman and Teital (2009) describe the importance of aligned curriculum materials referred to as instructional tasks:

> What determines what students know and are able to do is not what the curriculum says they are supposed to do, or even what the teacher thinks he or she is asking the students to do. What predicts performance is *what students are actually doing*. (p.30, City et al.'s emphasis)

This notion of the importance of curriculum materials is also underscored by Black and Wiliam (1998b), who indicate that curriculum materials "have to be justified in terms of the learning aims that they serve, and they can only work well if opportunities for pupils to communicate their evolving understanding are built into the planning." (p.10) Curriculum materials are the vehicles through which the students develop and demonstrate their understanding, and so they must be adequately linked to intended outcomes.

The lack of curriculum materials addressing the ELOs presents significant difficulties for the faithful implementation of the curriculum. When teachers are required to produce their own materials, often without models to copy and adjust, the fidelity of the course is compromised. The curriculum is only as strong as its ability to be faithfully implemented in the classroom; even though this study has focused on the intended curriculum, the significant gaps in curriculum

resources would indicate that it would be difficult for an individual teacher in this school to be able to represent the curriculum faithfully, despite their best intentions. Consequently, it is difficult to imagine that students across all classes would be making significant progress when the implementation of the course is likely to be vastly different between classrooms and between year levels. This difference in the implemented curriculum from one classroom to another may help to explain why some classes, in particular, feature students who are making less than optimal progress according to the ICAS assessments.

The materials which were most well aligned required students to practise the skills and demonstrate the knowledge that they would require to successfully complete the intended learning of the course. Models of demonstrably effective curriculum materials, such as those developed by Adey and Shayer (1990), could be used to develop a greater number of materials which accurately align with the intended outcomes of the course. In particular, the consistent format of lessons and materials, where students carry out experiments in which the analysis of results produces conclusions which conflict with the mental models they have developed, provide opportunities to both learn skills of drawing conclusions but also of redrafting and refining hypotheses based on data. It must be noted, however, that the lessons of Adey and Shayer, and any developed in their image, are not intended to teach skills of investigation design, so lessons which do address the design aspect of the scientific process would need to diverge from this model.

The alignment model indicated that the assessment tasks used to assess student progress were, generally, also closely aligned with the intended goals. Some tasks, however, were more representative of authentic science inquiry than others and assessed scientific literacy with greater reliability and validity than other tasks in the same program. Although the assessment programs as a whole would provide the information necessary to track student progress in scientific literacy over time, the interview responses indicate that improvements could be made to several facets of the assessment program.

The number of tasks in each year level could be altered depending on the need for feedback to students on the development of their science literacy skills. At Year 7 the assessment schedule consists of many tasks, and it was suggested by some respondents in the semi-structured interviews that there could be a reduction in the number of tasks. Although the reduction in the number of

tasks is an option, most of the literature on formative feedback would indicate that the feedback cycle works most effectively when tasks are shorter and more frequent than longer, less frequent pieces (Black & Wiliams, 1998b; Broadfoot, 1996; OECD, 2005; Wiliam, 2006). This is best summarised by Black & Wiliams (1998b), who indicate that "(i)t is better to have frequent, short tests than infrequent and longer ones" (p. 12). A reduction in the number of summative tasks used to generate a grade with the introduction of more frequent formative tasks may be a worthwhile compromise.

When examining those tasks that did have a strong alignment to the intended goals of the program, the aligned tasks shared a number of general features. Four features were identified in assessment tasks which were closely aligned to the goals of the science program. First, the tasks are explicitly linked to scaffolded instruction that describes to the student the learning path that needs to occur, which appears as a continuum in the case study science program, and provides them with the necessary skills to take the next step in their learning. Multiple studies have shown that assessment is only really useful when they are accurately linked to the path of intended learning for the students (Black and Wiliams, 1998a, 1998b; Hattie, 2003; OECD, 2005; Rothman, 2006).

Second, these aligned tasks were identified as open ended in order to provide students with more freedom to generate a response which utilised a variety of skills. The opportunity for students to construct and communicate ideas as part of the task itself aligns more closely with the goals of the case study science program. Although speaking primarily about continua in language studies, the view of Masters and Forster (2000) is applicable:

> Open ended tasks which permit different levels of response can also be useful for estimating students' achievement levels along a continuum. For example, the same essay prompt usually can be administered to students with very little writing ability, and then performances on several prompts can be used to locate students along a continuum of increasing writing competence. (p. 7)

Also, the tasks requiring students to relate experimental ideas to contexts showed a greater alignment with the epistemological goals of the program. Tasks which are more closely related to authentic science inquiry seem to lend themselves better to both more effective learning and meaning-making. This view is consistent with those of Chinn and Malhotra (2002), as well as the Australian Curriculum (2011) developed for Science, which devotes a particularly large component

of curriculum space (and hence teaching time) to the development of inquiry skills (ACARA, 2011).

Fourth, the tasks should be deliberately designed with the continua in mind; they require application of a number of skills that increase in difficulty. The tasks need to be at a difficulty that allows both the least progressed student to give a response and the highest performing students to display the full extent of their understanding. This view is well-supported by the literature. Masters (2001) states that for assessment to be truly valuable, it must capture the level of understanding of students in the full extent of the range. To test a narrow range of comprehension and skills, which was sometimes the case in the evaluated science assessment program, means that the level of comprehension of many students will not be adequately measured, and this would make it difficult to adjust teaching strategies in order to help each child improve.

Evaluation of the Model

The second aim of the research project was to determine how effective the curriculum evaluation model developed and implemented in this study was in evaluating the alignment of intended outcomes, curriculum materials and assessment. The responses from the semi-structured interviews showed that, although some aspects of the scoring model need altering, it was generally successful in that it developed the type of information that would be useful for schools as they tried to align their programs due to the focus on external testing. The limitations in the effectiveness of the model stemmed from both the sheer volume of data generated through the evaluation, as well as the usefulness of the data produced.

Effectiveness of the scoring method

The scoring instruments used in the analysis were considered reliable, a view based upon the response of the reviewers and the relatively strong Fleiss' (1971) kappa co-efficient scores. The responses from the interviews indicated that, providing that the associated training programs were adequate, the scoring criteria in most areas could be confidently used to accurately assess the alignment of the intended outcomes, curriculum materials and assessment of a science program. The Webb (1997) scoring system was considered to be reasonably easy to apply, with the exception

of two of the criteria. The reviewers encountered a similar problem to that recorded by Martone and Sireci (2009) when using the analysis: by averaging reviewers' ratings across a large number of assessment tasks, or a broad set of criteria against which the task is assessed, the degree of alignment score can be inflated, and can mask the different views of the reviewers.

The Categorical Concurrence criterion from the Webb (1997) model was particularly difficult to use effectively. "The wording actually makes it difficult to understand, and I sometimes had to go back through the [materials] I had marked before to find one that was similar to the one I was actually marking so that I could get a score." [R1] By rephrasing the Categorical Concurrence criterion it could be easier to identify levels of alignment, and hence improve the reliability. This contrasts with the analysis of the reviewer responses made by Webb himself, in his 1999 study. He indicates that this criterion was consistently scored, while identifying the weakest as the Depth of Knowledge Consistency and Range of Knowledge criteria. "If [an intended outcome was] very broadly stated, it was still considered assessed if it had an item matched to it, regardless of what else within that [outcome] was not measured" (p.18).

The major weakness in the scoring method appears to be the Krathwohl (2002) section of the instrument, used to measure the Cognitive Process dimensions. Reviewers indicated that, while this instrument collects a large amount of data, the data collected does not provide useful information with which to alter the curriculum in order to bring it closer to its epistemological goals. This is particularly true when the nature of the reference scale is based on a progression of applications of knowledge and skill. When the time taken to score these materials according to the Krathwohl (2002) scale is considered, the value of the information in determining the alignment of this particular case study course is questionable. It could be argued, however, that when the scoring method was used to score programs using a scale less dependent on a developmental paradigm (perhaps norm- or percentage achievement-based), this element of the model may be more useful.

Implementation of the scoring method in schools

Two significant challenges exist for the implementation of this evaluation model in schools. The first is the amount of time required to complete the scoring for a particular curriculum. Although the time spent on each of the criteria decreases as the reviewer becomes more familiar with the

process, the entire process requires in excess of five hours for a semester long course (19 – 22 weeks). This is generally prohibitive in a school setting. The time spent on the review needs to be reduced significantly to make it more manageable for teachers and administrators to use. If the program took between three and four hours to complete then it becomes more manageable. As the Cognitive Process dimensions do not provide data which are particularly useful in determining alignment, the potential exists to remove the scoring of these dimensions, which would greatly reduce the amount of time spent scoring.

The second challenge is developing the expertise required to evaluate the resources. The reviewers must be subject matter experts, knowledgeable in the pedagogical implications of a particular set of concepts and skills, and have a solid grasp of the underlying theory that the intended outcomes requires. The reviewers also need a strong understanding of the intended outcomes of the course. Many schools have teachers with the requisite subject matter expertise, but the ability of a school to conduct an evaluation will hinge on the quality of the training program. The amount of time spent outlining the intended learning of the course in the training program was greater than anticipated, particularly considering the reviewers were all employees of the case study Middle School. This observation matches with those of Sireci (1998), who indicates that, for measures of content validity and alignment, it is important for highly knowledgeable subject matter experts to be involved. In addition, he states that it is critical for the reviewers to be familiar with the standards against which the materials are going to be measured. Inconsistent interpretation of standards, particularly those with broad phrases, across the reviewers conducting an alignment study can cause error in expert judgement (La Marca et al., 2000; Webb, 1997).

One way to improve the quality of reviewers' knowledge of the process is the implementation of a comprehensive training program. In the case of this project, each of the reviewers indicated that the training program, although quite helpful in conducting the review, had significant gaps in the development of understanding required for reliable judgements about materials and assessment. Participants found it difficult to score materials in formats with which they were unfamiliar – no real direction had been given for the scoring of entire sets of materials, or with assessment tasks which differed significantly from those used during the training session. This matches the problems identified by Webb (1999),who indicated that a large number of materials of different types need to be scored in the training sessions (certainly more than he had intended) and the standards (intended

outcomes) needed to be put into context so that reviewers knew the purpose of the standards. Although the second of these was not a problem encountered in this project, the first certainly matches comments made by reviewers in the semi-structured interviews. The selection of materials to be scored by all participants during the session that were representative of all of the assessment materials present in a course would make a significant difference to the effectiveness of both the training program and the scoring itself. With the improvements outlined in the review of the training program above, teachers operating in schools can gain the requisite expertise in order to accurately and reliably score according to this method.

Summary

This chapter summarised the research findings in relation to the research questions and compared and contrasted them to the relevant literature. The first section described a broad overview of the aims of the research. The next section indicated that the case study curriculum was generally aligned to the intended goals of the course. However, the scarcity of curriculum materials at both year levels was identified as a particular concern, as it is difficult to adequately implement a curriculum faithfully when supporting materials are lacking. A discrepancy was described between the views of the reviewers and that found in the literature concerning the ideal frequency of assessment. Reviewers recommended that the number of assessment tasks in the curriculum be reduced, which contrasts sharply with the views of the literature, which recommends more frequent and shorter assessment events. This section also described the common features of assessment and curriculum materials which were aligned: that they should be open-ended in nature, explicitly linked to the scaffolded instruction and the related curriculum materials, match the epistemological goals of the program by relating directly to the elements of real-world scientific research, and designed to directly assess the intended goals of the course. Following this was a discussion of the potential changes which could be made to the case study program to improve the alignment. The inclusion of more frequent formative tasks would allow scope for teachers to adjust their instruction to better meet the needs of students within the group. Another significant change would include the development of a greater number and quality of curriculum materials that were explicitly linked to scientific literacy, and consistency in formatting of these materials.

The final section described the effectiveness of the scoring method and its implementation in schools. It described the significant training requirements for accurate use of the scoring criteria that were not featured in this particular research (greater range of assessment materials scored in the training sessions, greater frequency of sessions), and outlined the importance of reviewers understanding both the subject matter and intended outcomes/standards in the reliable implementation of the alignment scoring program. This section also described several areas in which the criteria used to judge the level of alignment were not effective, particularly the criteria for the alignment of epistemological goals and the Categorical Concurrence criterion in the Webb (1997) analysis.

The final chapter provides major recommendations which have emerged from the research and also concludes the current study.

CHAPTER SEVEN: CONCLUSION AND IMPLICATIONS

Introduction

This final chapter is divided into four distinct sections. The first section describes an overview of the study, while the second provides a conclusion to the research. The third section analyses how the study contributes to the body of education research knowledge. Finally, implications of the findings of the study are discussed, and future research considered.

Overview

The purpose of this study was to develop and evaluate a method by which the alignment of curriculum materials and assessment tasks with the intended goals of a Middle School science program could be evaluated. An evaluation model was then developed and implemented to ascertain the degree of alignment of the case study science program, and to describe how the materials were aligned to the intended goals by identifying the commonalities of aligned curriculum materials and assessment tasks.

The increased focus on the inquiry skills that contribute to scientific literacy in the Australian Curriculum for Science (ACARA, 2011), means that a large number of schools will be required to change the pedagogical approach to teaching science. The ability to use a model to evaluate and then adjust materials to better suit the intentions of the curriculum would be useful to many schools.

The conceptual framework of the study considered the various definitions of scientific literacy and then linked them to methods by which curricula are ideally developed through Constructive Alignment and Backwards Design. The literature emphasised the importance of the alignment of a program's intended goals, curriculum materials and assessment tasks (Biggs, 1999; La Marca et al., 2000; Ramsden, 1992; Tytler, 1949; Wiggins & McTighe, 2001) and presented some methods by which the alignment could be evaluated. These scoring methods were then used to develop the alignment scoring method applied in this analysis.

Three reviewers evaluated curriculum materials and assessment tasks from a Middle School science program, producing quantitative scores that revealed the degree to which the science program had achieved alignment. After the completion of the scoring process, the reviewers participated in semi-structured interviews, discussing the implementation of the evaluation model. The interview responses were transcribed and then signed-off by the reviewers. From the interviews and the scoring, the features of aligned curriculum and the effectiveness of the evaluation model were determined.

Conclusions

This study's research questions provide the framework on which the conclusions of this study are based.

To what extent are the intended outcomes, curriculum and assessment in this Middle School science curriculum constructively aligned?

The current study established that the Middle School science program in the case study had general alignment of the intended learning outcomes, curriculum materials and assessment. The reviewers' scores generally indicated good alignment through the material that had been developed and implemented in the classroom. However, there were a number of materials and tasks which were not adequately aligned, and there was a lack of curriculum materials to support some of the ELOs addressed by the curriculum. This lack of materials makes it more difficult to maintain fidelity of implementation for teachers as they attempt to implement the intended curriculum.

The features of the aligned assessment tasks generally matched those identified in the literature: open-ended tasks which are explicitly linked to the scaffolded instruction, with assessment that directly assesses the intended goals of the course. The inclusion of these features into a greater proportion of the tasks, as well as the related curriculum materials, would provide a basis for improving the assessment program.

How effective is the curriculum evaluation model developed and implemented in this study for evaluating the alignment of intended outcomes, curriculum materials and assessment?

The evaluation model was deemed to be effective in determining the alignment of the science program. However, several evaluation criteria were identified as problematic and there were concerns about the amount of time required to score the materials. It was also noted that the training provided to the reviewers was valuable but inadequate to ensure consistency in judgement for all types of materials.

Contribution to Knowledge

With the implementation of the mandated Australian Curriculum for Science which has a much greater focus on the development of inquiry skills that contribute to scientific literacy, there is a need for substantial change in both the science curricula being offered at many schools and the pedagogy of the teachers implementing the curricula. In addition, there is greater recognition of the need for the curriculum and intended goals to be aligned; this is driven partially by economics (government funding will be tied to the implementation of the Australian Curriculum), and partially by the increased transparency of student performance through the MySchool website.

Much of the current literature emphasises the need for the curriculum to be focused on intended goals of the course, and yet, if the case study school is representative of the overall school system, the knowledge and understanding of what that looks like is still underdeveloped. There is an absence in the literature, particularly for Australian schools, of methods by which by teachers working within schools can review science curricula for alignment. The ability of schools to be able to independently analyse their curriculum is not only beneficial for the learning of their students, but also provides an excellent professional development exercise. The alignment model implemented and evaluated in this case study addresses the need to independently analyse the curriculum by providing a system by which science teachers serving in a particular school can review their curriculum in light of the intended goals of the course (i.e. Australian Curriculum) without requiring external auditors or experts.

Implications

The findings of the current study, in conjunction with the reviewed literature, have resulted in the series of implications presented below. These are divided into three categories: implications for future research, implications for the case study school and implications for the refinement of the evaluation model.

Implications for future research

The findings of the current research have shown that the case study science program is partially aligned to the intended goals of the program. Also, the alignment scoring model is effective at determining alignment in Middle School science programs. However, the alignment scoring program might need to be altered in order for it to be effectively implemented in schools or small school systems by existing staff. In particular, the number of criteria scored should be reduced in order to selectively collate the information most pertinent to determining alignment.

Implications relating to further research include four specific investigations. First, it would be useful to expand the use of the alignment framework to examine other science programs in the case study school. This may indicate whether the findings of the current study, based on two of the 20 science programs taught in the case study Middle School, accurately represents the alignment of the entire science program.

Second, the application of the alignment scale in its refined version (based on the recommendations of this study) to Middle School science programs that aim to develop scientific knowledge rather than inquiry skills might allow comparisons to be drawn between the effectiveness of the alignment program in theory-based and skill-based courses.

Third, further research could examine the effectiveness of various training programs on the reliability of reviewer judgement. Both the literature and the findings of this study indicate that the quality of the training program has a marked effect on the effective implementation of the scoring process. Research could be undertaken to identify the features of an effective training program which could be applied to a variety of alignment methodologies.

Finally, several of the reviewers also discussed the fact that the process of evaluating alignment itself also presents a significant professional learning exercise. Their comments indicate that, to score materials accurately, teachers need to have a strong understanding of what the science course is attempting to achieve and how the materials should be structured to ensure the alignment. The opportunity exists for further research into the professional learning aspects of the training and scoring program, and whether it results in a stronger understanding of the pedagogy underlying the science program.

Implications for the case study middle school and its science program

The current research established that the Middle School science program varied in the degree of alignment of curriculum materials and assessments with the program's intended goals. Materials and assessments which were significantly aligned shared characteristics that set them apart from the materials that were not aligned. In most cases, these alignment characteristics matched the features of effective tasks identified in the literature. Based on these traits, several recommendations can be made which should result in a science program that has a greater degree of alignment to its intended goals.

First, the development of a greater range of curriculum materials which are directly tied to the intended learning goals of the program should be considered for all of the science programs in the case study school. It is likely that the materials for the rest of the science programs will feature similar levels of alignment, but it would be necessary to check whether there are differences in the alignment of materials as they pertain to the older year levels. Although the current project only examined a small subset of the number of science programs available at the Middle School, they do present an accurate representation of the rest of the middle years program. Curriculum materials which are directly linked to the intended goals of the program enhance the learning, and give the students an opportunity to develop and practise the skills that the course aims to develop (Black & Wiliam, 1998b; City, Elmore, Fiarman, & Tietel, 2009; Mortimer & Scott, 2003; Wiggins & McTighe, 2001).

Although the development of this type of curriculum material requires both expertise and time, the fact that examples of these materials can be found already in the curriculum indicate that this Middle School has the capacity to construct adequate materials, and should rely less on inadequate or less-specific commercial materials.

Second, the assessment tasks, which make up the assessment program, should be adjusted to align with the findings of the current research and the features of effectively aligned assessment tasks identified in the literature. Each of the assessment tasks should be rewritten in order to feature the four major elements of the assessment alignment identified in this research: they should be open-ended in nature; explicitly linked to the scaffolded instruction and the related curriculum materials; match the epistemological goals of the program by relating directly to the elements of real-world scientific research; and designed to directly assess the intended goals of the course, allowing the students to demonstrate a wide range of achievement of a particular skill.

These changes would result in an assessment program that accurately measures the degree to which the students have developed their inquiry skills. Although the current case study assessment program as a whole adequately tracks student performance against the continua, the weakness of several tasks within the program highlights the need to ensure that each of the assessment pieces is carefully considered as to what it measures and to what extent it measures the level of performance of each student in the cohort.

Implications for the refinement of the alignment scoring method

Some of the findings of this research indicate that the alignment scoring method used in this study could be refined to both improve the accuracy of the scoring and the ability for individual schools or small school systems to use the method *in situ*. Several recommendations can be considered for revising and improving the evaluation methods.

The model could be improved by reducing the scoring of the epistemological goals (Chinn & Malhotra, 2002), and the removal of the alignment scoring associated with the Cognitive Process dimensions of the program provided by Krathwohl (2002). In the semi-structured interviews the reviewers indicated that the information gained from the epistemological analysis took a large

amount of time, but a similar effect could have been gained with a quick checklist that considered the assessment program as a whole. The implementation of this checklist would reduce the amount of time required to complete the analysis, while still providing the required data.

The reviewers also identified that the intended goals of the program (as defined by the Essential Learning Outcomes) are already measured on a continuum of increasing competency and sophistication of scientific literacy application rendering the Cognitive Process dimensions redundant. These two dimensions took a significant amount of time to score, resulting in a large amount of data that had little productive use. The elimination of this aspect of the method, particularly when dealing with a curriculum designed to improve scientific literacy, should enable the alignment scoring method to be completed in a shorter time.

It should be noted, however, that there may be situations or curricula in which the Cognitive Process dimension may be useful. It is anticipated that knowledge-based curricula, which do not use developmental scales, could make some use of this element of the dimension to identify the degree to which different types of cognitive process are addressed in the intended curriculum.

Some of the criteria used to score the curriculum materials and assessment could be rephrased in order to improve the reliability of reviewers' judgements. The reviewers indicated that several of the criteria were worded in such a way that it made it quite difficult to accurately differentiate between different levels of attainment. In particular, the Categorical Concurrence indicator from the Webb (1997) model was difficult to apply. Rewording of the criteria to remove some of the broader terms and greater referencing of the specifics of the task to be judged could make this indicator more meaningful for the reviewers.

The program should include an interview component as part of the scoring process. The information gained during the interviews helps to elaborate on some of the detail of the scoring. The provision of explanations for some scores, along with discussion of the features of aligned materials, would enhance the understanding of the data gained from the analysis.

The reviewer training program associated with the alignment scoring method needs to be properly delineated, with careful consideration given to the number of hours of training provided, the types

of curriculum material used to practise scoring and the amount of contact between the reviewers as they score the materials. The critical nature of training programs in alignment analysis has been observed by both Webb (1997) and Martone and Sireci (2009). During this particular research, the training was not planned and implemented as carefully as it should have been, and it resulted in some confusion on the part of the reviewers. The careful prescription of the training program would be particularly pertinent if the scoring method was to be used by schools and systems not included in this study.

Several changes should be made to the way in which the alignment scoring method training program was constructed: scoring a more representative range of materials during the training program, and giving guidelines on the amount of time which should be spent scoring particular materials. The training program needs to provide guidance in scoring all types of material present in the curriculum. This research project showed that the reviewers found it difficult to score materials with which they had not had any experience in the training program. Expanding the range of materials scored during training should help to improve the reliability of the scores given to curriculum and assessment materials which are different to those that were typical of the research project's training sessions.

In addition, guidelines about the amount of time which should be spent on each particular element of the curriculum would be useful. As the reviewers became more familiar with the criteria the scoring process time was accelerated; however, the reviewers indicated that they had spent an inordinate amount of time on some elements of the scoring system at the expense of others. An indication of how much time should be spent on each element of the framework would enable the reviewers to be more efficient in their work.

Wider Implications

Ultimately, the effectiveness of a curriculum is limited by the quality of the curriculum materials and the method by which they are implemented in the classroom. City et al. (2009) describe the essential components of teaching and learning as the instructional core: interactions between teachers and students in the presence of content. The importance of the intended curriculum content cannot be overstated; it determines what the students are learning, how it is being taught and what aspects are assessed. Without aligned curriculum resources, students are exposed to a disjointed and disparate curriculum that is inconsistently applied from one classroom to the next, wasting

valuable time and limiting the learning of the students. For middle school science curricula to improve, there should be alignment in planning and delivery with the intended goals and allowance for assessment tasks which are open-ended and directly related to these intended goals.

The purpose of a school or schooling system is to provide every student with the best educational opportunities. A feature of unaligned curriculum and assessment is that instruction and subsequent learning will vary greatly from one classroom to the next as the curriculum lacks the coherency to describe and influence instruction in the classroom. The idea that students enter a lottery in which their learning for a school year will be greatly influenced by the chance event of which class they are assigned to is unacceptable. Increased fidelity of implementation of the intended curriculum would reduce variation in instructional quality (particularly at the mediocre end of the spectrum) and students would be clear on exactly what they are supposed to be learning. As schools become more capable at reflecting on the alignment of their current offerings and altering them to reflect the features of aligned curriculum, the resultant learning of students should become more effective.

References

Australian Curriculum, Assessment and Reporting Authority (ACARA). (2009). *Curriculum Design Paper*. ACARA: Canberra.

Australian Curriculum, Assessment and Reporting Authority (ACARA). (2011). *Australian Curriculum: Science*. ACARA: Canberra.

Adey, P.L. & Shayer, M. (1990). Accelerating the development of formal thinking in middle and high school students. *Journal of Research in Science Teaching,* 3 (27), 267 - 285.

Adey, P.L. & Shayer, M. (2001). *Thinking through Science (3rd Edition).* Cheltenham: Nelson Thornes.

Aikenhead, G. (2006). *Science education for everyday life: Evidence based practice*. New York and London: Teachers College Press.

Anderson, L.W. (2002). Curricular Alignment: a re-examination. *Theory into Practice*, 4 (41), 255 – 261.

Anderson, L.W. (Ed.), Krathwohl, D.R. (Ed.), (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's Taxonomy of Educational Objectives.* New York: Longman.

Beane, J. (1993) Teachers of uncommon courage. In Stevenson, C. and Carr, J. F. (Ed.) *Integrated Studies in the Middle Grades.* New York: Teachers College Press.

Bell, J. (2005). *Doing your research project* (4th ed.). Berkshire England: Open University Press.

Bhola, D.S., Impara, J.C., & Buchendahl, C.W. (2003). Aligning tests with states' content standards: Methods and issues. *Educational Measurement: Issues and Practice*, (22) 3, 21 – 29.

Biggs, J.B. (1996). Enhancing teaching through constructive alignment. *Higher Education*, 32, 1-18.

Biggs, J.B. (1999). *Teaching for quality learning at university.* Buckingham: Open University Press.

Black, P. & Wiliam, D. (1998a). Assessment and Classroom Learning. *Assssment in Education (5)* 1, 7 – 74.

Black, P. & Wiliam, D. (1998b). *Inside the black box: Raising standards through classroom assessment.* London: NferNelson Publishing Company.

Blank, R.K., Porter, A.C., & Smithson, J.L. (2001). *New tools for analyzing teaching, curriculum and standards in mathematics and science.* Washington DC: Council of Chief State School Officers.

Boudett, K.P., City, E.A., & Murnane, R.J. (Eds.) (2005). *DataWise: A Step-by-Step Guide to Using Assessment Results to Improve Teaching and Learning.* Cambridge: Harvard Education Press.

Bransford, Brown & Cocking. (2000). *How People Learn: Expanded Edition.*. National Research Council.

Broadfoot, P. (1996). *Education, Assessment and Society.* Buckingham: Open University Press.

Burns, R. (2000). *Introduction to Research Methods.* London: Sage Publications.

Cavagnetto, A.R. (2010). Argument to Foster Scientific Literacy: A Review of Argument Interventions in K-12 Science Contexts. *Review of Educational Research,* 80, 336 – 371.

Carey, S., Evans, R., Honda, M., Jay, E. & Unger, S. (1989). 'An experiment is when you try it and see if it works'; a study of Grade 7 students' understanding of the construction of scientific knowledge. *International Journal of Science Education*, 11, 514 – 529.

Chiapetta, E., Sethna, G., & Fillman, D. (1993). Do middle school life science textbooks provide a balance of scientific literacy themes? *Journal of Research in Science Teaching*, 30, 787 – 797.

Chinn, C. A. & Malhotra, B. A. (2002). Epistemologically authentic inquiry in schools: A theoretical framework for evaluating inquiry tasks. *Science Education,* 86, 175 – 218.

City, E.A., Elmore, R.F, Fiarman, S.E., & Tietel, L. (2009). *Instructional Rounds in Education.* Cambridge, MA: Harvard Education Press.

Cornbleth, C. (1990). *Curriculum in context.* Basingstoke: Falmer Press.

Creswell, J.W. (2005). *Educational Research: Planning, Conducting and Evaluating Quantitative and Qualitative Research.* Merrill.

DeBoer, G.E. (1991). *A history of ideas in science education: Implications for practice.* New York: Teachers College Press.

Dochy, F.J.R.C. & McDowell, L. (1997). Assessment as a tool for learning. *Studies in Educational Evaluation,* (23) 4, 279 – 298.

Driver, R., Asoko, H., Leach, J., Mortimer, E. & Scott, P. (2004). Constructing scientific knowledge in the classroom. *Educational Researcher,* (23) 7, 5 – 12.

Eltinge, E. & Roberts, C. (1993). Linguistic content analysis: A method to measure science as inquiry in textbooks. *Journal of Research in Science Teaching,* 30, 65 – 83.

Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin,* 76 (5), 378 -382.

Gallagher, J.J. (1991). Prospective and practicing secondary teachers' knowledge and beliefs about the philosophy of science. *Science Education,* 75, 121 – 33.

Germann, P. J., Haskins, S. & Auls, S. (1996). Analysis of nine high school biology laboratory manuals: Promoting scientific inquiry. *Journal of Research in Science Teaching,* 33, 475 – 466.

Goodrum, D., Hackling, M. & Rennie, L. (2001). *The status of teaching and learning of science in Australian schools: a research report.* Canberra: DEET.

Grundy, S. (1987). *Curriculum: Product or Praxis*. Lewes: Falmer Press.

Hackling, M.V. & Prain, V. (2005). *Primary Connections Stage 2 Trial: Research Report.* Australian Academy of Science.

Hackling, M.V. & Prain, V. (2008). *Primary Connections Stage 3 Interim research and evaluation report 15: Impact of Primary Connections on students' science processes, literacies of science and attitudes towards science.* Australian Academy of Science.

Hattie, J. (2003). Teachers Make a Difference: What is the Research Evidence? Paper presented at ACER Research Conference *Building teacher quality: What does the research tell us?* 19 – 21 October, 2003, Melbourne.

Hodson, D. (1998). Towards a philosophically more valid science curriculum. *Science Education*, 72, 19 – 40.

International Competitions and Assessment for Schools. (2008). Case Study School Report, UNSW.

InterAcademies Panel. (2006). *Report of the Working Group on International Collaboration in the Evaluation of Inquiry-Based Science Education (IBSE) programs*. IAP: Washington.

Jiminez, M. P. A. (1994) Teaching evolution and natural selection: A look at textbooks and teachers. *Journal of Research in Science Education*, 31, 519 – 535.

Krippendorf, K. (1980). *Content Analysis.* London: Sage.

Kesidou, S. & Roseman, J. E. (2002) How well do middle school science programs measure up? Findings from Project 2061's curriculum review. *Journal of Research in Science Education*, 39, 522 - 549.

Krathwohl, D. R. (2002). A revision of Bloom's Taxonomy: An overview. *Theory into Practice*. (4) 41, p. 212-219.

Kuhn, D. & Phelps, E., (1982). The development of problem solving strategies. In H. Reece (Ed.), *Advances in Child Development and Behaviour, Volume 17.* New York: Academic Press.

La Marca, P.M., Redfield, D., Winter, P.C. & Despriet, L. (2000). *State standards and state assessment systems: A guide to alignment. Series on standards and assessment.* Washington, DC: Council of Chief State Offices.

Lederman, N.G., (2006). Students' and teachers' conceptions of the nature of science: A review of the research. *Journal of Research in Science Teaching. (29) 4*, 331 – 359.

Marsh, C. (1996). *Perspectives: Key concepts for understanding curriculum 1.* London: Fullman Press.

Martone, A. & Sireci, A.G. (2009). Evaluating alignment between curriculum, assessment and instruction. *Review of Educational Research. (79)* 4, 1332 – 1361.

Masters, G. N. (2001). Standards and assessment for students and teachers: A developmental paradigm. In Zbar, V. & MacKay, T. (ed.) *Leading the Education Debate.* Melbourne: ACER Press.

Masters, G. N. & Forster, M. (2000). *The Assessments We Need* . Camberwell: Australian Council for Educational Research Limited.

Millar, R. & Osborne, J. (1998). *Beyond 2000: science education for the future.* London: King's College London.

Mortimer, E. & Scott, P. (2003). *Making Meaning in secondary science classrooms.* Maidenhead, Philadelphia: Open University Press.

MySchool results for Ballarat Clarendon College. (2010). Retrieved July 27, 2011, from http://www.myschool.edu.au

National Curriculum Board (2008). *National Science Curriculum: Framing Paper*. Carlton South: National Curriculum Board.

National Research Council (NRC) (1996). *National Science Education Standards.* Washington, DC: National Academy Press.

Organisation for Economic Cooperation and Development (OECD). (2005). *Formative Assessment: Improving Learning in Secondary Classrooms.* Paris: OECD.

Organisation for Economic Cooperation and Development (OECD). (2006). *Assessing scientific, reading and mathematical literacy*. Paris: OECD.

Porter, A.C. & Smithson, J.L. (2001). *Defining, developing, and using curriculum indicators.* Philadelphia, PA: Consortium for Policy Research in Education.

Print, M. (1993). *Curriculum development and design (2nd Ed.).* St Leonards: Australia: Allen & Unwin.

Ramsden, P. (1992). *Learning to teach in higher education.* London: Routledge.

Richmond, W. K. (1971). *The school curriculum.* London: Methuen.

Roberts, D. A. (2007). Scientific literacy/Science literacy. In Abell, S.K. & Lederman, N. G. (Eds.) *Handbook of research on science education* (pp. 729 – 780). Mahwah, NJ: Lawrence Erlbaum Associates.

Ross, A. (2001). *Curriculum studies and critique.* London: Falmer.

Rothman, R., Slattery, J.B., Vranek, J.L., & Resnick, L.B. (2002). *Benchmarking and alignment of standards and testing* (CSE Technical Report No. CSE-TR-566). Los Angeles, CA: National Centre for Research on Evaluation, Standards, and Student Testing.

Rothman, R. (2006). (In)formative Assessments: New test and activities can help guide student learning. *Harvard Education Letter*, (22), 6.

Rothman, R. Slattery, J.B., Vranek, J.L., & Resnick, L.B. (2002). *Benchmarking and alignment of standards and testing.* Los Angeles, CA: National Centre for Research on Evaluation, Standards and Student Testing.

Sireci, S.G. (1998). Gathering and analyzing content validity data. *Educational Assessment, 5*, 299-321.

Shuell, T.J. (1986). Cognitive conceptions of learning. *Review of Educational Research*, 56, 411 - 436.

Stern, L. & Ahlgren, A. (2002) Analysis of students' assessments in middle school curriculum materials: Aiming precisely at benchmarks and standards. *Journal of Research in Science Education*, 39, 889 – 910.

Stringer, E. & Dwyer, R. (2005). *Action Research in Education.* Upper Saddle River, NY: Pearson Prentice Hall

Tyler, R.W. (1949). *Basic principles of curriculum and instruction.* Chicago: University of Chicago Press.

Tytler, R. (2007). *Reimagining science education: Educating students in science for Australia's future.* Camberwell: Australian Council for Educational Research.

Tytler, R., Waldrip, B. & Griffiths, M. (2004). Windows into practice: constructing effective science teaching and learning in a school change initiative. *International Journal of Science Education*, (26) 2,171 - 194.

United States Department of Education. (1999). *Peer review guidance for evaluating evidence of final assessments under Title I of the Elementary and Secondary Education Act*. Washington DC.

Victorian Curriculum and Assessment Authority (2000). *Curriculum and Standard Framework II.* Melbourne: VCAA.

Victorian Curriculum and Assessment Authority. (2005) *Victorian Essential Learning Standards Overview.* Melbourne: VCAA.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessment in mathematics and science education.* (Research Monograph No. 6). Washington: National Institute for Science Education Publications.

Wiersma, W., and Jurs, S.G. (2004). *Research Methods in Education.* Pears.

Wiggins, G. & McTighe, J. (2001). *Understanding by Design (2nd Ed.).* Alexandria, VA: Association for Supervision and Curriculum Development.

Wiliam, D. (2006). Assessment for Learning: Why, What, and How? *Orbit* 36 (2006), 2-6.

Yinn, R.K. (1984). *Case study research: Design and methods.* California: Sage.

Zimmerman, C. (2000). The development of scientific research skills. *Developmental Review*, 20, 99 – 149.

Appendices

Appendix A: Essential Learning Outcomes (ELOs) for the Middle School in the proposed case study.

ELO 1: Hypothesis and Contention

1.1 Generation of an Hypothesis/Contention

Is the ability of a student to use prior knowledge in order to make a prediction or begin investigation of a point of view about an issue. The hypothesis and contention should ideally be related to the focus question or aim, and should be supported with a brief outline of the reasoning behind it. Students can show understanding of this ELO through written work or through verbal responses.

1.2 Number and variety of hypotheses/claims/ideas

Is the ability of a student to generate a number of ideas surrounding a theme or problem. Ideally, the student should be generating large numbers of ideas which have relevance to the problem and have some variety in composition or approach. This aspect is normally identified as the "creative thinking" aspect.

ELO 2: Collecting and Evaluating Evidence

2.1 Evaluation of the reliability of data

Is the ability of a student to able to assess the reliability of the source of data in the investigation. This data can be sourced from an experiment or through primary and secondary sources. Students should be able to check sources for accuracy, either in controlling variables or through the veracity of the statements. They consider extraneous factors such as motivations for testimonies, whether the source is primary or secondary in nature and reproducibility of data.

2.2 Effectiveness of collection procedure

Is the ability of a student to tailor their collection of evidence to the Hypothesis or Research Question. The student should show no prejudgement or bias in their consideration of evidence, and should be fluent and efficient. The procedure should include appropriate strategies for gaining information, whether through a strong experimental method or a search strategy on the internet.

2.3 Procession of Data

Is the ability of a student to be able to both present a set of data. The presentation of the data involves processing the information into a suitable graph, table or similarly appropriate form. More advanceded students should be able to quantify the complex trends and patterns in the data.

2.4 Interpretation of Data

Is the ability of a student to be able to interpret data sets.. To interpret the data, the student will be searching for trends and patterns in the data, any inferences that are made by the data systems themselves. More advanced students should be able to quantify the complex trends and patterns in the data.

ELO 3: Argument and Conclusion

3.1 Develop a coherent and well supported argument

Is the ability of a student to produce an argument with supporting evidence. This argument may take many forms, including an argumentative essay, experiment or debate. The evidence must be strongly related to the aim/contention of the argument at all times, with any erroneous data acknowledged. The argument considers information which either can support or refute the contention/hypothesis as appropriate.

3.2 Develop a strong conclusion

Is the ability of a student to be able to develop a strong and unambiguous conclusion relating to the data. This conclusion need not be all-encompassing; where appropriate, a good conclusion can also include comments or caveats which point to a lack of data or to the surety of a decision. The conclusion should always be relevant to both the argument posed and the contention which it answers.

ELO 4: Implications of Decisions

4.1 Further investigation

Is the ability of a student to both transfer knowledge and principles to near/far situations and to determine the next step in a process. This transfer of understanding begins as applying the understanding to closely related situations, and at the higher levels involves the student's ability to apply the skill or understanding to dissimilar contexts.

4.2 Ethical Judgements

Is the ability of a student to be able to empathise and articulate the views of others. Judgement in a situation should always take into account the ethical considerations of a problem. The decisions made in a situation should be based not only on the student's ideas of right or wrong, but should also show an awareness of the views of others within the community.

4.3 Metacognition

Is the ability of a student to reflect on thought processes used in any given situation. At the higher levels, a student will modify their thinking to suit a particular strategy and will be able to articulate the changes and the reasons for them. This metacognition may be determined by written journals or through questioning during class.

Appendix B: Sample continuum used to measure student progress.

ELO 2: Working with Data

Aspect 2.1 Evaluating the Reliability of Data

The student is able to evaluate data stemming from complex experiments involving multiple variables; can consistently identify experimental errors stemming from more complex sources (placebo effects, statistical significance, bias); can isolate misinterpretation of scientific terminology/theory which undermines the experiment. The student is able to suggest changes to the collection method which could eliminate these errors, or suggest alternative hypotheses about a flawed set of data.	
The student is able to evaluate data stemming from complex experiments involving multiple variables; can identify some experimental errors stemming from more complex sources (placebo effects, statistical significance, bias). The student is able to suggest changes to the collection method which could eliminate these errors, or suggest alternative hypotheses about a flawed set of data.	
The student is able to evaluate data stemming from simple multi-variable experiments; can identify more complex experimental errors (placebo effects, statistical significance, bias). The student is able to suggest changes to the collection method which could eliminate these errors.	
The student is able to make comments about the reliability of data collected in simple multi-variable experiments. The student can identify variables which have not been controlled, and can suggest changes to the collection method in order to control them.	
The student is able to make more sophisticated comments (problems in calculations, errors in types of data collected) about the reliability of data collected in simple single variable experiments. The student can identify variables which have not been controlled, and can suggest changes to the collection method in order to control them.	
The student is able to make some basic comments (absence of steps, lack of specificity, nor result step) about the reliability of data collected in simple single variable experiments. The student can identify variables which have not been controlled, and can suggest changes to the collection method in order to control them.	
The student is able to make some basic comments (absence of steps, lack of specificity, no result step) about the reliability of data collected in simple single variable experiments. The student can identify variables which have not been controlled.	
The student is able to make some basic comments (absence of steps, lack of specificity, no result step) about the reliability of data collected. The complexity of these ideas is limited to simple statements about missing steps or nonspecific instructions.	
The student is able to recognize basic format flaws in a given collection method.	
The student is unable to recognize basic format flaws in a given collection method.	

Appendix C: Relationship between aspects of scientific literacy and testing statements in the ICAS program.

ELOs addressed by the ICAS test	**ELOs not addressed by the ICAS test**
Generation of an hypothesis	Number and variety of hypotheses
Identification of the most promising hypothesis	Evaluation of the reliability of sources
Effectiveness of collection procedure	Ethical Judgements
Processing data	Metacognition
Interpreting Data	
Developing a coherent and well-supported argument	
Developing a strong conclusion	
Further Investigation	

Appendix D: Essential Learning Outcomes classified according to scientific literacy (Vision I and Vision II) and/or science literacy.

ELO	Aspect	Type of Literacy Involved	Vision I or Vision II (Robert, 2007)
1. Hypothesis and Contention	Generation of an Hypothesis/Contention	Scientific literacy and Science literacy	Vision I
	Number and variety of Hypotheses/claims	Scientific literacy	Vision I
2. Collecting and evaluating evidence	Effectiveness of collection procedure	Scientific literacy	Vision I
	Evaluating the reliability of sources	Scientific literacy and Science literacy	Vision I
	Processing data	Scientific literacy and Science literacy	Vision I
	Interpreting Data	Scientific literacy	Vision I
3. Argument and Conclusion	Developing a coherent and well-supported argument	Scientific literacy and Science literacy	Vision I
	Developing a strong conclusion	Scientific literacy and Science literacy	Vision I
4. Implications of decisions	Further Decision	Scientific literacy	Vision II
	Ethical Judgements	Scientific literacy	Vision II
	Metacognition	Scientific literacy	Vision I

Appendix E: Descriptors for levels of alignment in assessment according to Webb (1997).

Criteria		Scale of Agreement		
		1. Insufficient	2. Acceptable	3. Full
1A	Categorical concurrence	Important topics are excluded from assessment to the extent students can perform acceptably on assessments and still lack understanding of intended goals.	Assessments cover a number of skills so that a student judged to have acceptable knowledge on the assessment will have demonstrated some knowledge on nearly all curriculum goals.	A one-to-one correspondence between topics given in expectation and topics by which assessments results are reported.
1B	Depth of knowledge consistency	Students can be judged as performing at an acceptable level on the assessments without having to demonstrate for any topic the attainment of the most cognitively demanding expected performance for each student.	For nearly all major topics, nearly all of the most cognitively challenging expected performance for all students is comparable to or can be inferred from the most cognitively demanding task taken by all students.	For each major topic, the most cognitively challenging expected performance for all students is comparable to the most cognitively demanding task taken by all students.
1C	Range of knowledge tested	Important forms or specific cases of major concepts and/or ideas given in the expected performance are excluded from or ignored on assessments or their	Assessment specifications account for nearly all forms or the full range of each major concept or idea expressed in the expected performance so there is a	Students are required on all assessments to show knowledge of all forms or the full range of each major concept or idea expressed in the expected range of

		specifications.	strong likelihood that students' knowledge and use of all forms will be assessed.	performance.
1D	Balance of representation	Weights on assessments by topic are sufficiently different from the assigned importance in the expectations such that a student could be judged as meeting the performance expectations without knowledge of highly emphasised topics.	Distribution of importance by topics in performance expectations nearly matches the weight in assessments without major exclusions.	The proportion of assigned importance of topics in performance expectations is equivalent to the weight in assessments.
2	Cumulative growth in procedural knowledge	Assessment instruments across the grades do not represent a logical or sequential growth in student knowledge over time implied in the expectations. Assessments in lower grades require a more advanced understanding than do those in later grades as depicted in performance expectations. Or, important stages in the development of skills are excluded from assessment events.	Assessment instruments elicit information according to general patterns according to how students' knowledge develops over time and how students relate these ideas.	Assessment instruments elicit information compatible with how students' knowledge develops over time and how students relate these ideas.

Appendix F: Information Letter for Participants

Information Letter to Participants

Alignment of Intended Learning Outcomes, Curriculum and Assessment in a Victorian Middle School Science Curriculum

Student Researcher:

Name: Reid Smith
Faculty: Faculty of Education and Arts
Edith Cowan University, WA
Phone: 03 5330 8200
Email: smithre@bcc.clarendon.vic.edu.au

Supervisor:

Name: Dr. Graeme Lock
Faculty: Edith Cowan University (CRICOS Code 00279B)
School of Education
2 Bradford Street
Mt Lawley 6050
Room 17.144
Phone: 08 9370 6529
Email: g.lock@ecu.edu.au

I am a student currently undertaking a Masters of Education by Research degree at Edith Cowan University. I wish to invite you to be a participant in my study of the alignment of Intended Learning Outcomes, Curriculum and Assessment in a Science Curriculum.

Description of the research project

This research project will focus on a case study of a regional Victorian, independent Middle School. Recent measures have indicated that the current science curriculum of this Middle School may not develop students' skills in science literacy as effectively as possible. One hypothesis is that there is a misalignment of intended outcomes, curriculum materials and assessment. This research project has two purposes: to determine the extent to which the intended curriculum and assessment performed in this Victorian middle years science program are aligned to its stated goals and objectives; and to design, implement and evaluate a model for assessing the degree of alignment of intended outcomes, curriculum and assessment. The research project will utilise modified versions of three existing curriculum evaluation tools and will use both qualitative and

quantitative analysis methods to determine the extent of the alignment of curriculum materials. It is anticipated that this research project will provide a model for analysing the extent to which the assessment and instruction are aligned to intended learning outcomes in a middle years science curriculum, as well as producing a realignment of the course materials in the case study school.

You have been selected to participate due to your familiarity with the purpose of the Middle School Curriculum featured in this study, as well as the scientific themes explored in each curriculum. Your participation would include:

- *Training in the scoring of selected curriculum materials against a series of rubrics.*
- *Actual scoring of selected curriculum materials.*
- *Two semi-structured interviews which will be recorded using videotape. The interviews will be conducted in order to ascertain how accurate the scoring process is and whether the scoring process is easily applicable to the secondary school environment.*

It is estimated that your involvement would consist of approximately 8 hours duration, and is entirely voluntary.

Ethical Clearance of the research:

This research project has gained ethics approval from the ECU Human Research Ethics Committee.

Confidentiality of information

The information you provide will be used to evaluate the effectiveness of both the curriculum materials being assessed and also the effectiveness of the scoring methods developed in the study. All information you provide will be stored in a locked cabinet, and used only for the purpose of this study. The results of the study will be used to produce a thesis paper for submission.

Withdrawing consent to participate

As a participant, you are free to withdraw their consent to further involvement in the research project at any time. If you choose to withdraw, any materials relating to your work in the project will be destroyed.

If you have any questions or require any further information about the research project, please contact:

Reid Smith
Faculty of Education and Arts
Edith Cowan University, WA
Contact:
Email: smithre@clarendon.vic.edu.au
Ph: (03) 5330 8200

If you have any concerns or complaints about the research project and wish to talk to an independent person, you may contact:

Research Ethics Officer
Edith Cowan University
270 Joondalup Drive
JOONDALUP WA 6027
Phone: (08) 6304 2170
Email: research.ethics@ecu.edu.au

Appendix G: Consent Form for Research Participants

Consent Form for School Leaders and Teachers
(Research Participants)

Consent Form

- I have read this document, or have had this document explained to me in a language I understand, and I understand the aims, procedures, and risks of this project, as described within it.
- For any questions I may have had, I have taken up the invitation to ask those questions, and I am satisfied with the answers I received.
- I understand that participation in the project is entirely voluntarily.
- I am willing to become involved in the project, as described.
- I understand I am free to withdraw from participation at any time within 5 years from project completion, without affecting my relationship with the school, with the research team or Edith Cowan University.
- I give my permission for the contribution that I make to this research to be published in academic journals, presented at conferences and presented in research reports, provided that I or the school is not identified in any way.
- I understand that a summary of findings from the research will be made available to me upon its completion.
- I understand by consenting to this interview, I might be contacted for another interview.

Name of Participant (printed): _____

Signature of Participant: _____ Date: / /

Appendix H: Individual Reviewer Scores for the Alignment of Year 7 and Year 9 curriculum materials

Criteria	Score							
	Year 7				Year 9			
	Reviewer 1	Reviewer 2	Reviewer 3	Mean Score	Reviewer 1	Reviewer 3	Reviewer 3	Mean Score
Are the key goals of the intended curriculum addressed?	2.5	2.5	2.5	**2.5**	2	2.5	2.5	**2.3**
What is the extent of curriculum materials supporting the key goals?	2	1.5	2	**1.8**	1.5	1.5	2	**1.7**
Is there an identification and maintenance of a sense of purpose towards the intended learning goals?	2.5	2.5	2.5	**2.5**	2	2	2	**2**
Do the curriculum materials take into account student ideas on scientific literacy?	3	3	3	**3**	2.5	1.5	1.5	**1.8**
Does the intended curriculum engage students with the key goals?	2.5	2.5	2	**2.3**	2	1.5	2.5	**2**
Does the intended curriculum develop and use scientific literacy?	2.5	2.5	2.5	**2.5**	2	2	2	**2**
Does the intended curriculum promote student thinking about science literacy?	3	3	3	**2.7**	2	2	2	**2**

Appendix H: Individual Reviewer Scores for the Alignment of Year 7 Assessments

Criteria	**Score**																											
	Dog's bark				**Safety Task**				**Running Race**				**Camping on the Range**				**Candy Co.**				**Reflection Booklet**				**Overall Assessment Materials**			
	Reviewer 1	**Reviewer 2**	**Reviewer 3**	**Mean Score**	**Reviewer 1**	**Reviewer 2**	**Reviewer 3**	**Mean Score**	**Reviewer 1**	**Reviewer 2**	**Reviewer 3**	**Mean Score**	**Reviewer 1**	**Reviewer 2**	**Reviewer 3**	**Mean Score**	**Reviewer 1**	**Reviewer 2**	**Reviewer 3**	**Mean Score**	**Reviewer 1**	**Reviewer 2**	**Reviewer 3**	**Mean Score**	**Reviewer 1**	**Reviewer 2**	**Reviewer 3**	**Mean Score**
Categorical Concurrence	2	2	2	**2**	3	2	3	**2.7**	2	2	2	**2**	1	2	1	**1.7**	3	3	3	**3**	1	1	1	**1**	2	2	2	**2**
Depth of knowledge consistency	3	2	3	**2.7**	3	3	3	**3**	3	3	3	**3**	3	3	2	**2.3**	3	3	2	**2.7**	2	2	2	**2**	3	2.5	2.5	**2.8**
Range of knowledge tested	1	1	1	**1**	2	1	1	**1.3**	1	1	1	**1**	1	1	1	**1**	3	3	3	**3**	1	1	1	**1**	2	1.5	1.5	**1.7**
Balance of representation	2	2	2	**2**	2	2	2	**2**	2	2	2	**2**	1	1	1	**1**	2	2	2	**2**	2	2	2	**2**	1.5	2	1.5	**1.7**
Cumulative growth in content knowledge	3	3	2	**2.7**	3	2	3	**2.7**	3	3	3	**3**	2	1	2	**1.8**	3	3	3	**3**	2	2	2	**2**	2.5	2	2.5	**2.3**

Appendix I: Individual Reviewer Scores for the Alignment of Year 9 Assessments

Criteria	**Score**																											
	Temp prac				**Conc Prac**				**Datsun Mystery**				**Murder Most Foul**				**Reflection Booklet**				**Exam**				**Overall Assessment Materials**			
	Reviewer 1	**Reviewer 2**	**Reviewer 3**	**Mean Score**	**Reviewer 1**	**Reviewer 2**	**Reviewer 3**	**Mean Score**	**Reviewer 1**	**Reviewer 2**	**Reviewer 3**	**Mean Score**	**Reviewer 1**	**Reviewer 2**	**Reviewer 3**	**Mean Score**	**Reviewer 1**	**Reviewer 2**	**Reviewer 3**	**Mean Score**	**Reviewer 1**	**Reviewer 2**	**Reviewer 3**	**Mean Score**	**Reviewer 1**	**Reviewer 2**	**Reviewer 3**	**Mean Score**
Categorical Concurrence	2	2	2	**2**	2	2	2	**2**	3	2	2	**2.3**	3	3	3	**3**	1	1	1	**1**	2	1	2	**1.7**	2	2	2	**2**
Depth of knowledge consistency	3	3	2	**2.8**	3	3	2	**2.8**	3	2	3	**2.7**	3	3	3	**3**	2	2	2	**2**	2	2	2	**2**	2.5	2.5	2.5	**2.5**
Range of knowledge tested	2	2	2	**2**	2	2	2	**2**	1	1	1	**1**	3	3	3	**3**	1	1	1	**1**	2	2	2	**2**	2	2	2	**2**
Balance of representation	2	1	2	**1.8**	2	1	2	**1.8**	2	2	2	**2**	2	2	2	**2**	2	2	2	**2**	1	2	1	**1.3**	1.5	2	2	**1.8**
Cumulative growth in content knowledge	3	3	3	**3**	3	3	3	**3**	2	2	2	**2**	3	3	3	**3**	2	2	2	**2**	2	2	3	**2.3**	2.5	2.5	3	**2.7**