EXPLORING THE IMPACT OF A LARGE-SCALE DIAGNOSTIC SCIENCE TEST AND FORMATIVE PRACTICES. A mixed-methods study.

James Scott MEd

Doctor of Philosophy C02041

University of Technology Sydney

Faculty of Arts and Social Sciences

Certificate of original authorship

I, James Scott declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy by Thesis in the Faculty of Arts and Social Sciences at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

29 August 2018

Acknowledgments

This thesis would not have happened without insights, support and encouragement from a number of people.

My thanks to Dagmar McCloughan, ESSA Team Leader, for the opportunity to be involved with the ESSA program in the early years of its development and implementation. Also, my thanks to Professor John Pegg (University of New England) and Associate Professor Debra Panizzon (then from the University of New England) for the opportunity to be a part of the research team investigating the potential of SOLO as a tool for improving assessment for learning. After that initial involvement, both provided me with their advice, encouragement and support which I sought at different times whilst I worked on this thesis.

I would like to acknowledge Doctor Geoff Barnes who encouraged me to run with the idea that the residual from a regression procedure was a measure of a real effect of teaching. My thanks also to the NSW Department of Education for providing me with access to ESSA and NAPLAN data in a form that I could use for the purposes of this thesis. Particular thanks in this regard are due to Doctor Nadine Smith and former colleague and dear friend Gerry McCloughan.

Associate Professors Nick Hopwood and Tapan Rai from the University of Technology Sydney have my gratitude for the time and advice they provided as I developed the approach I wanted to take with the research model and analysis of data.

I am extremely grateful to the science teachers who responded to the survey about their practice and particularly so to the teachers who made themselves available to participate in the case studies. I have undertaken to provide them with the results of my work in a form that I hope will be useful to them.

I had the support and advice of two excellent supervisors, both at the University of Technology Sydney, for this thesis. Professor Peter Aubusson who encouraged me to undertake this project in the first instance and Associate Professor Matthew

Kearney who took over in the later stages to assist me bring it to a conclusion. My gratitude and thanks to both for their patience, advice and support.

This thesis also had the benefit of the considerable editing skills of Doctor Terry Fitzgerald who is also at the University of Technology Sydney.

Finally, I want to acknowledge the forbearance of Daune my wife who, in the end, waited patiently and supportively for me to complete this project so that we could resume our lives together.

Thesis format

This is a conventional thesis comprised of title, front matter, glossaries (acronyms and terms used), table of contents, list of figures, list of tables, abstract, six chapters, appendices and references consulted in the preparation of this thesis.

List of Acronyms

AAS	Australian Academy of Science
ABS	Australian Bureau of Statistics
ACARA	Australian Curriculum Assessment and Reporting Authority
ACCI	Australian Chamber of Commerce and Industry
ACER	Australian Council for Educational Research
AE	At Expectation (see also WAE and WBE)
ANOVA	Analysis of Variance
AQF	Australian Qualifications Framework
ARG	Assessment Reform Group
BCA	Business Council of Australia
BOS	Board of Studies
BOSTES	Board of Studies, Teaching and Educational Standards
CC	Curriculum Corporation
CCII	Centre for Continuous Instructional Improvement
DEC	NSW Department of Education and Communities
DET	NSW Department of Education and Training
D of E	Department of Education
ESA	Education Services Australia
ESSA	Essential Secondary Science Assessment
EV	Acronym for the acronyms ESSA and VALID.
F	The Foundation or entry level for schooling (see K).
HSC	Higher School Certificate
ICSEA	Index of Community Socio-Educational Advantage
K	Kindergarten or entry level for schooling (see F).
NAP-SL	National Assessment Plan-Scientific Literacy
NAPLAN	National Assessment Plan Literacy And Numeracy
NESA	New South Wales Education Standards Authority
NGSS	Next Generation Science Standards (US)

NSES	National Science Education Standards (US)
OECD	Organisation for Economic Co-operation and Development
PCK	Pedagogical Content Knowledge
PIRLS	Progress in International Reading Literacy Study
PISA	Programme for International Student Assessment
SEA	Socio-Educational Advantage
SEAR	Science Education Assessment Resource
SET	Science, Engineering and Technology
SLPM	Scientific Literacy Progress Map
SMART	Schools Measurement Assessment and Reporting Toolkit
SME	Science, Mathematics and Engineering
SOLO	Structure of the Observed Learning Outcome
SPSS	Statistical Package for the Social Sciences
STEM	Science, Technology, Engineering and Mathematics
TIMSS	Trends In Mathematics and Science Study
US	United States of America
VALID	Validation of Assessment for Learning and Individual Development
VET	Vocational Education and Training
WAE	Well Above Expectation (see also AE and WBE)
WBE	Well Below Expectation (see also AE and WAE)

Glossary of terms as used in this thesis

artifact
Something made by human effort, in this context related to educational assessment.

assessment as learning
Assessment as learning occurs when students are their own assessors. Students monitor their own learning, ask questions and use a range of strategies to decide what they know and can do, and how to use assessment for new learning. (NESA, 2018)

assessment for learning
Assessment for learning involves teachers using evidence about students' knowledge, understanding and skills to inform their teaching. Sometimes referred to as 'formative assessment', it usually occurs throughout the teaching and learning process to clarify student learning and understanding. (NESA, 2018)

assessment of learning
The use of evidence of learning to make a summative judgment of achievement against outcomes and standards. Sometimes referred to as 'summative assessment'. It usually occurs after a period of instruction. The judgment is often expressed as a mark, percentage or grade. The usefulness of the grade or mark depends on validity and reliability of the processes used to gather and assign value to the evidence gathered. (NESA, 2018)

assessment-related work
Is the purposeful collecting of evidence of learning, creating the means by which that evidence was obtained (if not by direct observation of behaviour), the assumptions used to interpret that evidence, the choice of text forms used to represent and communicate results of assessment, and subsequent uses for those results.

capabilities	A measure of the ability, capacity, power or potential to do something. The *Australian Curriculum, Science* includes seven general capabilities all students are expected to acquire as they progress through schooling.
Curriculum Corporation	A national educational support entity created by the Federal, state and territory governments in Australia to produce educational resources for Australian Schools. It was replaced by Education Services Australia (ESA) from 2010.
competencies	See capabilities.
curriculum	The documents teachers use to inform the learning activities they plan and deliver to students.
diagnostic assessment	Gathering evidence of learning to identify gaps, strengths and weaknesses in student learning.
education jurisdiction	States and territories in Australia manage the delivery of educational services to students in Australia. They provide for registration and regulation of public and private schools in their geographic areas of jurisdiction.
educational standards	Are the learning goals students are expected to achieve, usually after set periods of instruction typically associated with Year or Grade levels.
feedback	Information provided by an agent regarding aspects of one's performance or understanding.
formative assessment	See assessment for learning.
formative practices	Instruction informed by formative feedback.
high stakes assessment	Any assessment where the results have consequences for the recipient of those results.
key competencies	A set of competencies related to equipping students for work.

low stakes assessment	The use of evidence of learning in ways that reduces to a minimum unintended, usually negative, consequences for the learner.
outcomes	Measurable or observable behaviours intended as a result of instruction.
Primary Connections	A set of curriculum materials produced by the Australian Academy of Science designed to assist K-6 teachers to teach science.
proficiency areas	Areas of skill or expertise.
proficiency levels	Descriptions of response features that differentiate between levels of skill or expertise.
regression	Regression is a statistical process for estimating the relationships between variables.
Science by Doing	A curriculum support resource produced for secondary science teachers by the Australian Academy of Science.
scientific literacy	Scientific literacy is the ability to engage with science-related issues, and with the ideas of science, as a reflective citizen (OECD). It is also the specialized literacies that distinguish science literacy from general literacy and numeracy.
SEA quarters	Socio-Educational Advantage (SEA) proportions, relative to Australia, in school populations. (ACARA *MySchool* website)
SEA score	Socio-Educational Advantage (SEA) score is a composite measure of socio-educational advantage generated for the purposes of this project.
selective entry schools	A category of school in NSW, entry to which is determined by student results in tests of reading, mathematics, general ability and writing.

self-regulated learners	Students who can plan their own learning, monitor their performance and then reflect on the outcome of that learning.
Skills, cognitive	Include remembering, thinking logically and reasoning, explaining and describing.
Skills, employability	Skills related to communicating, working in teams, problem solving, initiative and enterprise, planning and organising and self-management.
Skills, generic	Groups of skills variously described as basic/fundamental, people-related, conceptual/thinking, personal skills and attributes, skills related to the business world and skills related to the community.
SOLO model	Structure of the Observed Learning Outcome (SOLO) theory that involves two learning cycles within a mode of thinking
SOLO taxonomy	Structure of the Observed Learning Outcome (SOLO) theory that describes a single learning cycle within a mode of thinking
standards framework	Descriptions of levels of performance in a number of categories relating to curriculum, teaching or other profession.
statistically significant	Is the probability of finding a given deviation from a null hypothesis, or a more extreme one, in a sample. (SPSS definition)
STEM system	Science, Technology, Engineering and Mathematics institutions in a country or larger group that prepares people for work in, and including, the institutions that produce STEM outputs in society and related economies.
summative assessment	See assessment of learning.

syllabus	A detailed curriculum that in NSW may be used to define the scope of an external test.
The Board	A generic term for the statutory authority in NSW with responsibility for determining the curriculum and related assessment requirements schools need to comply with so that students satisfy requirements for receipt of credentials. In the course of this project that authority began as the NSW Board of Studies (BOS), became the NSW Board of Studies Teaching and Educational Standards (BOSTES) before becoming the NSW Education Standards Authority (NESA) in 2017.
The Department	A generic term covering the NSW government authority responsible for delivering public education services to students in NSW. It went from being at the beginning of this project (2012) the NSW Department of Education and Training (DET) to the Department of Education and Communities (DEC) to the NSW Department of Education (D of E).
Year 8	The year of schooling in Australia (Grade in other places); in this case the ninth year of schooling.

Table of contents

List of Figures

List of Tables

Abstract

Researchers working with schools in the UK and elsewhere are finding that explicitly teaching students the "five strategies of formative assessment" (Black and Wiliam, 2009, p. 8) is helping to re-engage students with science. This thesis presents findings about the impact of two major interventions on the assessment-related work of junior secondary science teachers in the New South Wales government school system (the largest in Australia) and on student science results. The first intervention took the form of advice to teachers about formative assessment in the official science curriculum (introduced in 2003), where it is called assessment for learning. The second took the form of a mandatory low-stakes, large-scale, test-based diagnostic assessment program involving Year 8 students. This program was fully implemented across NSW from 2007. The assessment framework used to inform the development of test items and tasks and that informs the comprehensive feedback provided to students, parents and teachers is underpinned by Structure of the Observed Learning Outcome (SOLO) theory. Three research questions guided data collection. The research design employed mixed methods, including both quantitative and qualitative methods as well as case studies involving sixteen purposively chosen school sites. Descriptive and inferential statistics were applied to the analysis of both state-wide and school-specific, teacher-provided survey data about their practices and school-level test results. An interpretive approach was used to generate assessment-related work narratives from audio-recorded interviews and artefacts of assessment practice provided to the researcher by volunteering science teachers in the case study schools. The findings show that teacher use of three of five dimensions of formative practice and an explicit focus on teaching students the skills of writing to learn science produced science test results that were above expectation. Less certain was the hoped-for finding that students were also acquiring the skills of learning how to learn. An unexpected finding was that students in regional schools where science results were well above expectation were less positive about their school science experience than their metropolitan counterparts.

CHAPTER ONE: OUTLINE OF MY PROJECT

1.1 Introduction

This thesis reports in six chapters how I used a mixed methods research design to explore the impact of two assessment initiatives on teachers' assessment-related work and student results in the largest government-run school system in Australia. The findings are then used to argue in the final chapter for the retention of both initiatives and to support recommendations to enhance their future effectiveness.

Education in Australia is the responsibility of the eight states and territories that make up the Commonwealth of Australia (Commonwealth of Australia Constitution Act, 1901). The state and territory governments in those jurisdictions have established government (or public) school systems which are managed by education departments responsible to those governments. Education departments allocate and manage the human and physical resources provided by governments to deliver educational services to students in the government school system. Students enrolled in the government school system are entitled to free education from age 5 to 17 years.

Private interests have also established schools in each of the state and territory jurisdictions. The majority of those schools are affiliated with organized religions. The Catholic Church supports the largest number of schools affiliated to a religious organization. Private schools with a common philosophy or religious affiliation have formed themselves into systems for the purposes of efficient and effective use of resources. Parents pay school fees directly to private schools to send their students there. However, all school systems in Australia receive money collected by government tax systems.

The governments of the eight states and territories have established autonomous authorities to manage the registration and accreditation of schools established by both government and private interests. Registration ensures the community that their children are educated in appropriate physical surroundings and provided with adequate human and other resources to support their learning. Accreditation ensures that students have access to educational programs based on a high-quality curriculum and related assessment and credentialing processes. Registration and accreditation processes are determined and managed independently of direct government influence. In recent times state and territory governments have added registration of teachers and accreditation of tertiary education courses preparing people for teaching to the responsibilities of those autonomous education authorities.

Over the past four decades, the eight state and territory governments, with support from the national government, have been working toward a shared national policy agenda for education in Australia. In 2008, by cooperative agreement of the national and all state and territory governments, the *Australian Curriculum Assessment and Reporting Authority* (ACARA) was established to perform the following functions "development of national curriculum, administration of national assessments and associated reporting on schooling in Australia" (ACARA, 2016a). ACARA is responsible to the Council of Australian Governments (COAG) Education Council.

New South Wales (NSW) is the most populous state in Australia and around 20% of all secondary school students in Australia attend a government school in NSW (ABS, 2018). It's Department of Education ("the Department" in this thesis) manages the largest government school system of all eight states and territories. The autonomous education authority in NSW is at the time of writing this the *NSW Education Standards Authority* (NESA) and is referred to as "the Board" in this thesis. It was variously the *NSW Board of Studies* (BOS) then the *NSW Board of Studies, Teaching and Educational Standards* (BOSTES) before becoming NESA on January 1, 2017. Data used in this research was provided to me by the Department and by science teachers working in government secondary schools across NSW. It was supplemented by school data available on the national *MySchool* website managed by ACARA.

The following section, Section 1.2, will outline the two assessment initiatives that are the focus of interest for this thesis. Section 1.3 will outline the research questions and methodology. Section 1.4 will provide an overview of the findings. Section 1.5 explains the importance of the research. Section 1.6 will explain my interest in the two initiatives and Section 1.7, the final section in this chapter, will outline the structure of my thesis.

1.2 The two initiatives

The phrase 'formative practices' in the title of this thesis is taken from a paper by two researchers, Black and Wiliam (2009) titled *Developing the theory of formative assessment.* They used the phrase to cover theorizing about instruction informed by feedback from assessment. The paper had its origins in work the pair had been commissioned to do some thirteen years earlier by the UK based Assessment Reform Group (ARG) with funding from the Nuffield Foundation. Black and Wiliam were commissioned to review the literature on the use of assessment to support learning, also known as formative assessment. The results of that review were published in a booklet for teachers called *Inside the Black Box* (Black & Wiliam, 1998b).

The ARG had used the phrase "assessment for learning" (ARG, 2002a, p. 3) to differentiate it from "assessment of learning" (ARG, 2002a, p. 3). A full explanation of the distinctions between the two will be provided in Chapter Two, the literature review for this thesis. This thesis will explore the assessment-related work of teachers in the early years of secondary schooling to find out the extent to which that work could be described as "formative" in Black and Wiliam's (2009, p. 8) theory of formative assessment. In other words, the extent to which instruction or teaching is explicitly informed by the results of assessment-related work of teachers.

Assessment-related work of science teachers is defined here as the purposeful collecting of evidence of learning, creating the means by which that evidence was obtained (if not by direct observation of behaviour), the assumptions used to interpret that evidence, the choice of text forms used to represent and communicate results of assessment, and subsequent uses for those results. ‘Student results’ as used in the title refers to the representation of the judgment made by teachers about the value of the teacher collected evidence of student learning. It is typically represented by a grade, a mark (sometimes expressed as a percentage) or a level (in this project, six levels were common). This form of result is what is meant by assessment of learning. It becomes assessment for learning when it is used to inform the next step in teaching or instruction (feedback) while it is happening.

The first of the initiatives used in this study was assessment advice for science teachers titled: “Assessment for Learning?” (BOS, 2003, p. 70). It was embedded in the 2003 release of the official science curriculum documents that secondary science teachers are expected to use when preparing teaching and learning programs for their students. The initiative took the form of advice to teachers about how to gather and use evidence of learning to inform the next steps in instruction as it was happening. In other places ‘assessment for learning’ is referred to as “classroom assessment” by Shepard (2001, p. 2) or “formative practices” by Black & Wiliam (2009, p. 6) in their paper on the theory of formative assessment. The curriculum document (also referred to as a syllabus in NSW) summarises the scope of assessment for learning for science teachers in these terms. It:

1. is an essential and integrated part of teaching and learning
2. reflects a belief that all students can improve
3. involves setting learning goals with students
4. helps students know and recognise the standards they are aiming for
5. involves students in self-assessment and peer-assessment
6. provides feedback that helps students understand the next steps in learning and plan how to achieve them
7. involves teachers, students and parents in reflecting on assessment data. (BOS, 2003, p. 70)

The focus on assessment for learning in official curriculum documents was a strong signal to teachers about the need to shift the emphasis from using evidence of learning for reporting achievement after instruction to improving instruction itself. Other implications are that curriculum intentions, instruction and assessment should be aligned and that students and the wider school community need to be more involved more. The current curriculum documents (BOSTES, 2012) continue with that emphasis and have extended it to include advice on "assessment as learning" as well as "assessment of and for learning" (NESA, 2018). All three will be discussed further in the literature review (Chapter Two). The current (2018) curriculum for science in NSW replaced the 2003 curriculum beginning with Year 7 and 9 in 2014.

The second initiative was a test-based intervention called at the time of its introduction the Essential Secondary Science Assessment (ESSA) program. The test was delivered to students at the midpoint of a mandatory four-year science course commencing in their first year of secondary schooling (Years 7 to 10 in Australia). After piloting (2005) and trialing (2006), the first test for the full cohort of Year 8 students was in 2007. In its initial form, it was a pen-and-paper test with the same 'look and feel' as other pen-and-paper tests students were used to doing. It was subsequently delivered online from 2010 and continues to be delivered this way. It was the first cohort test to be delivered online by an education jurisdiction in Australia.

The test was designed to do much more than provide a report to parents on student achievement at the midpoint of a four-year science course. It was designed as "a diagnostic tool to identify what students know and can do and where teaching needs to be directed to enhance scientific understanding" (Panizzon, Arthur, & Pegg, 2006, p. 1). To better support that goal, the assessment framework was informed by the Structure of the Observed Learning Outcome (SOLO) model. SOLO is a "cognitive structural model" (Panizzon, 2003, p. 1428) developed from empirical studies of the structure and sophistication of the language used by students in their responses to test items and tasks. The SOLO model used in NSW is based on the SOLO taxonomy originally published by Biggs and Collis (1982, 1991).

The assessment framework developed for the ESSA program enabled a map to be created that puts syllabus expectations along one axis and levels of understanding about those expectations along a second axis. How this works will be explained further in Chapter Two. The test was also accompanied by a survey designed to find out what students thought about science, their school science experience and the test itself. The results of the survey analysis were provided to science teachers along with detailed feedback about student responses to every item and task in the test.

The ESSA program was compulsory for all Year 8 students in the government school system and for Year 8 students in non-government schools that had opted into the program. The program was expanded in 2015 to include a test for Year 6 and Year 10 students and renamed Validation of Assessment for Learning and Individual Development (VALID). The addition of two extra tests provided schools with a way of mapping the progression of student learning in science from Years 6 to Year 8 and then Year 10.

VALID8 remained compulsory for all government schools, but the new VALID6 and VALID10 tests were (and still are) optional for both government (and non-government schools wanting to participate). As the program name change took place before data collection began in this project (second half of 2016) and schools were already calling it the VALID program, I chose to use the acronym EV in this thesis to reflect both the original (ESSA) and new (VALID) acronyms. I will refer to the EV program or EV test from this point onwards (unless it is more appropriate to refer to either ESSA or VALID).

The period of interest for this project is from 2011 to 2014 inclusive which were the last four years of data linked to teachers' work using the 2003 curriculum. The EV program is appropriately described as an external, large-scale, low-stakes, diagnostic intervention. 'External' refers to the source of the test, which is external to the school. 'Large-scale' refers to the size of the program, which includes all

NSW government schools with Year 8 students (465 schools at the time of this research). The student cohort size in Year 8 from 2010 to 2015 numbered around 47,000 students. The statistics quoted for the size of the government school system and the size of the EV program were sourced from the NSW Department of Education.

'Low-stakes' is a relative descriptor for the impact of the EV program on students, their parents and their teachers as explained further in Chapter Two.

Diagnostic assessment refers to the intended use of assessment results to identify strengths and weaknesses in student learning (Goodrum, Rennie, & Hackling, 2001; Hackling, 2004; Masters, 2013; Millar & Hames, 2003; Treagust, 2006).

The wider context for the two initiatives described in this section will be described in the first section of the literature review.

1.3 Research questions and methodology

This section outlines the specific research questions, the research design and related methodologies used to guide this research project. A full account of the methodology will be provided in Chapter Three.

The objective of this study is to answer the motivating question of what impact are the two initiatives of formative assessment and the diagnostic EV test having on the assessment-related work of science teachers in NSW government schools and why it matters?

Three research questions provide the focus for this research:

1. What use are science teachers making of the EV program including SOLO and why is it used or not used?
2. What formative practices are evident in the work of science teachers and why are they used or not used?

3. Is the use of formative practices by teachers linked to improvement in students' EV results and later achievement in and engagement with science?

The first question is about identifying the extent to which EV tests, EV results or related resources (including SOLO theory, student survey results and professional learning opportunities) have been accessed and used by science teachers to inform assessment-related work at their schools.

The second question is about the extent to which formative practices are evident in teachers' assessment-related work. Chapter Two will elaborate the theoretical framework of five dimensions of formative practice used in that exploration of teachers' work. "Formative practices" is a phrase used by Black and Wiliam (2009, p. 8) in their discussion of the theory of formative assessment. In that discussion Black and Wiliam explore the links between what they call the five strategies of formative assessment and their relationships to pedagogy or instruction. I decided to use Black and Wiliam's phrase and invented "five dimensions" of formative practice as the basis for characterising teachers' responses to items in an online survey about their work. The five dimensions were based on the five strategies of formative assessment as articulated by Black and Wiliam (2009) in their paper.

The third question is about investigating the association between formative practices and achievement (as measured by EV results and other assessments in science) and later take-up of science courses in the senior years of secondary schooling (a measure of ongoing or later engagement).

Also explored in relation to the third question was the extent to which the formative practices observed in the assessment-related work of science teachers may have assisted learners' acquisition of self-regulation (Boekaerts & Corno, 2005). Self-regulation describes students who are good managers of their learning, like learning and continue their involvement in learning. The expectation that some students had developed those attributes as a result of exposure to formative practices used by science teachers was based on the work being done in the UK by Black, McCormick, James and Pedder (2006), and James et al. (2007).

Three predictions were made to test the assumption of acquisition of self-regulation by some students. Confirmation of the three predictions would be taken as evidence that the assumption of self-regulation for some of the students was reasonable. The three predictions and the thinking behind them is discussed in Chapter Three. Analysis of data provided by case study schools are reported in Chapter Five. 16 schools identified themselves as willing to be involved in a case study as outlined below and fully in Chapter Three.

The capacity to manage one's learning is an essential skill in the context of the knowledge society and related economy where the capacity to learn new skills and adapt to change is increasingly important for maintaining a job and wider life satisfaction (UNESCO, 2005, p. 27). Chapter Two describes some of the work being done to teach students the strategies of formative assessment as one means for producing student self-regulation. It is for this reason that helping teachers to adopt formative practice as their default pedagogy "matters" (see the motivating question for this research project stated at the beginning of this section).

The research design involved mixed methods executed in three phases. An outline of the phases follows. Full details will be provided in Chapter Three and subsequent chapters.

The first phase employed a quantitative inferential statistics procedure where EV results were regressed over an EV result predictor and the residuals from that regression were used to identify three groups of schools. One group had schools with large positive residuals, a second group with zero or close to zero residuals and a third group with large negative residuals. As will be explained in Chapter Three, schools in these three groups are associated with EV results that were well above, at or well below expectation respectively.

Expectation was relative to the EV result predictor. The EV result predictor was developed from a combination of reading and numeracy results obtained by students in national testing in Year 7 and again in Year 9. The reasoning for using such a predictor is explained in Chapter Three.

The Department accessed its records of test results for schools with 10 or more students in Year 8 who had sat the EV test in four successive years from 2011 to 2014 inclusive. It also matched those students with their Year 7 and Year 9 reading and numeracy results from national testing and retained those results for students who had sat the tests at the same school in successive years. Reading and numeracy results were used to generate four predictors of EV results for the 10 or more students in each school. In the end the Department provided me with four sets of regression residuals from 394 schools (out of a potential 465).

Using one of the four sets of residuals, I identified three groups of between 80-90 schools using the size and polarity of their residuals as the basis for allocation to one of the three groups. Science teachers at the selected schools were invited to complete an anonymous online survey about their teaching and assessment practices. Responses were collated according to the school group the science teachers had been assigned to.

The second phase employed a quantitative method to analyse teacher responses in each of the three groups and then to compare the results from each group for statistically significant differences. The procedure used was Analysis of Variance (ANOVA). Its purpose was to find out whether there were statistically significant differences in assessment-related practices of teachers in each of the three groups.

Analysis and findings from the first and second phase of the research were reported in Chapter Four.

The default position for responses to the online survey was respondent anonymity. However, respondents who wished to be considered for involvement in a case study (the third phase of the research design) were invited to identify themselves and their school. Teachers at 36 schools spread across the three groups identified themselves. Between four and six of the identified schools from each of the three groups were invited and subsequently participated in case studies.

Teachers at case study schools were invited to provide school level EV and Year 10 results, numbers of students completing Year 12 science courses and artifacts of teacher-produced assessment-related work considered by them to be exemplary practice (including test or assignment items and tasks, related marking rubrics, sample school reports, assessment plans or science department programs where assessment was explicitly described). Teachers were asked to bring the artifacts to a semi-structured interview at the school which was planned to take up an hour of their time. The interviews with teachers were audio recorded. Access to students was not part of the research design.

Case study schools provided Year 8, Year 10 results and Year 12 completion data. I sourced and collected case study schools' socio-educational advantage profile data from the ACARA managed *MySchool* website. That data and the residual (from phase one) were collated and analysed using inferential statistics to establish the strength of correlations to confirm (or disconfirm) three predications relevant to answering the third research question. Interviews and artifacts were qualitatively analysed and assessment-related work narratives were developed from that analysis for each of the case study schools as well.

The proposition that instruction consistent with formative practices may have supported students' self-regulated practices was also tested in the context of answering research question three.

Findings from quantitative analyses performed in the case study third, phase of the research along with supporting evidence and examples from the assessment-related work narratives for those schools were reported in Chapter Five.

Anonymity for participating teachers and their schools was guaranteed for this research.The steps taken to protect the identities of participating schools, data and teachers are described in Chapter Three.

1.4 Overview of findings

In terms of the methodology, the reading-numeracy predictor chosen accounted for 89.2% of the explained variation averaged over the four years (2011-2014) of results. This is a very strong correlation given that other large-scale testing programs involving predictors and regression analysis, such as ACARA's Index of Community Socio-Educational Advantage (ICSEA), accounted for 81% of explained variation in 2013 NAPLAN results (ACARA, 2014b) and 80% of the 2014 results (ACARA, 2015). When Rowe (2006) analysed the 2003 PISA results for the Australian sample of 15 year-old students, he found that the boys (n =6335) reading results accounted for 77.4% of the explained variation in their science results; the comparable figure for girls (n =6216) was 75.3% (Rowe, 2006, p. 8). The same students sat both the reading and the science tests.

When the residuals for all schools and different school categories were analysed, it was found that EV results "were better than expected" (i.e. the residual was positive) in:

- 53% of the 394 schools in the study;
- 67% of the provincial schools (n = an estimated 90 schools);
- 68% of the fully selective entry schools schools (n = 19); and
- 23% of the partially selective entry schools (n = 24).

In relation to the first research question about teacher use of EV resources and SOLO, some findings were that:

- 67% of science teachers made use of EV resources to support their assessment programs and in-class work;
- 25% of teachers rated their understanding of SOLO as good or very good; and
- 18% of teachers said they used SOLO as a basis for feedback to students on their learning.

When it came to student survey results (the survey accompanied the EV test and a new feature of external testing in NSW):

- 67% of science teachers said they had looked at the results
- 49% had discussed the results with their colleagues; and
- 18% of teachers said they had discussed the results with their students.

In relation to the second question, there were statistically significant differences in the use by teachers of three of the five dimensions of formative practice. The teachers at schools where results had been identified as being “well above expectation” (or WBE schools, compared to their colleagues in the other two groups of schools, were more frequent users of activities involving:

- discourse that elicits evidence of learning;
- the provision of feedback known to progress learning; and
- the use and modeling of “good learning behaviours” (Boyle, Fahey, Loughran & Mitchell, 2001, p. 200).

For the third research question, the answer to the first part of the question (Is the use of formative practices by teachers linked to improvement in students’ EV results…) was a strong yes. When it came to extrapolating that result beyond Year 8 to Year 10 achievement (…later achievement), uncertainty about the comparability of Year 10 data across schools was too great to have reasonable confidence in between school comparisons. The within school correlations for Year 8 and Year 12 science course completions and Year 10 achievement and Year 12 science course completions was highly positive and statistically significant.

The assumption that schools where results were ‘well above expectation’ would have more self-regulated students than other schools was the basis for three predictions about later achievement and later engagement. The terms achievement and engagement as used in this project are defined in Chapter Three. The predictions related to comparable schools (schools with the same socio-educational advantage). None of the predictions could be confirmed beyond reasonable doubt which in turn rendered the underlying assumption of self-regulation doubtful as well.

Contributing to the uncertainty about self-regulation was the finding that students at the three provincial case study schools that had ‘well above expected’ EV results were less positive about their school science experience than students in the metropolitan case study schools, most of whom were in schools where EV scores were ‘at’ or ‘well below expectation’.

1.5 Importance of the research

Two claims about the importance of this thesis are made. The first claim is that this project was the first large scale study in Australia using the results from an external science test to provide confirmation that formative assessment and related instruction (formative practices) are associated with better learning outcomes in science.

Here “better” means that the school’s overall science results had a higher mean than the science results of the school it was being compared to. In this context ‘comparable’ means a school or schools with the same socio-educational advantage score (a measure of the collective learning potential of students at a school; its derivation is explained in Chapter Three).

The wording of the claim and the notion of comparable schools relate to the methodology involved in producing the evidence for the claimed association between teacher use of formative practices and student learning.

As a result, this study adds to the growing body of evidence from around the world about the effectiveness of formative practices. A synthesis of key literature linking formative practices to better learning outcomes is presented in Chapter Two, the literature review.

Specifically, my research showed that students attained better results in those schools where teachers provided students more frequently with ‘science-rich’ activities involving three of the five dimensions of formative practice (Black & Wiliam, 2009). The dimensions were: classroom discourse eliciting evidence of learning; teacher feedback known to progress learning of that content and teacher use and modeling of “good learning behaviours”.

The second claim for importance relates to the study’s methodology. The methodology involves taking a student’s results from national literacy and numeracy testing to generate a predictor for their result in a science test. As was discussed in the section above, the regression of science test results over the same students' set of science predictor scores produced a school set of individual residuals. The claim here is that the residual is a measure of the real and direct contribution of science teaching to the science learning of students at that school. A positive residual means that a student has learned more science than expected; a negative residual means students have learned less than expected. When individual residuals are summed and averaged, the individual student residuals produce a school score.

When the residuals from all schools where this process has been applied are standardized they can be compared. Schools with larger positive residuals have done more for student scientific literacy than those where the residuals are large and negative. The process from residual to comparing actual school EV results commences in Chapter Three and the findings reported in Chapter Four.

An unanticipated finding was that science teaching in provincial schools had produced better than expected results but that (for high performing case study provincial schools at least) students were not enjoying their school science experiences. This last finding was an important consideration in concluding that an assumption of self-regulation as a contributor to later achievement and later engagement was not warranted.

1.6 The researcher

I began my career in science education as a secondary school science teacher (1967 to 1979) before taking on the role of head teacher, science in the NSW government school system (1980 to 1993). I accepted the role of senior science manager in the then newly created curriculum support directorate of the NSW Department of Education (1994 to 2005). In that role, I provided curriculum support to science teachers in government schools across NSW, managed the development of a number of science curriculum support resources, provided policy advice on science education to senior management in the Department and led professional development for a statewide network of science consultants.

I represented the Department at the national level as a member of steering committees and as a contributor to national, science teaching, curriculum, assessment, professional standards and curriculum support reviews and projects. I also participated regularly in the annual conferences of the Australian Science Teachers Association and Australasian Science Education Research Association.

Commencing in the mid-1970s, I had a number of roles with the NSW curriculum and assessment authority as a science curriculum writer, curriculum policy officer on secondment from school (1987 to 1990). I was a member, then chair, of the authority's science curriculum committee, a HSC examination marker, chair of a HSC examination committee and supervisor of marking for a HSC science course. Later I had a role with ACARA as both a curriculum writer for the F-10 Australian science curriculum and subsequently as an officer assisting with development of the senior Chemistry and Physics curriculum documents.

I joined the Science Teachers Association of NSW in 1967 and was elected Vice-President on two separate occasions. I was also a convenor of their professional development committee and annual conferences, contributor to those conferences and the senior judge and marking trainer for their Young Scientist Award. I also represented STANSW as a member of the team engaged by ASTA to write their professional standards document for Highly Accomplished Teachers (of science) which became a model for later professional standards documents. I was awarded an honorary life membership of STANSW in 1997.

I became a casual lecturer and then coordinator for the Bachelors and Masters pre-service science teacher courses at the University of Technology Sydney (2004 to 2015). I was also a member of teams that researched, developed, piloted, trialed and marked the first EV tests. During that time (2005 to 2008) I led the training for markers of the extended response tasks as well as being the key liaison person between the Department and the agency contracted to manage the online marking of the extended response tasks.

This thesis is the culmination for me of five decades of work in science education, starting with part-time degrees at Macquarie University (BA majoring in

Curriculum and Geophysics, completed in 1974) and The University of Sydney (MEd majoring in Curriculum, awarded in 1991).

It is my intention to use the results from this study to satisfy criteria for the award of a PhD and for future advocacy work. The latter will be achieved when I provide feedback from this study to participant schools and policy advice to the Department of Education, NSW. To the extent that my advocacy produces support from the Department for teacher professional development leading to more confident and accomplished uptake of formative practices, then the transformative intent of this study will be realised. In addition, I will be offering my support to the schools that participated in this study, should teachers there wish to implement advice provided in my feedback to them.

From the above resumé it is appropriate to say that I bring both an insider and outsider perspective to this doctoral study (Fensham, 2013). I was an insider in the following ways:

- as a research participant in the initial evaluation of the suitability of the SOLO model in informing the development, implementation and marking of the EV extended response tasks in the first four years of its life;
- as a member of reference groups for the review into the status and quality of school science in Australia (Goodrum et al., 2001), for the review of options for a national test for primary science (Ball, Rae, & Tognolini, 2000) and for the Science Education Assessment Resources (SEAR) project (ACER, 2004a)
- as a writer of both state and national science curriculum documents (BOS, 2003; ACARA, 2014c).

My outsider perspective is "like that of other interested science educators [who access] projects' reports of their findings and to their aftermath influence (in so far as findings are published) on the policy and practice of science education" (Fensham, 2013, p. 13).

I have titled this first chapter *Outline of My Project* and written it in the first person to ensure that readers recognise what I bring to this study. Subsequent chapters are written in the passive voice. This underpins my wish to be seen as an independent researcher who has taken appropriate steps (see section 3.7 in Chapter Three) to conduct the research in the full knowledge of issues related to participant researchers / observers that arise in the context of qualitative research in education and psychology (see Denzin & Lincoln, 2011 and Hammersley, 2008).

My last involvement with teachers in the context of supporting their implementation of the syllabus (BOS, 2003) was in 2004 and the EV program was in 2008. Data collection in case study schools for this project took place in 2016. I had previously worked with one of the case study teachers some 12 years prior to that. He was then a participant at a one-day workshop I was running at that time. His school was invited to participate as a case study school in 2016 because it met the criteria for inclusion as an outcome of the phase one quantitative methodology.

1.7 Structure of this thesis

Chapter Two explores the research and other literature consulted for this thesis. It provides the theory for conceptualising five dimensions of formative practice, that comprise the framework for investigating the impact of assessment for learning advice and the EV program on assessment-related work of science teachers.

Chapter Three explains the three-phase, mixed methods design used to investigate the impact of the two initiatives (the EV test and expectations for greater use of assessment for learning) on assessment-related work of science teachers in the NSW government school system. The five dimensions of formative practice, which is the framework against which impact will be investigated, are described there.

Chapter Four reports the findings from the first and second phases of the study. The first phase used an EV result predictor to identify schools where EV results were well above, at and well below expectation (relative to the predictor). Teachers in those schools were invited to complete an online survey about their work. The second phase involved the analysis of survey responses to find out whether better than expected EV results were associated with formative practices.

Chapter Five reports findings from the third phase of the project. The third phase involved testing (both quantitatively and qualitatively) the propositions that students at comparable schools (schools with the same socioeducational advantage scores) who had more frequent exposure to formative practices, compared to students at schools not so exposed, would have

- better Year 8 EV results
- better Year 10 results
- more students (as a proportion of the Year 12 cohort) complete senior science courses.

Chapter Six summarises the study's findings and provides qualified confirmation for the claims made about the importance of the research. It also provides some suggestions, supported by findings in this project, for future research and recommendations to relevant education authorities about changes to enhance the ongoing effectiveness of the two interventions.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

Australia is one of the most advantaged and advanced countries in the world (OECD, 2018; UNDP, 2018). The reviews commissioned by successive Australian and other governments around the world, research and related agencies (See section 2.2) have argued that the best way to retain this advantage is to develop the creativity and cognitive skills of its people, with a particular emphasis on Science, Technology, Engineering and Mathematics or STEM as it is also known as (JFF, 2007; DES, 2003; OCS, 2014) and to aim for world's best practice in doing so.

That Australia's aspirations are global is evidenced by its membership of and active participation in OECD projects related to assessment, for example,

- ongoing participation in its Programme for International Student Assessment (PISA) since its inception in 2000 (OECD, 2014);
- case studies of classroom practice in Queensland schools were included in their *What Works* series of publications, for example, a study on Formative Assessment (CERI, 2005, pp. 191-204); and
- participation in the *OECD Reviews of Evaluation and Assessment in Education* series (OECD, 2011).

Section 2.3 reviews the research literature on assessment and discusses the idea that schools are enmeshed in a complex web which is appropriately called an assessment system.

Section 2.4 discusses the purposes of assessment and how theories of learning and cognition impact what and how we assess.

Section 2.5 examines the concept of assessment as measurement and explores that idea in relation to summative and evaluative purposes for assessment.

Section 2.6 looks at the new emphasis being given to formative assessment and its contextualisation in teaching known as formative practice and why this may be the key to helping students become life-long learners.

Section 2.7 describes the evolution of the SOLO model and positions it as a generic, developmental learning progression that enhances the feedback potential of summative tests such as the EV test.

Section 2.8 reviews the main ideas discussed above and that have informed this study.

2.2 A curriculum, teaching and assessment for the twenty-first century

In April 2005, Carmel Tebbutt, the Minister for Education in New South Wales (NSW), Australia, announced to the NSW Parliament:

> There is no doubt that science and technology are integral to our modern society [and] we must do all we can to encourage students to take up science and to continue its study in years 11 and 12. The Government is introducing for year 8 an essential secondary science assessment to help improve learning outcomes and generate student interest in studying science (Tebbutt, 2005, p. 14956).

The first sentence from this quote is a strong statement of the need, at least in the eyes of the then NSW government, to ensure that more students engage with science until the end of their senior secondary schooling. The basis for this claim will be outlined later in this section. The second sentence is a reference to the EV program described in Chapter One. As will be explained later in this chapter, imposing a test is a tool used by governments to signal to the community the importance placed by government on aspects of the curriculum, in this case science (along with literacy and numeracy as will also be explained below).

In her speech announcing the introduction of the EV program, minister Carmel Tebbutt, explicitly referred to a report from a review into innovation, science, technology and mathematics teaching and teacher education in Australia (CRTTE, 2003), also known as the Dow Report. That review was a contribution to the then national government's broader agenda to

> promote research, development and innovation [because the Australian economy was transitioning from one based on] land, labour and capital to one based on human and intellectual capacity. (Australia, 2001, p. 4).

This emerging new economy was referred to as the knowledge economy in many of the reports prepared for governments in Australia such as the Dow report referred to above (CRTTE, 2003) and around the developed world at that time (OECD, 1996). All were anxious to ensure that all continued to prosper into the future.

In one such report, the then chief scientist for Australia, Robin Batterham, wrote: "Science, engineering and technology underpins our future as a thriving, cultured and responsible community" (Batterham, 2000, p. 9). His report identified that more investment must be made in people and culture, ideas and commercialisation if Australia was to keep up with the rest of the developed world. His recommendations for doing so were based on his analysis of "initiatives and consequential structural changes underway…in OECD and Asian countries" (p. 41), including the United States, the United Kingdom, Canada, Japan, Finland, Ireland, Singapore and The People's Republic of China.

Batterham's proposed strategies and recommendations for keeping up with the changes going on in the world economy were aimed at ensuring that a growing number of students were prepared for science, engineering and technology (SET) related work. Among the strategies he identified were: making lifelong learning a key strategy for education providers and employees, inspiring students to study SET-based subjects, rewarding excellent SET teachers, providing specialist intensive training for teachers, and providing opportunities for SET graduates already in the workforce to enter the teaching system. The need for more students in Australia to engage with STEM in the later years of school and beyond was affirmed in the Dow report (CRTTE, 2003) referred to above and, in fact, most of the strategies and related recommendations from Batterham's report were repeated and endorsed in the Dow report (CRTTE, 2003).

The national Australian Education Council was pursuing an agenda to broaden the school's curriculum to better equip the growing number of students completing six years of secondary schooling with skills that better prepare them for work as well as success in tertiary studies. The Matters and Curtis (2008) report to the Australian Government Department of Education, Employment and Workplace Relations (DEEWR) described how five competences first proposed by the Karmel review (QERC, 1985) ended up as "Key Competencies" (AECRC, 1992) which were then handed over to state and territory education systems. A summary of this Key Competency work is included in Appendix A.

There was a trial of the key competencies in NSW schools, TAFE institutes and workplaces, which were defined as "the integrated application of knowledge, skills and understandings" (Ryan, 1997, p. 5). The trialing in schools was found to be "broadly supported by practitioners involved in the field testing [...but there was] little support for a separate additional layer of assessment and reporting that focuses on key competencies". (Ryan, 1997, p. 7) As will be apparent from a reading of the fourth section of the Table in Appendix A, the key competencies were later written into the NSW science syllabus (BOS, 2003) which contained the curriculum of interest for this project. Thereafter, the extent of Key Competency acquisition was assessed by teachers in the context of content and skills related to the separate learning area syllabuses, including science.

Of note too was the syllabus expectation that after four years of science teaching in NSW, students would emerge as independent learners who were "creative, responsible, scientifically literate, confident, [and] ready to take their place as a member of society." (BOS, 2003, p. 10) This aspiration was mentioned in the *Adelaide* Declaration (see Appendix B) as well as in Batterham's (2000) report.

The push from employers and government to broaden the curriculum's purpose from preparation for tertiary study (Connell, 1985) to preparation for life in the twenty-first century was expressed in three agreements between the national, state and territory education ministers about national goals for education which were subsequently endorsed by governments. The first of these was the *Hobart Declaration* (MCEETYA, 1998) with ten *Common and Agreed National Goals for Schooling* released in 1989. The goals were subsequently revised and endorsed in the *Adelaide Declaration* (MCEETYA, 1998) which was released in 1998. Following a review some ten years later a further iteration was published in the *Melbourne Declaration* (MCEETYA, 2008). Each *Declaration* was accompanied by an action plan. ACARA was created as a consequence of government commitments to the action plan attached to the *Melbourne Declaration*. The three sets of goals are included as Appendix B.

The above outlines the influences being brought to bear on the curriculum for schooling, including the science curriculum. The Dow report (CRTTE, 2003) also included reference to a recently completed comprehensive review into science teaching in Australian schools titled *The Status and Quality of Teaching and Learning of Science in Australian Schools* (Goodrum et al., 2001).

Goodrum et al. (2001) included a table adapted from the USA *National Science Education Standards* (NRC, 1996). The table summarised traditional science teaching practices found around the world and in Australia (left hand column) with practices supported by the research literature as being more effective (right hand column). The table from the review is published here as Table 2.1. The more effective approaches are summarized in the right-hand column. Three of the last four points in the right hand column are italicized and bolded by the thesis writer to highlight specific references to assessment and how it needs to change when compared to modal practices (see corresponding points in the left-hand column) at that time.

Table 2.1

Summary of needed changes to teaching and assessment

Teaching for scientific literacy requires:	
Less emphasis on:	More emphasis on:
memorising the name and definitions of scientific terms	learning broader concepts that can be applied in new situations
covering many science topics	studying a few fundamental concepts
theoretical, abstract topics	content that is meaningful to the student's experience and interest
presenting science by talk, text and demonstration	guiding students in active and extended student inquiry
asking for recitation of acquired knowledge	providing opportunities for scientific discussion among students
individuals completing routine assignments	groups working cooperatively to investigate problems or issues
activities that demonstrate and verify science content	open-ended activities that investigate relevant science questions
providing answers to teacher's questions about content	communicating the findings of student investigations
science being interesting for only some students	science being interesting for all students
assessing what is easily measured	***assessing learning outcomes that are most valued***
assessing recall of scientific terms and facts	***assessing understanding and its application to new situations, and skills of investigation, data analysis and communication***
end-of-topic multiple choice tests for grading and reporting	***ongoing assessment of work and the provision of feedback that assists learning***
learning science mainly from textbooks provided to students	learning science actively by seeking understanding from multiple sources of information, including books, Internet, media reports, discussion, and hands-on investigations

Source: Figure 7.1 in Goodrum et al.,2001, p. 168.

In the Australian context, Goodrum et al. (2001) had identified that

> most secondary science teachers are concerned about the final assessments for students which determine access to tertiary education and they regard covering the content likely to be assessed as of paramount importance, the repercussions of which echo right down to the early years of high school (p. 145).

The reviewers were concerned about that focus on "final assessments" and their recommendations for change identified assessment as an area for reform. Three of the last four points in the right-hand column are about assessment. The third one received special mention in their recommendations.

> Recommendation 7: It is recommended that the Commonwealth assist educational jurisdictions to reform assessment practice so that assessment more effectively serves the purpose of improving learning. Assessment must focus on the learning outcomes associated with scientific literacy. (Goodrum et al., 2001, p. xiii)

Subsequently, two of the review report authors were commissioned to prepare a five-year action plan (2008 to 2012) to manage the continuing implementation of recommendations from that initial report. (Goodrum & Rennie, 2007) Assessment was one of eight areas for action. The overriding objective of assessment reform, they wrote, was to "improve the quality of student assessment by ensuring that it was aligned with intended learning outcomes." (p. 15).

Two priority actions to achieve this objective were described in their report. The first was for "effective [use of] diagnostic, formative and summative assessment approaches" (p. 16) to be embedded in curriculum resources developed to support science teaching. The second was to "monitor performance in science at the national level" (p. 16). Goodrum et al. (2001) recommended that the latter be done by national sample testing of students.

In response to the first proposed action, two major Australian curriculum support initiatives subsequently modelled the use of assessment for diagnostic, formative and summative purposes as recommended. These were Primary Connections (AAS, 2016), which provides comprehensive support materials for science teaching in the K-6 years, and Science by Doing (AAS, 2017), which provides similar support for junior secondary science teaching.

The proposal for national monitoring of science performance was, in effect, an endorsement of current programs using existing sample testing programs, one an Australian initiative and the other two were international in origin. These programs test samples of NSW students in Years 4, 6, 8, 9 and 10 (fifteen-year-olds) and will be described further in this chapter.

(Broadfoot, 2009) observes that around the world externally set, test-based assessments are being used

> ubiquitously to provide for selection, for certification, for accountability and for international comparisons of educational standards. The advent of the 21st century also heralded the early stages of a movement to promote the use of assessment as a tool to support learning itself. (p. vii)

The NSW syllabus being used at the time of this project with its emphasis on assessment for learning was an example of the later, as was the introduction of the EV program and the *Quality Teaching in NSW public schools* (*QT*) initiative. The *QT* initiative was a professional learning initiative of the Department to support and improve teaching and assessment in government schools. The QT initiative was a professional development program widely supported in NSW schools in the first decade after 2000. The syllabus message about assessment for learning was reinforced in the Department's QT initiative.

> Assessment is the process of identifying, gathering and interpreting information about students' learning. The central purpose of assessment is to provide information on student achievement and progress and to set the direction for ongoing teaching and learning. (DET, 2006, p. 5)

In addition to NSW resources supporting assessment for learning, three national projects had additional resources online for science teachers by 2005, including:

- material about assessment for learning (CC, n.d.);
- a range of diagnostic assessment items and tasks for science (ACER, 2004a); and
- online learning objects specifically targeting science learning that could be used by teachers and students for diagnostic purposes as well as to support science learning more generally.

Elements of these three programs are evident in teaching and learning support resources currently available to schools on Education Services Australia managed websites (ESA, n.d.), including *Improve* and *Scootle* (ESA, 2012).

By 2010, science teachers in NSW should have been very aware of expectations for their use of assessment for learning strategies and resources, including the use of assessment data from the EV program. As explained in Chapter One, NSW has chosen to retain and expand its EV test from Year 8 to include both Year 6 and Year 10, though for now the latter two tests are not mandatory. (DET, 2015) While there is considerable evidence that summative tests contribute to disengagement with learning (Darling-Hammond, 2003; Harlen & Deakin-Crick, 2002; Osborne & Dillon, 2008; Stiggins, 2007; Tytler, 2007), this thesis uses the context of the EV program to provide important insights into how large-scale, summative, externally designed tests are being used to improve both achievement in and engagement with learning.

2.3 Assessment and assessment systems

Some definitions of assessment and assessment systems are provided to introduce this section. These will be followed by a discussion of the literature relating to three common purposes for school assessment. The impact of current understandings about learning and cognition on assessment and the need to ensure that what is done in the name of assessment is fit for purpose complete the section.

The following five definitions of assessment are found in the literature.

The first is:

> The terms *educational measurement, assessment, and testing* are used almost interchangeably in the research literature to refer to a process by which educators use students' responses to specially created or naturally occurring stimuli to draw inferences about the students' knowledge and skills. (Popham cited in NRC, 2001, p. 20, italics in original)

The second is:

> [Assessment is] the process of gathering and interpreting information about the progress of students' learning. (Hackling, 2004, p. 127)

The third is:

> Assessment is a term that covers any activity in which evidence of learning is collected in a planned and systematic way and is used to make a judgment about learning. (Harlen & Deakin-Crick, 2002, p. 1)

The fourth is science specific:

> [Assessment is] the collection and interpretation of information about learners' knowledge, understandings, skills and attitudes relating to the science outcomes. (Goodrum et al., 2001, p. 20)

The fifth has alternative names for assessment, depending on what is being assessed:

> [Assessments are] judgements on individual progress and achievement of learning goals [from] classroom-based assessments, as well as large scale, external assessments and examinations … appraisal refers to judgements on the performance of school-level professionals, e.g. teachers, school leaders … evaluation refers to judgements on the effectiveness of schools,

> school systems, policies and programmes. (Nusche, Radinger, Santiago, & Shewbridge, 2013, p. 59)

The last definition relates to the system of assessments that schools are expected to participate in. The assessments involve collecting evidence of learning and evidence of performance that goes well beyond writing responses to pen and paper test items.

Participants in any discussion about assessment need to understand more than the literal interpretations of the words "evidence of learning". Two examples illustrate this: The first is the NSW Department of Education and Training's Quality Teaching (QT) initiative (DET, 2003) mentioned above. It suggests four questions.

1. What do you want the students to learn?
2. Why does that learning matter?
3. What are you going to get the students to do (or to produce)?
4. How well do you expect them to do it? (DET, 2006, p. 10)

A more sophisticated version of the context for assessment is provided in a National Research Council (NRC, 2001) report. The NRC manages seven programs for the US Academies of Science and Engineering, including their Behavioural and Social Sciences and Education programs. It draws on expertise from within and outside the academies as needed. For the NRC

> Assessment is always a process of reasoning from evidence … [and] is imprecise to some degree [and assessments] are only estimates of what a student knows and can do. (p. 2)

Every assessment involves three foundational elements (which the writers call the vertices of an assessment triangle):

> *a model of how students represent knowledge and develop competence in the subject domain* [cognition]; *tasks or situations that allow one to observe student's performance* [observation] *and an interpretation method for drawing inferences from the performance evidence thus obtained*

> [interpretation] … *These three elements—cognition, observation, and interpretation—must be explicitly connected and designed as a coordinated whole.* (p. 2, italics in the original)

A fundamental premise of the NRC (2001) report is:

> Most widely used assessments of academic achievement are based on highly restrictive beliefs about learning and competence not fully in keeping with current knowledge about human cognition and learning. Likewise, the observation and interpretation elements underlying most current assessments were created to fit prior conceptions of learning and need enhancement to support the kinds of inferences people now want to draw about student achievement. (pp. 2-3)

The NRC (2001) report makes this observation about assessment too.

> Much greater value and credibility [is attributed] to external assessments of individuals and programs than to classroom assessment designed to assist learning … *More of the research, development, and training investment must be shifted toward the classroom, where teaching and learning occur*. (p. 9, italics in the original).

This last sentiment was echoed in the Goodrum et al. (2001) review and recommendations mentioned in the previous section.

Since the beginning of the 1990s, students in Australia have been asked to sit tests imposed by education authorities outside the immediate school before their final year of schooling. In Years 7 and 9 all students sit literacy and numeracy tests once sat by state and territory education authorities. ACARA in the context of its National Assessment Plan Literacy and Numeracy (NAPLAN) program has taken over management of the tests since 2008. Year 8 students in NSW government schools at least sit EV tests for science. In many schools, science department buy tests developed by private testing companies (such as ICAS science tests produced by Education Assessment Australia (EAA). (EAA, 2018) These ICAS tests provide independent feedback on the level of science process skills students possess at the time they sit the test. The Australian Council for Educational Research (ACER) also provides comparable tests for science that schools can purchase to support their teaching and learning programs (Masters, 2009).

It is also possible, but less likely, that students could be asked to sit tests produced by two international agencies in reading literacy, numeracy and scientific literacy. The first organisation to bring these tests to Australia (in 1995) was the International Association for the Evaluation of Educational Achievement (IEA) (IEA, 2013). These provide testing and reporting in reading literacy (PIRLS) over a pentennial cycle and in mathematics and science (TIMSS) in a quadrennial cycle. The TIMSS tests are currently sat by Year 4 and Year 8 students; only Year 4 students sit PIRLS tests.

The second program is the OECD’s Programme for International Assessment of Students (PISA), which provides tests in literacy, numeracy and scientific literacy over a triennial cycle for 15-year-old students (OECD, 2014). Australia has participated in PISA since it began in 2000. The ACER manages the test processes in Australia for the IEA and OECD and it writes the reports for Australia from their analysis of the results and related surveys (Thomson, De Bortoli, & Underwood, 2017; Thomson, Wernert, O'Grady, & Rodrigues, 2017).

Figure 2.1 is a representation that Nusche et al. (2013) used to examine and report against in their exploration of the assessment systems of participating OECD members, including Australia. The figure shows the complexity of the assessment system schools are now enmeshed in.

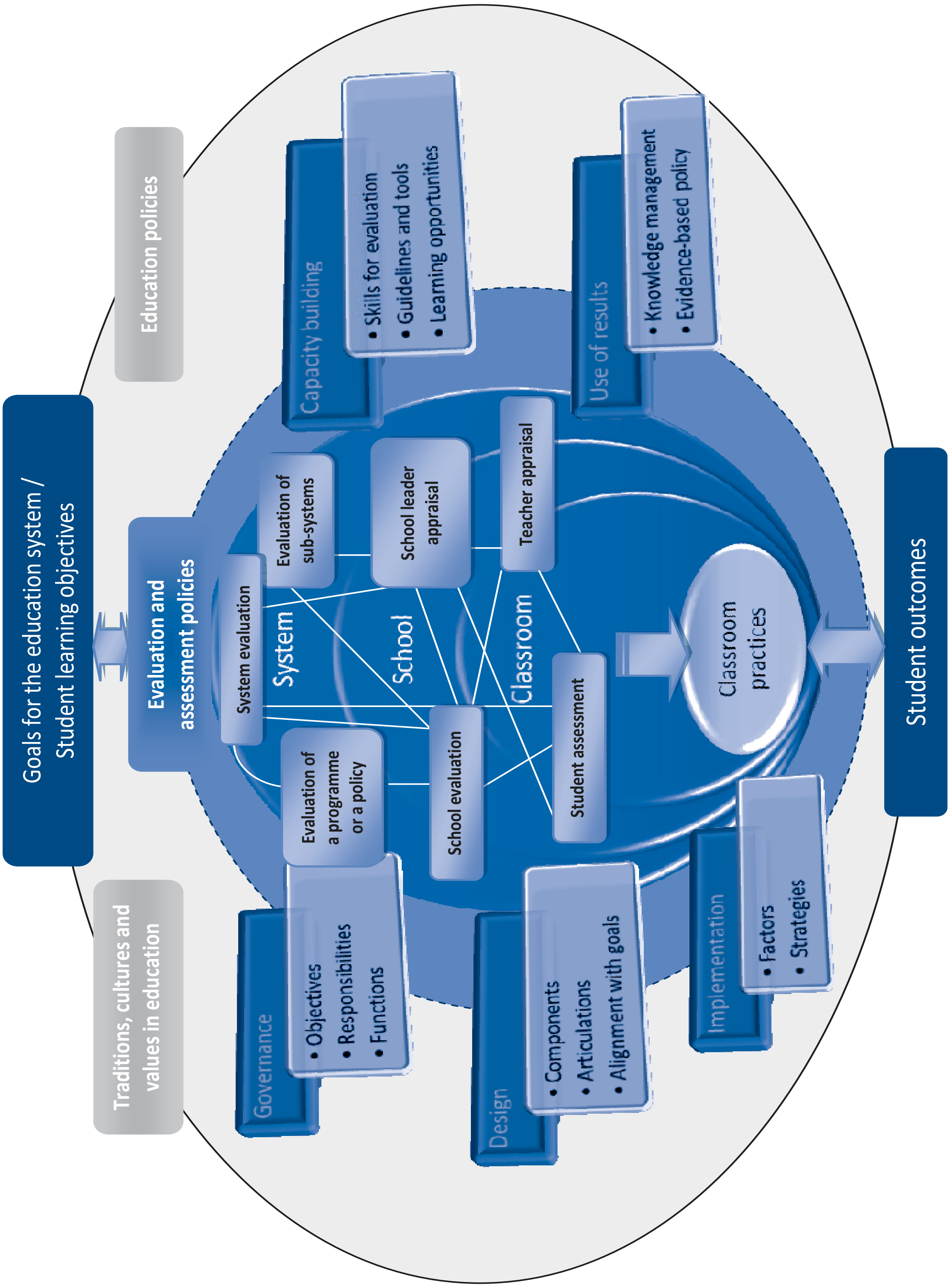


Figure 2.1 Components of an evaluation and assessment framework.

Note their articulation with each other to show how education system goals, student learning objectives and outcomes are aligned. Source: Nusche et al., 2013, p. 60.

In relation to that system, the NRC (2001) report says:

> Aspects of learning that are assessed and emphasized in the classroom should ideally be consistent with (though not necessarily the same as) the aspects of learning targeted by large-scale assessments. (NRC, 2001, p. 3).

This is a call for vertical alignment of assessment intent. The claim here is that classroom assessments and externally imposed tests should all be defensible in terms of the national or state or territory goals the tests are supposed to be providing evidence of learning about.

The NRC (2001) report also asserts: "*Educational assessment does not exist in isolation but must be aligned with curriculum and instruction if it is to support learning*" (p. 3, italics in the original). This is a call for the horizontal alignment of assessment practices, learning expectations (as described in the curriculum) and instruction. Others expressing a similar view include Biggs (1999), Mansell, James & the ARG (2009) and Masters (2013). Alignment means that what is intended to be learnt (curriculum)and how it is acquired (instruction) and demonstrated as being acquired (assessment) are connected by a coherent and consistent view of learning and cognition. The Trends in Mathematics and Science Study (TIMSS) assessment model collects data based on the premise of horizontal alignment. (Mullis, Martin, Ruddock, O'Sullivan, & Preuschoff, 2009)

It is these alignments that the NRC (2001) says are often missing in the real world of practice. The report by the Council of the Great City Schools in the US provides examples of the consequences when those alignments are weak or missing (CGCS, 2015). The CGCS (2015) report findings are listed under six headings: assessments required of all students in a given grade; sample and optional assessments; assessments for special populations; looking at testing in a district context; costs of testing in a sample district; and parents. A summary of 23 separate points includes the following: mandated, external testing of Grade 8 students took up at least 2.34% of the school year; there was no correlation between the amount of mandated testing time and the reading and math scores in grades four and eight on the National Assessment of Educational Progress (NAEP) test program; some tests are not well aligned to each other or college or career reading standards and often do not assess student mastery of any specific content; and, parents support replacing current tests with “better” tests. Despite these issues, 82% of the school parents polled expressed support or strong support for “[having] an accurate measure of what my child knows” (pp. 9-11).

Broadfoot (2009) describes a four-dimensional characterisation of assessment systems for analysing the links between the assessment system and the social context in which it is embedded. The components are purposes; mode (means used to gather evidence of learning); content (what is being assessed); and organisation (how assessments are conducted). She argues that the prevailing social context in Western societies at the end of the 20th century was dominated by enlightenment and modernist sentiments to do with “individual rights and responsibility, rationality and scientific progress” (p. vi). There was also enormous investment made during the 20th century into “the pursuit of maximum accuracy in educational measurement” (p. vii).

Broadfoot sees measurement as the purpose of assessment in this social context. The higher the score the more social merit was bestowed on that person, who then, presumably, could go on to be anything they wanted to be in life (scientist, doctor, lawyer, pilot and any other high social status job they desired). The content to be assessed was the curriculum content that could be measured. The dominant mode of assessment was by ‘paper and pencil’ testing. The evidence of learning it delivered ranged from a letter representing the best response (from several options provided) to writing a few words or the result of a calculation, or an extended response involving (one, some, or all of) calculations, annotated graphic representations (flow charts, diagrams, tables and graphs) and text types characterising description, explanation, justification or a creative synthesis. TIMSS, PISA and NAPLAN tests use a mix of short response items and extended response tasks; the balance being in favour of short response items (typically between 60-80%). Correct responses were counted and summed. In this context the bigger the number the better the result.

The typical organisation for external, standardised tests assumes responses will be from individuals and provided within a strictly imposed time limit, and that the test and answer booklets would be produced, printed, delivered, collected, collated and coded in processes managed by the agency responsible for the test (or their delegate). Large-scale test scores (raw scores) once obtained would often be standardised in a variety of ways using statistical procedures to ensure a fair basis for comparability.

Many teachers and others in the community beyond schools believe that this mode of assessment provides an objective, unbiased and thus fair assessment of individual performance at the time the test is taken. Support for this generalisation has been expressed in international and local (Australian and NSW) reports and research papers reviewing large-scale assessment programs, such as those mentioned above and the recently abandoned Year 10 tests in NSW. Examples include Cooney (2006), Smith (2005) and Wasson (2009) in respect of the NSW literacy and numeracy tests; BOS (2011) for the now abandoned Year 10 tests in NSW; and Thomson, Wernert, et al. (2017) and Thomson, De Bortoli, et al. (2017) for the latest TIMSS and PISA reports respectively. In the US, the NRC (2001) supports the NAEP (2011) test model.

Broadfoot (2009) goes on to identify a change developing in how the education community views assessment that she associated with post-modernism.

> [This] movement sees assessment as a tool to support learning … involvement of human beings in every aspect of its design, execution and use makes [testing] irrevocably a social project and thus subject to all the vagaries that any kind of human activity implies … assessment in the 21st century shows signs of a growing preoccupation with 'fitness for purpose' and impact on learning. (p. vii)

This emerging view supports the move away from seeing assessment as a summative program to a formative one (as evidenced in the NSW Science syllabus of interest here). The EV program is an attempted shift in that direction. It uses a summative test to provide feedback on learning with the expectation that test results be used formatively by teachers to improve science learning and engagement. How this has worked out in practice in NSW is reported on in the concluding chapter of this thesis.

The OECD's (2011) report and recommendations on the Australian evaluation and assessment system mentioned at the beginning of this chapter was based on Australia's submission to the OECD review process and the observations of an independent OECD panel that visited Australia in June 2010. The panel concluded:

> The overall evaluation and assessment framework [in Australia] appears as highly sophisticated and well conceptualised, especially at its top level (national and systemic levels). However, there is a less clear articulation of ways for the national agenda to generate improvements in classroom practice through the assessment and evaluation procedures which are closer to the place of learning. (OECD, 2011, p. 9)

Of interest though, is the inclusion of two Australian case studies of formative assessment in one of the OECD's *What Works* publications on formative assessment (CERI, 2005). These local examples of good classroom assessment and school support for assessment are models that could be applied more widely in Australia to address the panel's conclusions.

2.4 The purposes for assessment

The locus of interest for this study is teachers' assessment-related work. The following discussion about purposes for assessment will focus on classroom and school assessment-related work.

The NRC (2001) report posits three purposes for assessment:

1. to assist learning
2. to measure individual achievement
3. to evaluate programs. (p. 3)

The NRC (2001) says that these three purposes hold for classroom and large-scale tests as well.

In the UK, The Economic and Social Research Council's (ESRC) report, *Assessment in Schools. Fit for purpose? A Commentary by the Teaching and Learning Research Programme* (Mansell et al., 2009), identified the same three uses (or purposes) for assessment:

1. to help build pupils' understanding, within day-to-day lessons
2. to provide information on pupils' achievements to those on the outside of the pupil-teacher relationship: to parents (on the basis of in-class judgments by teachers, and test and examination results) and to further and higher education institutions and employers (through test and examination results)
3. to hold individuals and institutions to account, including through the publication of results which encourage outsiders to make a judgment on the quality of those being held to account. (p. 8)

The first purpose in both reports is also referred to in the literature as:

- classroom assessment (Black & Wiliam, 1998b; Brookhart, 2003; Cowie, 2005, 2013; Marzano, 2000; Ruiz-Primo & Li, 2012; Shepard, 2001; Stiggins, 2004)
- formative assessment (Bell & Cowie, 2002; Black & Wiliam, 2009; Heritage, 2010; Panizzon, Callingham, Wright, & Pegg, 2007; Sadler, 1998; Stiggins & DuFour, 2009)
- assessment for learning (ARG, 2002a; Biggs & Collis, 1982; Hargreaves, 2005; Stiggins, 2002; Wiliam, 2011b)
- embedded assessment (mainly in the US) (Wiliam, 2011a; Wilson & Sloane, 2000).

The second purpose in both reports is often referred to as summative assessment (Biggs, 1998; Harlen & Deakin-Crick, 2002; Harlen, 2005) or assessment of learning (ARG, 2006; Hackling, 2004). Historically, summative assessment attracts considerable individual and/or public attention when results that have real consequences for those receiving them are publicised, delivered, used or recorded for later use. For that reason, summative assessment is also called high-stakes assessment (Au, 2007; Broadfoot & Black, 2004; Dulfer, Polesel, & Rice, 2012; Gipps, 1999; Klenowski & Wyatt-Smith, 2012; Lim, Tan Eng Thye, & Kang Lu-Ming, 2009). The last citation relates to Singapore's education system requirements.

The third purpose relates to accountability. Often the results of summative assessments are the basis for monitoring the performance of a school or school system. The issues related to high-stakes assessment are discussed by researchers listed for summative assessment also apply in this context.

2.4.1 Three purposes for assessment?

In an editorial reviewing the first 10 years of the UK journal *Assessment in Education: Principles, Policy & Practice,* Broadfoot and Black (2004) asserted:

> Educational assessment must be understood as a *social* practice, an art as much as a science, a humanistic project with all the challenges this implies and with all the potential scope for both good and ill in the business of education. (p. 8)

The editors go on to identify from the papers published in those years a "subset of subtle purposes, which serve to underline the pervasive [social] power of assessment to define and shape every aspect of educational life" (pp. 11-12), including:

- as a mechanism for controlling class behaviour and attention (the threat of poor results!)
- to describe achievement standards in terms of qualitative changes in the response capabilities of students over time. This was a reference to work done in Australia in the first half of the 1990s to develop subject 'Profiles' for a national curriculum (Rowe & Hill, 1996)
- the use of assessment to encourage 'deep' rather than 'surface' learning

- encouraging ownership (by both teachers and students) of assessment as an influence on their capacity and motivation to learn
- the growing use by policy makers of the social power of assessment in attempts to raise achievement levels, change the focus of curriculum priorities, in performance management for teachers, institutional quality assurance and control and, defining ‘standards’ through the publication of league tables.

Matters and Curtis (2008) refer to attempts by policy makers to change the focus of curriculum priorities as “signalling” (p. 17). The writers use the term in the context of government efforts to have key competencies embedded in school level curriculum documents assessed by teachers. The message from government was that this content was of equal value to the other content in, say, the NSW science syllabus. In its imposing of the EV program on schools, the NSW government was signalling to students, teachers and the wider community its view of the relative importance of science in the curriculum (see the quotation opening Section 2.2). The same could be said of the decision to introduce sample testing of Year 6 students in science literacy every three years (Ball, Rae, & Tognolini, 2000) and the decision to participate in international testing of science.

The second, third and fourth purposes are linked to formative assessment and will be explored in Section 2.6. The fifth cluster of purposes identified here relates easily to the third purpose of assessment identified by both the NRC (2001) report and Mansell et al. (2009) commentary.

It is evident that discussions about assessment and meanings of related terms can be a source of confusion. Newton (2007) is a UK-based expert and researcher with wide experience in assessment. Based on his experience of discourse about assessment purposes, he reports that the phrase ‘assessment purposes’ may be interpreted in at least three ways. The first is a reference to the technical aim of the assessment, which is to make a “judgment” (p. 150) that is typically referred to as the result (this he calls the first or judgment level). “Judgment” and the NRC’s (2001) “interpretation” in the context of the “assessment triangle” (NRC, 2001, pp.

2-3) are equivalent, but it is worth observing that the word "judgment" has moral overtones. "Interpretation" is a neutral, objective, technical word. The word judgment is perhaps an intended, if implicit, reminder of the social power vested in assessment. (Broadfoot & Black, 2004)

Newton (2007) analysed historical publications about assessment to explain how first level judgments might be better expressed to clarify the various forms assessment might take. To do this he resorted to technical descriptions of the various judgments a professional working in the assessment area might use. He distinguished between quantitative, summative judgments involving appraisal, and qualitative, descriptive judgments involving analysis at the two ends of a judgment dipole. The former might be either self-referenced or norm-referenced judgments. The latter may be either concept-referenced judgments or performance-referenced.

The second way is about the use to which the assessment result is put (the decision level). Newton (2010) produced a list of 22 "categories of uses for assessments", including social evaluation, formative assessment, diagnosis, screening, segregating, guidance, program evaluation, and institutional monitoring. The Mansell et al. (2009) commentary reference to uses for assessment rather than purposes acknowledged Newton's work in this area.

The same set of test results are sometimes used for multiple purposes, often inappropriately (James, 2009; Newton, 2007). The NRC (2001) report makes the same observation. Compare this with James's (2009) observation that "twenty years ago … test and examination results were predominantly meant to serve as indicators of what a pupil knew and understood about a subject" (p. 8). Multiple uses for the same set of results were acknowledged in evidence to a UK House of Commons Select Committee (SCCS&F, 2008) along with an acknowledgment the same test was not always the most appropriate for all purposes.

Newton's (2007) third way of interpreting 'assessment purposes' relates to the intended impact of testing, which is to signal the importance of the learning (so important that it will be tested!). Newton (2007) also recognises unintended, negative impacts for both second-and third-level uses and intentions. The notion of impact is an explicit recognition of the 'principle' that assessment is a social act because assessment results both convey information and influence what people do (Mansell et al., 2009).

Put another way Fensham and Rennie (2013), Jones and Buntting (2013), and Millar (2013) all agree that what is assessed has a powerful influence on what is taught (or not taught). An example is school reporting of achievement in science at the end of Year 10 to the NSW Board of Studies. Schools are advised that the results should not include any consideration of achievement of syllabus outcomes related to values and attitudes. On the other hand, the advice relating to investigation skills is explicit about what is to be included (BOS, n.d.).

Some writers have sought to frame assessment in terms of functions rather than purposes. For Hattie(2003a), assessment is not about the test itself, it is the function that matters. Test results, he asserts, function as feedback to

> …teachers and/or students … which they need to interpret when answering the three feedback questions: Where am I going?, How am I going? and, Where to next? Specifically, feedback is actions or information provided by an agent (e.g. teacher, peer, written report, book, parent, experience) that provides information regarding aspects of one's performance or understanding. (p. 2)

Hattie argues that these three questions work for all levels of the assessment system, and feedback combines judgment and action (either proposed or actual).

Masters (2013, p. 2) proposes that the overriding function of assessment is to provide understanding, not judgment. He uses the analogy of a doctor-patient consultation where the doctor is trying to elicit the symptoms from a patient in order to diagnose the illness and then propose actions to cure the patient. Extending this analogy, he says, "*The fundamental purpose of assessment is to establish where learners are in their learning at the time of the assessment*" (Masters, 2013, pp. 5-6, italics in the original).

In this scenario Masters wants to remove the pejorative judgment (of pass or fail) and replace it with understanding as the basis for further action. Both Hattie and Masters share a view of learning as a continuous process that can be assisted by a timely diagnosis and appropriate intervention. Both researchers see the primary role for assessment as improving student learning.

2.4.2 Theories of learning, cognition and assessment

At the beginning of this section, an overview of what a reader needed to bring to a productive discussion about assessment was outlined. What a stakeholder in education understands about learning and cognition informs what they believe is important to learn and how they explain why it matters. It also informs the construction or choice of tasks to provoke responses from students, the interpretation of those responses (in terms of the assessors understanding of curriculum intentions), and the representation and explanation of the judgment (the result) about learning inferred from the responses to assessment tasks.

Two examples of where teachers found theories of learning and cognition helpful follow. In the first, Black, Harrison, Lee, Marshall, and Wiliam (2004) found that UK secondary science, mathematics and English teachers the researchers were working with in an effort to improve formative assessment practices wanted to know more about "the psychology of learning" (p. 16). Teachers wanted a model of how students learn that would be useful for providing feedback to students. In the second example, Panizzon et al. (2007) found that when participating teachers were given the SOLO theory of cognition, teachers found it useful for planning assessment tasks and restructuring science learning programmes to reflect the developmental changes anticipated by the SOLO model.

A discussion of learning theories and their relationships with assessment follows. According to the NRC (2001) report,

> Most current tests, and indeed many aspects of the science of educational measurement, have theoretical roots in the differential and behaviorist

> traditions. The more recent perspectives—the cognitive and the situative—are not well reflected in traditional assessments. (p. 60)

Biggs (1995) wrote:

> Two basic conceptions of the nature of learning exist in our educational thinking, quantitative and qualitative … the quantitative tradition has the longest history [and stems from] the positivist tradition in the social sciences … The qualitative tradition has its roots in nineteenth century phenomenology [and] Gestalt psychology. [Both of which later contributed to a family of learning theories underpinned by] *constructivism*. (pp. 2-5, italics in the original).

The quantitative assessment tradition is associated with behaviourist theories of psychologists such as Edward L. Thorndike and B. F. Skinner who conceive learning as acquiring

> discrete quanta of declarative or procedural knowledge; as far as assessment was concerned, any one quantum is treated as functionally independent of any other. The curriculum becomes in effect a list of discrete units: facts, skills, competencies, behavioural objectives, performance indicators, and the like and assessment a matter of how many. (Biggs, 1995, p. 2)

From this perspective, teaching or instruction is

> conceived as transmitting knowledge from teacher to learner…the teacher's task is to know the subject and expound it clearly, the learner's to receive it accurately [and] assessment [involves the] correct units being summed to give an accurate score that yields an index of competence in what is learned. (Biggs, 1995, p. 2)

The quantitative assessment instrument of choice was the multiple-choice test. If essays were used, the marking rubric identified units that would be considered correct or acceptable and 'full marks' would be awarded when enough correct units were evidenced. A good test would have a range of units at varying levels of cognitive difficulty but the units would all be treated as having "mutual equivalence, independence, and additivity." (Biggs, 1995, p. 3)

The behaviourist perspective emerged in the 1930s "about the same time that theories of individual differences in intellectual abilities were maturing" (NRC, 2001, p. 61). According to behaviourists such as Thorndike (cited in NRC, 2001),

> People learn by acquiring simple components of a skill, then acquiring more complicated units that combine or differentiate the simpler ones. Stimulus-response associations can be strengthened by reinforcement or weakened by inattention. When people are motivated by rewards, punishments, or other (mainly extrinsic) factors, they attend to relevant aspects of a situation, and this favors the formation of new associations and skills. (p. 61)

By contrast, the qualitative, constructivist or cognitive perspective comprises

> a family of theories rather than any one, according to which students are assumed to learn cumulatively, actively interpreting and incorporating new material with what they already know. Different theories variously emphasize the individual, social, cognitive, saccadic, contextual or emergent natures of learning, but all agree on an active learner seeking meaning by constructing knowledge rather than by receiving and storing knowledge. (Biggs, 1995, pp. 3-4)

In this perspective, the teacher's role is to help students "construct understandings that are progressively more mature and congruent with accepted thinking" (p. 4). The teacher should also recognise that students everyday experiences and prior learning will inevitably lead to naïve or alternative conceptions (Driver & Easley, 1978) of how the world works, and these need to be challenged and reoriented to better reflect the scientific viewpoint. A constructivist model of teaching and learning is the 5Es approach, as advocated in the *Science by Doing* curriculum support materials produced by the Australian Academy of Science (AAS, 2017).

From the qualitative perspective, assessment

> implies aggregating units of learning taken cross-sectionally with respect to time, that from the qualitative tradition implies charting longitudinal growth over time, from relative ignorance to relative competence … If that growth in competence can be described in recognizable stages then so much the better, because these stages can then become assessment targets. (Biggs, 1995, p. 4)

Biggs (1995) then describes two kinds of assessment that have emerged from constructivist thinking. One he describes as ecological, which appears to equate with what others have called performance or authentic assessment (Frey & Schmitt, 2007); the other he describes as developmental assessment. It is the latter that he goes on to elaborate as "a generalized model of qualitative assessment" (p. 6) and associate with the SOLO Taxonomy (Biggs & Collis, 1982). The SOLO Taxonomy and SOLO model will be described later in this chapter. Whilst Biggs (1995) positions the SOLO taxonomy as a qualitative developmental model, the later SOLO model has been validated both empirically and in measurement model terms as well (Panizzon & Bond, 2007).

The situative view of learning provides support for those arguing that assessment should be authentic, such as Darling-Hammond (2003); Fensham and Rennie (2013); Hackling (2004); Tytler (2007); and Wiggins (1998). The NRC (2001) writers say of this perspective:

> Much knowledge is embedded within systems of representation, discourse, and physical activity. Moreover, communities of practices are sites for developing identity—one is what one practices, to some extent. (p. 89)

In addition, standard assessment models take a view of knowledge as "disembodied and incorporeal [and it] captures only a small portion of the skills actually used in many learning communities" (NRC, 2001, p. 89).

The situative view of learning supports recent efforts to provide contexts for both the learning of science in the syllabus of interest for this project (BOS, 2003) and the framing of science as a human endeavour, and to engage students with science and encourage them to see themselves doing STEM work, post-school.

Vygotsky's (1978) concept of the zone of proximal development has been influential in the situative or socio-cultural view of learning. Shayer (2003) provides a commentary on both Piaget's and Vygotsky's views of cognitive development in children to support his particular intervention aimed at accelerating cognitive development.

Another contribution to the discussion about learning and related conceptions of assessment is that by Sfard (1998). Her contribution bridges behaviourist and cognitive (constructivist) and socio-cultural views of learning. She suggests that two metaphors are useful for understanding learning: the learning as acquisition metaphor (AM) – we acquire concepts or knowledge; and the learning as participation metaphor (PM). In the context of AM, assessment is about the quantity of what has been acquired. In PM, assessment is about a process of knowing, with the permanence of *having* giving way to the constant flux of *doing*. This metaphor implies that learning a subject is about "becoming a member of a certain community" (p. 6). AM is about the individual; PM is about the social.

Millar (2013) strongly advocates that both curriculum intention (what has to be done) and the assessment task (the conditions under which it is to be done as a demonstration of the acquired learning) should be provided in curriculum documents. "The assessment instrument *becomes* an operational definition of the [science learning] objective" (p. 56). Also, doing that would require teachers to acknowledge (if the task involved performance) a view of learning that acknowledges both AM and PM (Sfard, 1998).

An example of a teaching sequence that demonstrates a view of learning where both AM and PM are acknowledged is provided as Appendix C

The cognitivist perspective and related developmental approaches to assessment have informed work being done to elucidate learning progressions that span the years of schooling and span a topic of work lasting from five to ten weeks. The NRC (2001) report has a comprehensive and detailed discussion about developmental assessment and related terms, including progress maps, progress variables, developmental continua, progressions of developing competence, and profile strands (p. 137).

Of progress maps in general, the NRC (2001) report says

> The Developmental Assessment approach represents a notable attempt to measure growth in competence and to convey the nature of student achievement in ways that can benefit teaching and learning. (p. 190)

Rowe and Hill (1996) draw on both behaviourist and constructivist views of learning to provide an insight into the development of the Australian subject curriculum profiles ((CURASS, 1994) and outline their strengths and weaknesses from a developmental perspective.

Tom Corcoran's team at the *Centre on Continuous Instructional Improvement* (CCII) (Corcoran, Mosher, & Rogat, 2009) work on science learning progressions in the US. The team refer to the definition in the NRC-funded school science text book *Taking Science to School* edited by Duschl, Schweingruber, & Shouse (2007), which is widely used in the US:

> Learning progressions are descriptions of the successively more sophisticated ways of thinking about a topic that can follow one another as children learn about and investigate a topic over a broad span of time (e.g. 6 to 8 years). They are crucially dependent on instructional practices if they are to occur (p. 214).

Corcoran et al. (2009) use the term "adaptive instruction [to capture the sense of] formative assessment in action" (p. 8). This appears to be synonymous with the phrase assessment for learning that appears in the syllabus relevant to this study

(BOS, 2003) and with what Black and Wiliam (2009) call "formative practice" (p. 8).

2.4.3 Criteria for evaluating the credibility of assessments

In the context of explaining how to ensure the quality and credibility of assessments, researchers referred to a number of criteria that need to be addressed. Four examples of lists of criteria are provided in Table 2.2. The criteria apply from the level of classroom assessment to large scale external assessment.

Table 2.2
Issues to resolve when planning, constructing and using assessments

NRC (2001)	**Harlen (2005)**	**Matters and Curtis (2008)**	**Ruiz-Primo (2009)**
Identification of the targets for assessment Item and test design Validation Reporting Fairness	Validity Reliability Dependability	Validity and related constructs Reliability and related constructs Objectivity Feasibility Usability Credibility	Choose an approach to science instruction (eg inquiry… Identify the critical skills Define assessment purposes Define an appropriate approach for: Validity Reliability Fairness Issues of practicality

The change of state example (described in Appendix C) consider the "constructs" (NRC, 2001, p. 112) of physical and chemical change. An assessment task related to that example might involve providing students with access to a series of short video clips showing natural and 'made' changes. The task is to identify in each clip a process where either a physical or chemical change is occurring and to justify the choice.

The first consideration is *validity* (see Table 2.2 for the list of criteria). Do the video clips contain examples of the two types of changes? Do the images show aspects (construct dimensions) of the phenomena that are actual pointers or indicators of the changes to be recognised and associated with either a physical or chemical change and not something else? Is there evidence of other important learning that could be the subject of assessment (such as the practical value of the knowledge for safe use of materials and chemicals that could be inferred from the contexts on display in the video footage)?

Mansell, James & the ARG (2009) summarise the issue of validity in the context of teachers' summative assessments as "about whether the assessment measures all that it might be felt important to measure" (p. 12). In the above example, choices have to be made about whether the focus is on the processes of chemical change in isolation or whether students should be prompted to say something about its usefulness as well.

Messick's (1995) views on validity are widely cited in the research literature (e.g. Broadfoot and Black, (2004); Hattie, Jaeger, & Bond (1999); Masters, (2013); NRC (2001); Shepard (1993). Messick (1995) defines validity in the context of psychological and educational assessment as

> nothing less than an evaluative summary of both the evidence for and the actual-as well as potential-consequences of score interpretation and use (i.e., construct validity conceived comprehensively). This comprehensive view of validity integrates considerations of content, criteria, and consequences into a construct framework for empirically testing rational hypotheses about score meaning and utility. (p. 742)

He contrasts this more comprehensive approach to score interpretation with the historical

> primary emphasis in construct validation…on internal and external test structures—that is, on the appraisal of theoretically expected patterns of relationships among item scores or between test scores and other measures. (p. 743)

In essence, Messick (1995) is saying that the original construct for validity was located in the measurement paradigm for assessment (Biggs, 1995; Broadfoot, 2009) and he broadened it to encompass the concepts (constructs) that classroom teachers engage with every day and are looking to assess in the context of performances. Messick says this broader view of construct validity (see Table 2.3) depends on an appraisal of six aspects he identfies as "content, substantive, structural, generalizability, external, and consequential" (pp. 744-745).

Table 2.3
Messick's aspects of construct validity

content	includes evidence of content relevance, representativeness, and technical quality
substantive	refers to theoretical rationales for the observed consistencies in test responses, including process models of task performance, along with empirical evidence that the theoretical processes are actually engaged by respondents in the assessment tasks
structural	appraises the fidelity of the scoring structure to the structure of the construct domain at issue
generalizability	examines the extent to which score properties and interpretations generalize to and across population groups, settings, and tasks including validity generalization of test criterion relationships
external	includes convergent and discriminant evidence from multitrait-multimethod comparisons as well as evidence of criterion relevance and applied utility
consequential	appraises the value implications of score interpretation as a basis for action as well as the actual and potential consequences of test use, especially in regard to sources of invalidity related to issues of bias, fairness, and distributive justice

Source: Messick, 1995, pp. 744-5

Mislevy (2008) draws attention to research work that attempts to reconcile current psychometric models of assessment and recent views of cognition that include both cognitivist and sociocultural or situative perspectives.

> Cognition, in this view, is not just something that happens inside individuals' heads, but a coordinated interplay of actions within and among people in a socially-structured space. (p. 6)

Mislevy (2008) explores the impact of sociocultural views of learning on the traditional measurement models based on cognitivist views of learning and concludes that (latent) trait or item response theory "still holds under a sociocognitive metaphor, but with an interpretation quite different than that of the strict measurement metaphor" (p. 13). Latent trait theory ascribes a range of consistent behavioural responses to underlying, invisible but stable mental constructs such as ability, aptitude, expertise and intelligence. He also reports that another line of inquiry is finding that

> Models adapting features of generalizability theory, cognitive diagnosis, and standard measurement models would seem to be a suitable starting point for a psychometrics to support assessment under the sociocognitive metaphor. (p. 13)

The second criterion in Table 2.2 is *reliability*. Would different assessors score student responses the same way? Would the same assessor score a comparable response the same way? Would a student answering a comparable question on a different day answer the same way? And what does comparable mean in any case? Well-constructed marking criteria and rubrics help to ensure consistency of marking (an aspect of reliability), as would some prior practice using them before marking actually commenced. As well, check-marking by another assessor of a random sample of already marked scripts is another way of ensuring inter-marker reliability.

Fairness (see Table 2.2) is ensuring that students have had opportunities before the test to learn about physical and chemical changes and the differences between them. At one level, this can be an issue in Grade/Year cohort testing in schools where more than one class of students sit a common test. It can also be an issue with external testing when the curriculum used to prepare for the test is different across the various sites taking the test. In the UK and Australia, external testing is based on national curriculums that describe standards and related content that students taking the test have (or should have) been "taught". In the US, curriculum choice rests with individual school district boards. Large-scale external testing has to be more about general capabilities linked to assumed, common domain-specific knowledge that may or may not have been "taught" (Ruiz-Primo, Shavelson, Hamilton, & Klein, 2002).

Another consideration is the choice of assessment task and the opportunities it provides for different students (say from a background where English is not spoken at home) to respond. Returning to our chemical change/physical change example, would a deaf or blind student be able to score the same as a student with normal hearing and vision, given that the testees are not able to observe all possible evidence? (e.g. a deaf student would not hear heated corn popping, and a blind student would not see it). Some students may not have had any experience of pop-corn at all. Do testees need to write a response or simply tell the assessor what is going on? How many correct responses is required to demonstrate proficiency? The fairness and equity of assessment issues raised here are all related to assessment validity (Messick, 1995).

Dependability (see Table 2.2) involves making a defensible trade-off between validity and reliability. In the context of teacher summative assessment, Harlen (2005) says:

> Dependability is a combination of the two, defined in this instance as the extent to which reliability is optimized while ensuring validity. This definition prioritizes validity, since a main reason for using teachers' assessment rather than depending entirely on tests for external summative assessment is to increase the construct validity of the assessment. (p. 213)

In assessing student responses to the above task (distinguishing between physical and chemical changes), short response items (or even multiple-choice options) may increase reliability, but options for extended responses could include applications and reasons for choosing either chemical or physical change. The latter options improve validity but are harder to score reliably.

Objectivity is mentioned by Matters and Curtis (2008) as it is often raised as the bulwark for *fairness*. However, if the assessment is complex, such as marking an essay, it might be worth attending to the "objectivity and fairness of those who assess student work" (p. 15). The concern here is marker bias and how to ensure it does not affect or distort the application of the assessment criteria (see discussion about reliability and dependability above).

Feasibility/practicality (Table 2.2) also needs to be taken into account. According to Matters and Curtis (2008), "Feasibility means capable of being done, with the connotation of convenience and practicability in the doing. While many things are "doable, fewer are feasible" (p. 15). In the context of a national program, cost- and time-effectiveness are important considerations, and in a school context, resources and time factors are considerations.

In the example provided above, should the teacher provide the video clips and questions on a USB drive or allow students to access them via the school's intranet? The higher the cost in terms of time and resources, the more important it is to explain the benefits of what is being done. In large-scale testing of science in the Australian context, performance tasks involving an investigation were included in the national sample tests for Year 6 science (National Assessment Program-Scientific Literacy (NAP-SL) tests (ACARA, 2014a) but were replaced by an online simulation for the 2015 test (ACARA, 2017). The PISA, TIMSS or EV tests have no included performance tasks (the EV test has a simulated investigation as one of the extended response tasks). Some (e.g. Fensham & Rennie, 2013) would argue that this reduces the validity of test scores relating to science.

Usability (in Table 2.2) is another issue Matters and Curtis (2008) raise:

> The usability of assessment and reporting methods involves the capacity of the assessment and reporting system to be informative to stakeholders in meeting their diverse needs… An approach will be regarded as practicable if it works and imposes a justifiable yet limited load upon participants and yields valuable information to stakeholders. (p. 16)

The researchers discuss the notion that good assessment provides both summative and formative feedback and is credible. *Credibility* (see Table 2.2) inheres in the soundness of the assessment regime and the reputation of the issuing authority, which (for the purposes of employability skill credentials) may be the schools or the established state and territory curriculum and assessment bodies.

Education authorities publicise results from international tests and school-level aggregations of national test results. In Australia, results from international tests (TIMSS and PISA), NAP assessments and NSW Year 12 external school exit examinations are published for all to see. Some media outlets use the results to publish ordered lists of schools using whatever criteria the reporters believe supports the point they want to make about assessment results. Private coaching colleges also use the results in their advertisements to attract clients.

Poor results can encourage teachers to teach to the test (Au, 2007) as a mistaken response to social pressure for good results. The receipt by students of consistently poor assessments can discourage participation and engagement in learning who already have poor learning histories (ARG, 2002b). Testing or assessment that is consequential for stakeholders has been labelled 'high stakes' in the literature (e.g. Gipps (1999); Harlen & Deakin-Crick (2002); NRC (2001); Polesel, Dulfer, & Turnbull (2012)).

Messick (1995) insists that the impact of assessment results on individuals must be taken into account when interpreting assessment scores. On that basis, it is entirely appropriate for ACARA to explain the limitations of the information it provides about schools on the *MySchool* website that it knows people access to compare schools. TIMSS and PISA testing involves the collecting of contextual information to assist with interpreting the test scores PISA officials publish for each country (Thomson, Wernert, et al., 2017; Thomson, De Bortoli, et al., 2017). The NRC (2001) report identified four sets of concerns about the adequacy of assessments that were evident at that time:

1. the validity of evidence used to produce results
2. the reliability of inferences about the level of competence and overall proficiency demonstrated

3. the publishers' silence about interventions likely to improve achievement or performance
4. issues with equity and fairness. (see pp. 26 – 29)

Growing recognition of these concerns has prompted education authorities to better align large scale assessment with curriculum standards and to develop assessments for knowledge and skills not well addressed by existing test items and tasks that make up most standardised tests currently in widespread use (such as aptitude tests used to moderate individual school test results in Australia and in the US). Performance assessment has been another response. Students are presented with "open-ended tasks that call upon [them] to apply their knowledge and skills to create a product or solve a problem" (NRC, 2001, p. 30).

Harlen (2005), in work for the UK-based Assessment Reform Group (ARG), explored the issues teachers have to reconcile when attempting to use classroom assessments and results from tests, including large-scale external tests for both formative and summative purposes. Broadly speaking the trade off that has to be made is between validity and reliability, which was discussed above in the context of *Dependability*. The discussion in her paper is applicable to the NSW context where teachers are being asked to use assessments for both summative and formative purposes (the EV program).

A major report on high stakes testing in Australia was published in two parts by the Whitlam Institute in 2012. The literature review part (Polesel et al., 2012) considered "whether the tests themselves are reliable, valid and desirable on their own terms as a means of assessment" (p. 8) and cited research challenging the tests as a basis for educational decision making under the headings of reliability, student health and well-being, learning, teaching and curriculum.

The report itself (Dulfer et al., 2012) drew on the literature review and responses (N= 8353) to the very large online national survey to provide an "educators perspective" (in the report title) and concluded:

> NAPLAN is viewed by the teaching profession as 'high stakes testing'; findings…suggest that NAPLAN may be having a detrimental effect in areas such as curriculum breadth, pedagogy, staff morale, schools' capacity to attract and retain students and student well-being; and, concerns expressed…suggest that further research is required to examine carefully the uses, effects and impacts of NAPLAN (p. 9)

Whilst the tests reviewed in the Whitlam Institute-sponsored research are about literacy and numeracy testing, the technical issues related to validity, reliability, desirability and fairness apply to any one-off summative test, such as the national Year 6 science test, the EV tests, the now abandoned Year 10 tests in NSW and current Year 12 school exit tests in Australia and other parts of the world, as well as tests devised by teachers for students at their schools.

2.5 Measurement and summative and evaluative assessment

This section will describe summative assessment models – "the generation of summative data" (Broadfoot, 2009, p. x) – that epitomise the rigorous approach to measurement that underpins the TIMSS, PISA, NAP-SL and NAPLAN tests. It will include examples from the *MySchool* website.

Discussion of the above tests is relevant to this thesis for three reasons. The first is to get a sense of what is being measured. The second is to obtain a sense of whether test results can be used for formative purposes and, if yes, at what level/s (individual / class / school / school system (government or private) /state or territory / national / international might the information they provide be useful? The third reason is to understand what information about schools is available on the *MySchool* website and to explain how it was used in this research project.

In the context of a school, teachers' summative assessment (of individual's achievements) usually happens at the end of an episode of teaching. The phrase summative assessment refers to

> [the] process by which teachers gather evidence in a planned and systematic way in order to draw inferences about their students' learning, based on their professional judgement, and to report at a particular time on their students' achievements. (Harlen, 2005, p. 247)

At this point, it is perhaps worth recalling how evidence is gathered and used to inform reports to parents.

> Humans can only provide evidence in the form of what they *write, make, do* and *say* and it is from these four observable actions that all learning is inferred. This is the basic and fundamental role of *assessment*—to help interpret observations and infer learning. The more skills are observed, the more accurately generalised learning can be inferred. Hence, there is a need to document the discrete observable skills and find a way to blend them into cohesive evidence sets (Griffin, 2009, p. 195)

Griffin (2009) could equally have completed the above quote with the following addendum: "and interpret them in a conventional way to report progress in learning". In NSW, reporting conventions for students from Years K to 10 are described on the Board's website (NESA, 2017, *Awarding grades*). Information about assessment and related reporting procedures in the senior years for NSW and other Australian states and territories is available on the Australasian Curriculum, Assessment and Certification Authorities website (ACACA, 2018).

The assignment of a grade for reporting purposes is based on a teacher's judgment of the accumulated evidence of learning gathered since the last report. NSW government schools are required to formally report to parents twice a year. What is to be learned and assessed are syllabus outcomes and related content that defines the minimum expectations for achieving the outcome/s. This learning matters because it has been deemed appropriate by those empowered to create the syllabus for students at the age and stage of learning for which the proposed summative assessment is to be done. The expectation is that teachers will have provided students with access to the content that will be the target of the assessment.

As will be shown in Chapter Five, evidence of learning (in science) in the 16 case study schools was typically collected using pen-and-paper tests, responses to practical activities and research projects for which written reports or answers to specific questions and/or oral presentations are required. Typically, tasks were assessed in the course of the school year and a mark awarded to each based on criteria derived from the outcome/s targeted and its/their related content. The marks are recorded and then used as the basis for making an 'on-balance', holistic judgment that is then represented as a grade from A to E. It is the grade that is reported to students, parents and interested others at predetermined times in the year over the successive years of schooling.

In making this judgment, teachers are assisted by the Department (official policy and support material on the Department's intranet for government school teachers) and the NSW Board of Studies public website dedicated to assessment support (BOS, 2013). The Board's website includes the *Common Grade Scale* (CGS) and related advice about how to make a grade judgment. For a particular stage (e.g. Stage 4 covering school Years 7 and 8) an A grade would be awarded to a student who

> has an extensive knowledge and understanding of the content and can readily apply this knowledge. In addition, the student has achieved a very high level of competence in the processes and skills and can apply these skills to new situations. (BOS, 2013, *The Common Grade Scale*)

By comparison, an E grade would be awarded for work judged to demonstrate

> an elementary knowledge and understanding in few areas of the content and has achieved very limited competence in some of the processes and skills. (BOS, 2013, *The Common Grade Scale*)

The scope for judgment about the appropriate grade to apply is constrained to student demonstrations of knowledge and understanding; ability to apply that knowledge and understanding in new situations; and the level of skills and processes related to science. Depth is a relative term ranging in the case of skills and processes from very high to very limited. The capacity to apply those skills in new situations goes from an implied “almost all” to “most” for a B grade, then not mentioned after that. Thus, if there is no evidence of transfer, the best a student can achieve is a C grade. Judgments about syllabus-described Values and Attitudes (BOS, 2003, p. 11) were not to be included in these assessments.

The assignment of the grade is based on a holistic on-balance judgment applying to all of the outcomes assessed or for different “areas” of grouped outcomes. For Stage 5 science, the Board recommends reporting achievement for six areas: Knowing and understanding; Questioning and predicting; Planning and conducting investigations; Processing and analysing data and information; Problem-solving and Communicating (BOS, n.d.).

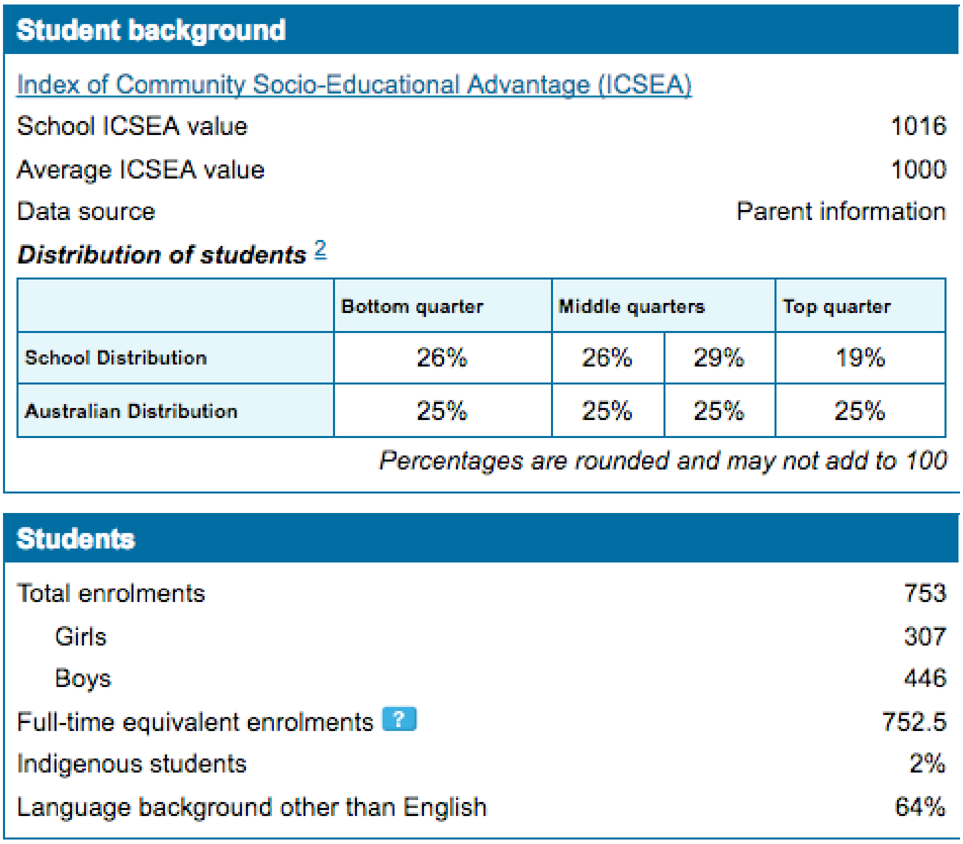
Given the methodology for collecting and scoring evidence of learning relative to syllabus standards, reporting in grades appears to be an appropriate trade-off prioritising construct validity over reliability. The award of a grade involves differentiating between five levels as compared to the dubious reliability of an implied differentiation if results were reported as percentiles.

The next example relates to the ways results from large scale national testing in literacy and numeracy of every eligible student in Years 3, 5, 7 and 9 in Australian schools are reported. The reference to literacy and numeracy testing and the *MySchool* website is included in this section of the thesis because both NAPLAN data and other publicly available data on the *MySchool* website pertaining to the case study schools involved was accessed for data relevant to addressing the research questions at the heart of this thesis. Those uses will be explained in subsequent chapters.

Standardised NAPLAN results for individuals are collated into school sets and used to generate reports for parents and schools. Aggregated school level data is published on the *MySchool* website in the form of a level related to a scale that has been established to cover the range of expected performances for the great majority of students sitting the tests up to Year 9. The scale includes 10 performance Bands. Year 3 students’ performance is reported against Bands 1 to 6;

Year 9 students' results are reported against Bands 5 to 10 (ACARA, 2013c), *Results and reports*). The school websites include a range of other information that is updated annually for the 9450 schools (ABS, 2018) across Australia. Information about each school is published on the school website (ACARA, 2016a, *About*).

An extract of the school data for a government, metropolitan, comprehensive school with some unclassified students (educationally disadvantaged) and a range of students from Years 7 to 12 is shown in Figure 2.2.



Student background

Index of Community Socio-Educational Advantage (ICSEA)

School ICSEA value	1016
Average ICSEA value	1000
Data source	Parent information

Distribution of students [2]

	Bottom quarter	Middle quarters		Top quarter
School Distribution	26%	26%	29%	19%
Australian Distribution	25%	25%	25%	25%

Percentages are rounded and may not add to 100

Students

Total enrolments	753
Girls	307
Boys	446
Full-time equivalent enrolments ?	752.5
Indigenous students	2%
Language background other than English	64%

Figure 2.2 Selected school data for a government, metropolitan, Years 7-12 school. Source: MySchool website (ACARA, 2017)

In addition to using NAPLAN data, school socioeducational advantage (SEA) profile (which is referenced to the national quartile profile) was used to find comparable schools, as will be explained in Chapters Three and Five.

ACARA produces an annual report titled the *National Report on Schooling in Australia,* which is available from the ACARA website (ACARA, 2016c, *Reporting*). The Report's main purpose is to report progress toward achieving the two "Educational Goals for Young Australians" (see Appendix B). ACARA does this by "collecting, managing, analysing, evaluating and reporting statistical and related information about educational outcomes from domains of learning deemed important by the national Education Council" (ACARA, 2016d, *National data collection and reporting*). The scope of this work currently includes literacy, numeracy, science literacy, ICT, and civics and citizenship. Science literacy of Year 6 students, for example, has been monitored triennially since 2003; the latest test cycle was completed in 2015.

The third example of summative testing to be discussed involves international comparative testing and the ways results from those tests are used, by whom and for what purposes. The two tests of interest here are the TIMSS and PISA tests described above. The tests provide summative assessments of performances by student cohorts in schools chosen by a stratified, random sampling methodology to deliver a sample of students for testing. The sample has to be representative of all targeted state and territory student populations in Australia as well as their school sectors and important demographic groups related to assessing equity and excellence (national Goal 1).

The tests are of literacy, numeracy and scientific literacy, but what is assessed within the domain constructs has to be accessible and comparable in cognitive demand for all participants across the jurisdictions taking part in the testing. Evidence of learning is collected by pen-and-paper tests and other, related, contextual information from surveys completed by students, teachers, principals and education authority officers.

Detailed reports on Australian students' performance in the international tests and on their considerable influence are available on the ACER website. The intended audience for the results from these international tests are high-level education policy officers, education advisers to government and the media, and education researchers. Data sets from the tests are available for download and independent analysis.

ACER has published a book for teachers about PISA (Thomson, Hillman, & De Bortoli, 2013) that explains the test and its purposes as well as providing some examples of assessment tasks. Both Fensham (2013) and Millar (2013) argue that this is a very worthwhile initiative because it provides good models for assessment items that teachers should use and replicate in the context of their own school-based assessment.

Scientific literacy was the domain of major focus for assessment in 2015 (as it was in the 2006 round of testing). In 2006 the constructs *interest in science* and *support for science* were included for assessment in the test. A third construct, *responsibility towards resources and environments*, was included in the student questionnaire. At that time, the inclusion of attitudes toward science in this sort of test was ground breaking. It was retained in 2015 but were addressed in the student questionnaire.

Of particular interest to this thesis is the PISA 2015 assessment framework, which included the new feature of cognitive demand "within the assessment of scientific literacy and across all three competencies of the framework" (OECD, 2017, p. 40). The test developers distinguish cognitive difficulty from empirical item difficulty. The latter is "estimated from the proportion of test-takers who solve the item correctly" (p. 40).

Cognitive difficulty relates to the type and level of mental processes demonstrated in responses to a question. Of relevance to this thesis is the international acknowledgment that the level of thinking demonstrated by a student is an important aspect of the competencies that define scientific literacy. From its inception in 2005, the EV program has included the measurement of cognitive difficulty (levels of thinking). The inclusion of cognitive difficulty in the PISA 2015 assessment framework is belated vindication of its incorporation into the EV program.

Before Webb's (1997) Depth of Knowledge (DOK) approach was chosen as the best to measure cognitive difficulty for the purposes of PISA 2015, a number of other theoretical frameworks were considered, including the SOLO Taxonomy (Biggs & Collis, 1982). In the view of PISA developers, DOK "is a simpler but more operational version of the SOLO Taxonomy" (OECD, 2017, p. 40). The EV program in NSW uses the SOLO model (Panizzon et al., 2006) as the basis for measuring levels of thinking. The reasons for using SOLO for this will be discussed in Section 2.6.

Given the high stakes involved here for the countries participating in these international tests, the assessment frameworks are subjected to scrutiny and need to be defensible. State of the art psychometrics are utilised to ensure dependability of scores (the appropriate trade-off between construct validity and reliability). For these international tests, given the diversity of curricula across the countries involved and the absence of any international agreement about what to test, "reliability [is] the dominant statistic in these international tests" (Fensham, 2013, p. 14).

Fensham's (2013) main criticism of TIMSS and PISA relate to the fact that pen-and-paper testing cannot assess the increasingly important expectations for science and technology learning, such as

> practical performance in science…decision making about socio-scientific issues, context-based science and science project work in and outside school…Neither TIMSS nor PISA acknowledges the absence of any testing of the science learnings associated with these newer goals. Such high-status silence can easily be interpreted as suggesting they are not of worth. (p. 18)

Despite acknowledging validity issues, over time TIMSS and PISA tests have provided important reliable (in the statistical sense) feedback to education authorities around the world on issues to do with gender equity in terms of science achievement, the impact of socio-economic background on achievement, and whether the gap between top and bottom performers is getting wider or narrower. In Australia, the sample size is deliberately large enough to provide reliable data on achievement of students in the different states and territories and school sectors (government school, catholic school and other independent school) as well. The Australian data shows girls and boys do equally well in science; the socio-economic status of parents is positively linked to achievement; and students of Indigenous background and students who are educated in geographically isolated places do much worse in science than their metropolitan counterparts. In short, these tests provide a picture over time of student progress in relation to the "Educational Goals for Young Australians."

2.6 Formative assessment and formative practices

The NSW science syllabus of relevance to this project (BOS, 2003) refers to assessment *of* and *for* learning (pp. 70-75). The current syllabus (BOSTES, 2012) talks about assessment *as* learning, as well as being *of* and *for* learning (p. 171). In the literature, assessment *as* learning (Dann, 2002; Earl & Giles, 2011; Hickey, Taasoobshirazi, & Cross, 2012) is linked to "assessment for learning" (Black & Wiliam, 2009, p. 8) when the researchers talk about activating students as the owners of their own learning. Advocates of assessment *as* learning accept that students should be valued participants in their own learning, anticipate receiving and utilising constructive feedback and feed-forward and be able to identify their own learning gaps and solve their learning needs, with teacher assistance. Through this practice students can develop skills for life-long learning and be self-motivated by learning self and peer assessment strategies. (Earl & Giles, 2011, p. 13)

Some researchers have warned that a simplistic view of assessment *as* learning could be misconstrued as endorsing teaching to the test, coaching to improve test answering skills, and the notion that testing counts as learning (e.g. Sadler, 2007; Torrance, 2007). Assessment as learning is also linked to self-regulated learning

(Boekaerts & Corno, 2005; Clark, 2012; Nicol & Macfarlane-Dick, 2006; Schraw, Crippen, & Hartley, 2006).

This thesis will attend primarily to formative assessment, self-regulated learning, learning how to learn, and learning independence or autonomy. because the research literature for these is more extensive. In Chapter Three this literature has been used to develop the dimensions of formative practice, which constitute the theoretical framework used for exploring the impact of assessment for learning and the EV program on assessment-related work of science teachers. This will be explained in subsequent sections and in Chapter 3.

According to Black, McCormick, James, and Pedder (2006), self-regulated learning is the key to "learning how to learn", which these researchers distinguish from learning to learn. Self-regulated learning underpins the capacity for "life-long learning" (Black et al., 2006, pp. 120-121). The importance to individuals of acquiring the capacity for independent, life-long learning has been identified as an important goal for preparing students for life in the knowledge society and its related knowledge economy. It was the over-riding goal for science education in NSW in the period of interest for this project (BOS, 2003).

Assessment for learning or formative assessment is attracting a lot of attention because it is perceived to be perhaps the single most important key to improving engagement with learning and related achievement in science. If it is properly implemented, students should be graduating from school well on their way to being self-managing learners. There are three reasons for making these strong claims.

The first reason is the wave of support for formative assessment and its pedagogical offspring, formative practices, sparked by two publications by Black and Wiliam in 1998: *Assessment and Classroom Learning* (Black & Wiliam, 1998a) and *Inside the Black Box: Raising Standards Through Classroom Assessment*. (Black & Wiliam, 1998b). The latter was written with science and other teachers in mind.

The second reason is the strong and growing confirmation that the teacher is "…the greatest source of variance that can make the difference [in achievement]" (Hattie, 2003b, p. 3). In his calculations of effect sizes for a large array of classroom interventions, Hattie identifies 14 influences of teachers, all but three of which are linked to what the teacher does in the classroom with students.

The third reason is that an analysis of what teachers do in the classroom that makes the most difference to achievement, are all linked to "formative practices" (so called by Black & Wiliam, 2009, p. 8). Each of these three reasons will be dealt with in separate subsections.

2.6.1 Support for formative assessment

The ARG with sponsorship from the Nuffield Foundation had commissioned Black and Wiliam in 1995 to review the literature on formative assessment. Their report was published in 1998 (Black & Wiliam, 1998a). Subsequently, the ARG published a brochure describing ten principles of *Assessment for Learning* and gave strong endorsement for two publications about assessment for learning arising from that review and later work (ARG, 2002a). The ARG defined assessment for learning as

> The process of seeking and interpreting evidence for use by learners and their teachers to decide where learners are in their learning, where they need to go and how best to get there. (ARG, 2002a, p. 2)

In a second publication, *Working Inside the Black Box,* Black et al. (2004) reprised the three questions the Black and William (1998b) review had set out to answer:

1. Is there evidence that improving formative assessment raises standards?
2. Is there evidence that there is room for improvement [in formative assessment practices]?
3. Is there evidence about how to improve formative assessment?

The research reported in 1998 had said yes to the first two questions. Black et al. (2004) provided an answer in the affirmative for question three. It reported the results of a two-year, school-based project involving the researchers working with science, mathematics and, later, English teachers to improve formative assessment practices and to develop new ones.

Assessment for learning was acknowledged in the US publication (NRC, 2001) as "*assessment to assist learning,* or *formative assessment."* (p. 38, italics in the original). The NRC (2001) report referenced the 1998 Black & Wiliam paper:

> [Black & Wiliam] also report…that the characteristics of high-quality formative assessment are not well understood by teachers and that formative assessment is weak in practice. (p. 227)

The NRC (2001) report appears to acknowledge this was an issue in the US as well because its Recommendation 11 said:

> The balance of mandates and resources should be shifted from an emphasis on external forms of assessment to an increased emphasis on classroom formative assessment designed to assist learning. (p. 14)

The OECD publication on formative assessment titled *Formative Assessment: Improving learning in secondary classrooms* (OECD, 2005) cited a journal version of the *Working inside the Black Box* narrative (Black & Wiliam, 2005). The OECD used the journal version as the main referent from the English-speaking world and linked it to eight case studies of formative assessment in practice from around the world, including Queensland, as indicated earlier in this chapter. Of formative assessment, the OECD report says:

Studies show that formative assessment is one of the most effective strategies for promoting high student performance. It is also important for improving the equity of student outcomes and developing students' "learning to learn" skills. (CERI, 2005, p. 13)

In Australia, the writers of *The Status and Quality of Teaching and Learning of Science* (Goodrum et. al., 2001) endorsed Black & William's (1998a) support for the provision of meaningful feedback to achieve improvements in learning outcomes.

Assessment for learning, as distinct from assessment of learning, implies an important shift in the ownership of assessment. The overwhelming message about assessment of learning is that it is done to someone (students?) by someone else (a teacher?) and the person 'done to' wears the judgment label assigned them (Newton, 2007). The language I am using is deliberately pejorative to signal that a proper understanding of assessment involves recognising it as a social act (Gipps, 1999; Broadfoot & Black, 2004)

In science learning, the teacher's role is to help students identify and own a progression in science learning appropriate to their needs as students in a science course. In the process of doing that, the teacher should provide students with the cognitive tools to construct their own learning maps, which they can use to navigate through life as a science student at school and in life generally. The goal for teachers is to make themselves redundant (Sadler, 1998).

Bell and Cowie (1997) wrote a report for the *Learning in Science Project (Assessment)* which ran in 1995 and 1996. Republished in 2002 (Bell & Cowie, 2002), the report suggested:

1. Pen-and-paper tests cannot provide data for many of the valued outcomes in science (such as inquiry tasks or working in teams).
2. There are many purposes for assessment (cf. Newton, 2007).
3. If learning is owned by the student, the teacher needs to be able to monitor student conceptual development and support the process by having a theory of learning that can be used to support that progress.
4. Formative assessment can provide evidence of learning for the gaps in assessment coverage (thus improving the dependability of the assessment) for a diversity of purposes and uses, and better quality feedback to support the progressing conceptual development from naïve to sophisticated understandings of science.

Cowie and Bell (1999) defined formative assessment as "the process used by teachers and students to recognise and respond to student learning in order to enhance that learning, during the learning" (p. 101). Wiliam (2011b) acknowledged the Cowie and Bell (1999) qualification "during the learning".

Wiliam (2011b) credits Stiggins with popularising the use of the phrase 'assessment for learning'. He also attributes to Stiggins the identification of four conditions that have to be satisfied for formative intent to be realised and for students to remain engaged with the learning process even when the assessment result is not what they would want to receive (Stiggins & Chappius, 2005). Wiliam (2011a) also elaborates the principles of formative assessment, going well beyond what was provided in the paper by Black & Wiliam (2009).

Wilson and Sloane (2000) described a system of embedded assessments—the so-called BEAR (Berkeley Evaluation and Assessment Research) Assessment System, or BAS. The BAS "is a comprehensive, integrated system for assessing, interpreting, and monitoring student performance" (p. 182). Its tools enable teachers to:

- assess student performance on central concepts and skills in the curriculum
- set standards of student performance
- track student progress over the year on the central concepts
- provide feedback (to themselves, students, administrators, parents, or other audiences) on student progress and on the effectiveness of the instructional materials and the classroom instruction. (p. 182)

The principles behind the design of this classroom-based assessment system are:

1. It should be based on a developmental perspective of student learning (ie that there is a definable pathway a student follows as they work through the topic […] a learning progression that describes intended learnings in a curriculum defined learning area, such as science.
2. There must be a match between what is taught and what is assessed (which means that other methods for assessing performance apart from responses to pen and paper tests must be used).
3. Teachers must be the managers of the system (if they are to use the results as effective feedback).

4. To be acceptable beyond the school, assessments have to be seen as fair, valid and reliable measures of the expected learning (evidence of high quality).

2.6.2 Teachers make the difference

Hattie (2003b) has summarised the findings from many studies, using Hierarchical Linear Modelling (HLM) (p. 1), which attributes the variation in student achievement at school to six main influences as measured by the results from large scale external testing. The last four are grouped as "combined effects" sometimes referred to as school environment factors. HLM also assigns the relative weight each has on achievement. The three contributors to variation are:

1. what students bring to school in the form of ability and social capital (50%);
2. the expertise of the teacher (30%)
3. the combined effects of school-principal, home and peer effects (20%).

Hattie (2003b) argues that supporting teachers would be the most productive way to improve achievement. He made that point by comparing the effect size differences of 16 influences on achievement (assessed using the SOLO Taxonomy) attributed to expert as opposed to experienced (which he defines in his paper) teachers (see Figure 2.3).

Given that effect sizes above 0.40 (vertical axis Figure 2.3 graph) are value adding above the average, teacher expertise is a very useful contributor to learning.

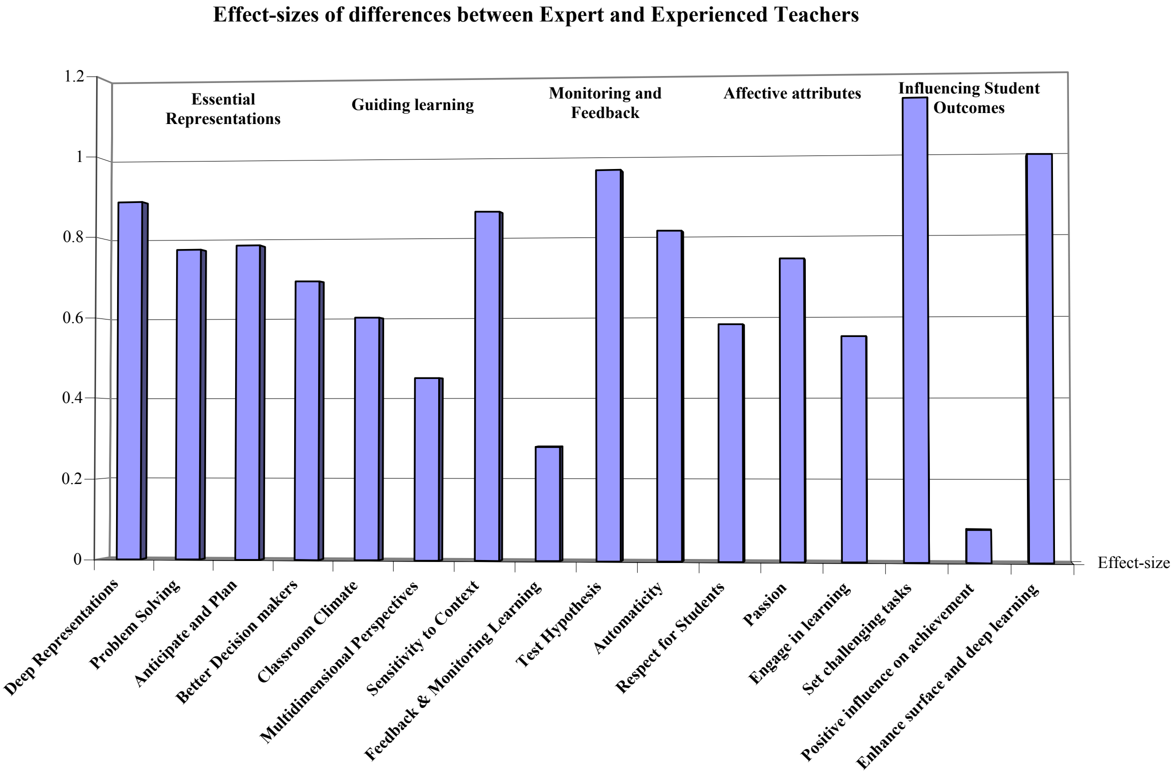


Figure 2.3 Effect-sizes of differences between Expert and Experienced Teachers. Source: Hattie, 2003b, p. 14.

When you drill down into the dimensions of expertise that provide the greatest effect size, the ability to use aspects of formative assessment feature highly. Examples given include the use of feedback, the capacity to manage classroom discussions productively, and working with students in ways that enhance their capacity for self-regulation.

2.6.3 Weight of evidence supporting formative practices

The sheer weight of evidence that emerged from meta-analyses of the huge body of research findings about interventions and strategies used by teachers to improve achievement is perhaps the most compelling reason for supporting formative practices. Meta-analysis is a statistical process that provides a comparable measure of effect size for interventions tried and tested in research projects where before and after studies produced a result. John Hattie's (2009) *Visible Learning,* *Tomorrow's Schools, The Mindsets that make the difference in Education* is extraordinary for two reasons.

The first reason is the huge number of research papers he analysed to produce the effect sizes for different interventions. The snapshot of the research projects included for publication numbered more than 800 meta-analyses of some 50,000 studies involving more than 200 million students.

The second was its revelation of consistently high effect sizes attributable to interventions associated with formative practices (this will be explained later in this section). The average effect size (ES) of all interventions Hattie reviewed was 0.40. Based on evidence from large-scale testing in the US, the UK, New Zealand and Australia, Hattie (2012) says this is the average ES on achievement of one years teaching. To have an impact on achievement above that, teaching needs to involve practices with an ES > .40.

Each of the influences is discussed and explained in the body of the text referenced in Table 2.4, which lists the 21 most powerful influences of student achievement as of 2012.

Table 2.4

Influences on learning and effect sizes

Influence	ES	Influence	ES	Influence	ES
Self-reported grades / Student expectations (STE)	1.44	Comprehensive interventions for learning disabled students (TGE)	0.77	Acceleration (SLE)	0.68
Piagetian programs (STE)	1.28	Teacher clarity (TRE)	0.75	Classroom behavioural (SLE)	0.68
Response to intervention (STE)	1.07	Feedback (TRE)	0.75	Vocabulary programs (CME)	0.67
Teacher credibility (in the eyes of the student) (TRE)	0.90	Reciprocal teaching (TGE)	0.74	Repeated reading programs (CME)	0.67
Providing formative evaluation (TGE)	0.90	Teacher-student relationships (TRE)	0.72	Creativity programs on achievement (TGE)	0.65
Micro-teaching (TRE)	0.88	Spaced vs mass practice (TGE)	0.71	Prior achievement (STE)	0.65
Classroom discussion (TGE)	0.82	Metacognitive strategies (TGE)	0..69	Self-verbalization and self-questioning (STE)	0.64

Source: Hattie, 2012, p. 266 / ES = Effect Size / STE = student effect / TRE = teacher effect / TGE = teaching effect / SLE = school effects / CME = curriculum effect

Eleven of the influences in Table 2.4 with the highest ES are linked to teacher use of formative practices (e.g. providing formative evaluation, classroom discussion, feedback, reciprocal teaching, and metacognitive strategies). Local variations of the curriculum effect (CME) influences were observed in the programs in some of the case study schools visited for this project.

2.6.4 Formative Practice

A consistent theme in Black and Wiliam's work is their interest in establishing a theory of formative assessment "to provide a unifying basis for the diverse practices that are said to be formative" (Black and Wiliam, 2009, p. 7). Their first proposition is that both the teacher and student are responsible for the outcomes from three key processes in teaching and learning:

- establishing where the learners are in their learning

- establishing where they are going
- establishing what needs to be done to get them there.

Black and Wiliam bring the three processes and the roles of the agents (teachers, peers and the students themselves) together on a grid to generate five key strategies for conceptualising formative assessment:

- clarifying and sharing learning intentions and criteria for success
- engineering effective classroom discussions and other learning tasks that elicit evidence of student understanding
- providing feedback that moves learners forward
- activating students as instructional resources for one another and their teacher
- activating students as the owners of their own learning. (Black & Wiliam, 2009, p. 8)

The researchers also provide an updated definition of formative assessment that conflates it with instruction:

> Practice in a classroom is formative to the extent that evidence about student achievement is elicited, interpreted, and used by teachers, learners, or their peers, to make decisions about the next steps in instruction that are likely to be better, or better founded, than the decisions they would have taken in the absence of the evidence that was elicited. (Black & Wiliam, 2009, p. 9)

The researchers explain that instruction means teaching and learning activities and, because the effect of "decisions about the next steps in instruction" is not certain, the qualification of "likely to be better or better founded" is an appropriate qualification for those decisions and related actions to improve learning.

For the purposes of this study, I have linked together in the following equation each of Black and Wiliam's (2009) five strategies of formative assessment and science teaching and call the combination, dimensions of formative practice.

Five Dimensions of Formative practice	=	**Formative assessment activity**	+	**Instruction in science**

The first dimension of formative practice involves activities that focus on clarifying and sharing learning intentions and success criteria related to learning science (LISC).

The second dimension involves classroom discourse in science contexts that elicits evidence of learning (CDEL).

The third dimension focuses on feedback used (by either or both the teacher or student) to progress the learning of science (FTAL).

The fourth dimension is about activating students as instructional resources for each other and the teacher in support of science learning and including peer assessment (ASIR).

The fifth dimension involves activating students (and teachers) as owners of their own learning in science and including self-assessment (ASTL).

In the methodology section of this thesis (Chapter Three), the activities related to each dimension are further differentiated into teacher or student focus/emphasis/agency. The reason for this differentiation is to provide an operational definition for self-regulated learning constructed in terms of the extent of student agency with the five dimensions of formative practice.

2.6.5 Formative practice and self-regulated learning

Black & Wiliam (2009) assign agency for assessment to teachers, peers, and individual learners, saying, "Formative assessment is concerned with the creation of, and capitalization upon, 'moments of contingency' in instruction for the purpose of regulation of learning processes" (p. 10). A narrow focus distinguished the formative assessment component in instruction from acts that follow, acts drawing from the teachers' knowledge of "instructional design, curriculum, pedagogy and epistemology." (p. 10). These 'moments of contingency' may be synchronous (immediately acted upon) or asynchronous (delayed action). To be effective, formative interactions have to result in learning. Black and Wiliam (2009) cite Sadler's definition of learning as "the activity of closing the gap between a learner's present state of mind and the state implied by the learning aim" (p. 12).

The feedback by a teacher may not do its intended job (change cognition) unless the teacher has some insight or understanding of how "students approach problem solving, and how they argue, evaluate, create, analyse and synthesise" (Sadler, 1998, p. 81). This refers to what teachers understand about the processes of metacognition that they and the student bring to the process of mediation occurring in the middle section of Figure 2.4.

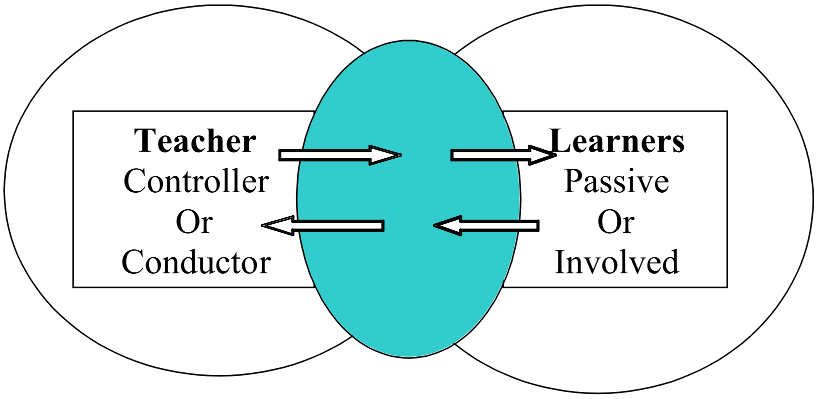


Figure 2.4 The three interacting domains of pedagogy (or instruction)
Source: Black & William, 2009, p. 11

To explain their theoretical models of mediation, Black and Wiliam (2009) begin by providing the definition of self-regulated learning used by Boekaerts, Maes, and Karoly (2005) who had completed a general review of this field:

> Self-regulation can be defined as a multi-component, multi-level, iterative self-steering process that targets one's own cognitions, affects and action, as well as features of the environment for modulation in the service of one's goals. (p. 150)

Boekaerts and Corno (2005) describe a

> dual processing self-regulation model where learning goals interact with well-being goals […] when students have access to well-refined volitional strategies manifested as good work habits, they are more likely to invest effort in learning and get off the well-being track when a stressor blocks learning. (p. 1)

Two possibilities operate in this context. One is described as a *top-down SR* or *growth* option pathway which

> has a focus on learning […] the student pursues the purpose of achieving learning goals that increase resources, i.e. knowledge and both cognitive and social skills. The process is motivated and steered by personal interest, values and expected satisfaction and rewards. (p. 14)

The other pathway is described as the well-being option, which may manifest itself as the learner choosing

> competitive performance goals or prioritis[ing] friendship with peers, which a focus on learning goals may put at risk. [It] may be triggered by […] some types of classroom feedback and reward, or merely by boredom. When cues from the environment have this effect, this second option is adopted—that of giving priority to well-being. (p. 14)

In the course of a learning episode, students may seek one or other of these options and choose at times to switch from one to the other. The choice of option is also influenced by the students "awareness of and access to volitional strategies (metacognitive knowledge to interpret strategy failure and knowledge of how to buckle down to work) helplessness, and failure of emotional control" (Vermeer, Boekaerts, & Seegers, cited in Black & Wiliam, 2009, p. 14).

2.6.6 Learning how to learn, self-regulated learning and life-long learning

In *Learning How to Learn and Assessment for Learning: a theoretical inquiry*, Black et al. (2006) write:

> The overall conclusion is that emphasis should be placed on practices that have potential to promote autonomy in learning, a common theme in the literature at all levels, and one reflected in our empirical work on teachers' attitudes and practices. (p. 119)

It is important to understand that the notion of learning how to learn is also consistent with the education agenda related to employers wanting employees with skills for work (and life more broadly) in the knowledge society. (OECD, 2003; CERI, 2008).

Black et. al.'s (2006) paper is an attempt to build a bridge between what we know about teaching and learning that might put students in charge of their own learning. See also Deakin-Crick, Broadfoot, and Claxton (2004); James (2006); James et al. (2007); Mansell et al. (2009) and Pellegrino (2009).

In the context of science education in Australia, *The Project for Enhancing Effective Learning* (PEEL), founded in 1985, anticipated the work reported on above. In the Australian context

> PEEL is about making significant changes in how students learn—generating learning that is more informed, purposeful, independent, interactive, and metacognitive. (Mitchell, Mitchell, & Lumb, 2009, p. 1)

The PEEL (2009) publication *Principles of Teaching for Quality Learning* describes 12 principles that teachers use to instil good learning behaviours. Good learning behaviours are those that operationalise metacognition and self-assessment, which are powerful contributors to learning how to learn. The list of activities and the ideas developed by science teachers and published in PEEL SEEDS (PEEL, 2009) are similar in type to the list of activities in the King's-Medway-Oxfordshire-Formative Assessment Project (KMOFAP) (Black & Wiliam, 2005). The list of procedures (using a PEEL term for learning activities) includes:

- sharing success criteria with learners
- classroom questioning
- comment-only marking
- peer- and self-assessment
- formative use of summative tests.

The point of drawing attention to PEEL is that it is an existing network that holds a considerable body of work that teachers can access themselves as they try to improve student learning behaviour and autonomy.

2.7 SOLO and the ESSA-VALID (EV) program in NSW

This section will discuss SOLO as a learning progression. It will explain the thinking involved in its creation, its use in science assessment and its contribution to the EV program in NSW.

2.7.1 The SOLO Taxonomy

The SOLO Taxonomy (Biggs & Collis, 1982) and its successor SOLO model (Panizzon, Arthur, & Pegg, 2006) are examples of developmental learning progressions in the cognitive tradition (NRC, 2001). The original SOLO Taxonomy was published by Biggs and Collis (1982). It was developed to assist teachers differentiate between quantity and quality in student responses to closed, classroom test questions. In their original construct for the Taxonomy, learning progresses through five levels, each one representing a higher level of learning as explained below.

Biggs & Collis (1982) were concerned that students could score highly by simply writing down a number of relevant responses (quantity) without any weighting being given to whether the thinking on display was of a higher order (which Biggs and Collis described as quality) than simple recall of related bits of information. Biggs and Collis examined Bloom's original taxonomy (Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956), Piaget's hypothecated cognitive structures (Ginsberg & Opper, 1979) and other post-Piagetian models (such as those put forward by Marton & Säljö (1976); Schroder, Driver& Streufert (1967) and Shayer (1976) as the basis for describing quality.

In the end Biggs and Collis (1982) proposed and developed an empirical model that classified answers in levels according to the increasing structural complexity evident in the answers. After working with thousands of student written responses to test items, they defined complexity as including

> progression from concrete to abstract; using an increasing number of organizing dimensions; increasing consistency within the response; the use of organizing or relating principles with hypothetical or self-generated principles being used at the most complex end. (p. 14)

The first version of SOLO was developed and published without reference to a traditional science subject such as biology or chemistry (the closest subject was geography). In 1991 Biggs and Collis published an updated version of the Taxonomy that took into account work done on academic and everyday intelligence, and on ideas related to multiple intelligences, novice-expert research and forms of knowledge (Biggs & Collis, 1991).

Biggs and Collis then showed how this version of SOLO could be applied to categorising representations provided by students attempting to "explain phenomena with as yet inadequately developed [scientific] constructs by using alternative frameworks to those used by scientists" (p. 71). They used Beveridge's

(1985) work on evaporation, supplemented with older students' responses to identify the structural elements in answers that exemplified the different levels in the SOLO Taxonomy. Figure 2.5 represents this updated SOLO Taxonomy.

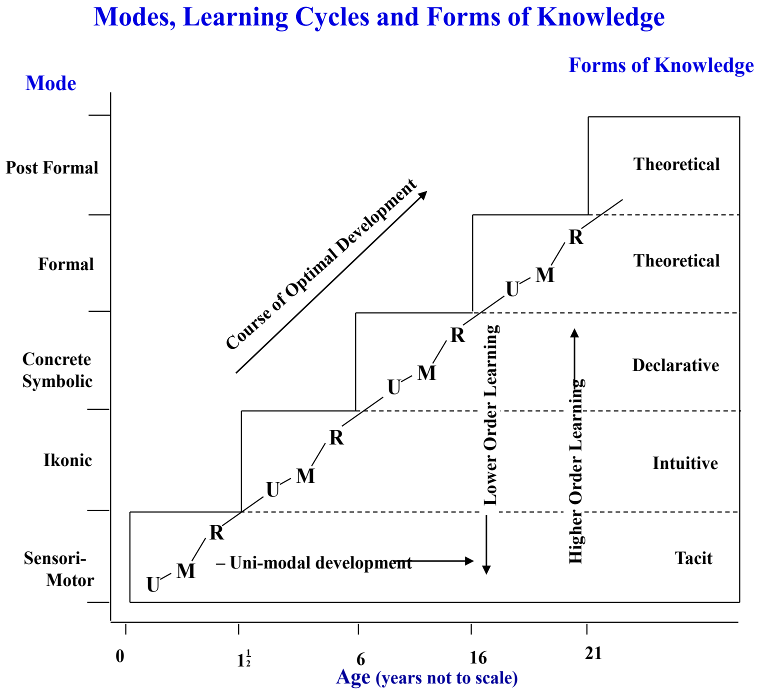


Figure 2.5 Representation of the Biggs & Collis (1991) SOLO Taxonomy (Source: Pegg, J., slide for a presentation at the ACER research conference, August 2010, in Melbourne)

Features of the SOLO Taxonomy include modes of representation/thinking (vertical axis) that are associated with age-related changes (horizontal axis) in student cognitive functioning. These changes enable students to construct different levels of response (denoted by the letters U-M-R within a mode) to questions using the knowledge forms associated with that mode of thinking (right-hand side labels within the modes).

Table 2.5 includes the examples used by Biggs and Collis (1991) to illustrate how the features of student responses change with age (column 2) and the five levels of learning descriptors student responses were mapped to.

Table 2.5

The concept of evaporation through modes of thinking and levels of thinking (SOLO Taxonomy)

Mode	Concept of evaporation	Level of learning
Postformal	U (EA) Developing and testing a new theory.	
Formal	R Working understanding of the discipline of physics M Other physical concepts involving principles of energy, matter U (EA)The heat energy supplied speeds particles so that water changes state into steam. The latent heat is the amount of energy supplied	5. *Extended abstract (EA)*-generalizes the structure to take in new and more abstract features representing a new and higher mode of operation
Concrete symbolic	R The heat turns the water into steam and it evaporates off, remaining invisible in the atmosphere (15 yrs) M The flame makes the steam come and the water goes (9yrs) U (EA) It soaks into the pan (7 yrs)	4. *Relational (R)*-integrates the parts with each other so the whole has a coherent structure and meaning 3. *Multistructural (M)*-picks up more and more relevant or correct features but does not integrate them 2. *Unistructural (U)*-the learner focuses on the relevant domain and picks up one aspect to work with
Ikonic mode	R The steam causes the water to disappear (7 yrs). This does not happen at our house. There's still water in the pan because my mum makes the tea with it (8 yrs) M You put the pan on top of the flame and the water goes U The flame does it (5 yrs)	1. *Prestructural (P)*-the task is engaged, but the learner is distracted or misled by an irrelevant aspect belonging to a previous stage or mode

The sensori-motor (mode) is not included here as it is related to motor-skills and, in this context, not knowledge of them.

Source: Adapted from Biggs & Collis, 1991, p. 65 / 66 (Their tables 5.1 and 5.2)

The five steps cover one learning cycle centred on the concrete symbolic mode (the 'target mode') which is the mode most relevant to the years of schooling. Note that the U level of one mode is the EA level for the mode below it.

The modes of thinking most relevant to schooling include the sensori-motor, ikonic, concrete symbolic and formal modes. As children age, they are able to access modes of thinking or representations that are progressively more complex and abstract. The modes in the SOLO Taxonomy do not progressively replace each other (as Piaget theorised) but are cumulative as explained below.

When attempting to learn a new skill set, such as Tai Chi for example, an already accomplished basketball player must begin at the sensori-motor level by learning (imitating and practicing) the basic foot moves or hand and arm moves of Tai Chi. Demonstrating either one is a unistructural response, demonstrating both separately is a multistructural response, and putting both feet and hand movements together with the correct breathing for one Tai Chi "move" is a relational demonstration. When the accomplished basketball-playing and now Tai Chi student takes their first driving lesson some ten or more years after starting school, he or she begins again at the sensori-motor mode to learn the actions involved in driving to the point of fluent relational execution needed to coordinate the many different component skills needed to pass, say, the safe overtaking part of the actual driving test.

From around age 18 months, children are able to link actions with imagined representations that they express in words an adult would interpret as "stereotypical characters and obvious plots" (Biggs & Collis, 1991, p. 63). In Tai Chi, while demonstrating a series of moves, a late primary-age student may use phrases like "horse-riding stance", "stroke the peacock's tail", "repulse the monkey", for example, as a way of representing the actions to themselves and others (these are actual examples from a Tai Chi support card used with all ages). This mode, called the ikonic mode,

> is evident in the intuitive knowledge displayed in … scientists [for example]. Kekule's realization of the structure of the organic ring compound was preceded by a hypnogogic dream of six snakes chasing each others' tails, and only later was his "truth" established to the satisfaction of the scientific community by evidence and argument. The ikonic mode is thus not merely a presymbolic mode of information processing restricted to early childhood. It continues to grow in power and complexity well beyond childhood. (Biggs & Collis, 1991, p. 63)

Most students begin their schooling with this (ikonic) mode of learning well developed at the multistructural and/or relational levels within the mode. Oral expression is dominant, but ikonic drawings and physical models representing people and things familiar to the student may also be produced.

From around age six years, students begin to show concrete-symbolic mode thinking. The knowledge, associated thinking and its representation within this mode is classified as declarative, which

> involves a significant shift in abstraction, from direct symbolization of the world through oral language, to written, second order, symbol systems that apply to the experienced world. There is logic and order between the symbols themselves, and between the symbol system and the world. The symbol systems of written language and signs give us one of the most powerful tools for acting on the environment, and they include writing itself, mathematical systems, maps, musical notation, and other symbolic devices. Mastery of these systems, and their applications to real world problems, is the major task in primary and secondary schooling according to any curriculum theory. Learning in the concrete-symbolic mode leads to declarative knowledge, demonstrated by symbolic descriptions of the experienced world. (Biggs & Collis, 1991, p. 63)

In their progress through this mode from incompetence to expertise,

> learners display a consistent sequence, or learning cycle, that is generalizable to a large variety of tasks and particularly school-based tasks. (Biggs & Collis, 1991, p. 64)

Responses observed may range from *prestructural* (not operating in the target mode, which is concrete symbolic in this situation) to *unistructural* (U) to *multistructural* (M) to *relational* (R) within the target mode or above the target mode (formal mode), where responses are classified as *extended abstract* (see column 3, Table 2.5). Biggs and Collis (1991) summarise observations and explanations provided by students related to the concept of evaporation to exemplify the five levels of thinking in the SOLO Taxonomy (see column 2, Table 2.5).

By the time students reach age 16 years (in Years 10 or 11), some are able to access the representational tools for formal thinking. The difference between a student operating at the concrete symbolic mode and formal mode as it was conceived then is illustrated in Table 2.5 (column 2) using the example of evaporation. In the concrete symbolic mode, explanations are tied to concrete situations and operational definitions (flames or sunlight for energy) for effects or changes observed.

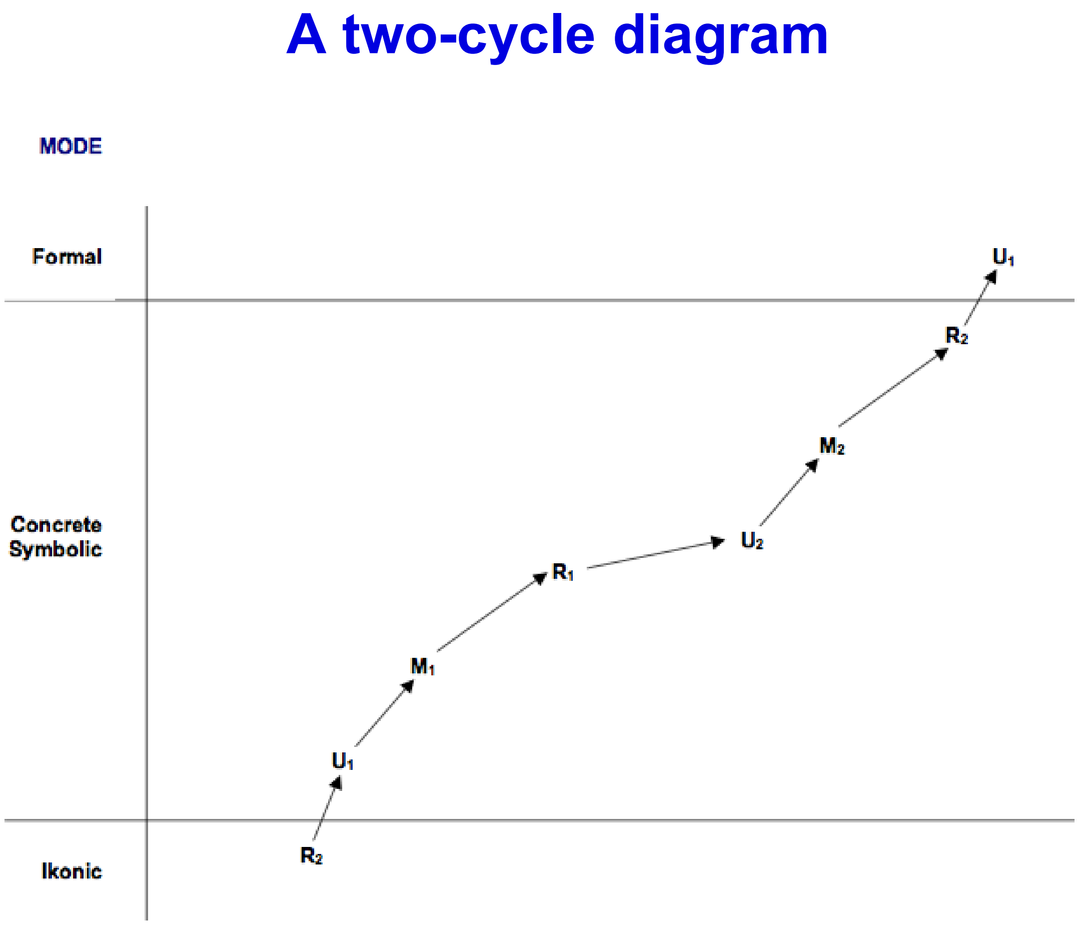
Students operating in the formal mode as it was conceived then are able to move away from particular concrete referents (ice, water or 'steam' and flames, sunlight, electricity, coal, gas) to discuss evaporation and boiling in terms of a "moving particle" model where energy is added to or taken away from a situation causing a change of state (from solid to liquid to gas and back again). "Thinking in the formal mode thus both incorporates and transcends particular circumstances" (Biggs & Collis, 1991, p. 63).

By Year 12 a number of students will be operating at the formal mode. Many may not enter the formal mode of thinking by the time they leave school at 17 or 18 years of age.

2.7.2 The SOLO model

The assessment framework developed for the EV program in NSW used an enhanced version of the 1991 version of the SOLO Taxonomy. Called the SOLO model to distinguish it from the original SOLO Taxonomy, the SOLO model includes a second learning cycle within the concrete-symbolic and formal modes of thinking. Like the Taxonomy before it, it has its roots in empirical evidence from thousands of written responses to test questions (Panizzon & Bond, 2007).

Figure 2.6 represents the two-cycle concrete symbolic mode of the SOLO model. The concrete-symbolic mode of thinking is the dominant mode of thinking throughout the years of schooling.



A two-cycle diagram

Figure 2.6. Representation of the “two cycles within a mode” SOLO model. Increasing age along the horizontal axis (L to R). Source: Pegg, J., slide from a presentation at the ACER research conference, August 2010, Melbourne.

The horizontal axis represents age in years. Most students enter school operating in the ikonic mode. Student’s capacity to use and mastery of the cognitive tools associated with the concrete symbolic mode of thinking develops over the years of schooling through two learning cycles (the second cycle is at a higher level than the preceding cycle). In the junior secondary years students acquire the language of science concepts which they are expected to use in explanations and justifications for the conclusions they come to.

At age 16 some students begin to think using abstract concepts not linked to particular situations (such as potential energy, properties of fields, latent heat, electro-magnetic radiation, mass, inertia and momentum).

The need to modify the single learning cycle approach emerged from a number of research studies where it was becoming increasingly obvious

> that a single unistructural-multistructural-relational cycle within a mode did not accommodate adequately the range of responses offered by students. In particular, it was difficult to interpret responses from many primary students, low-achieving secondary students, or adults new to a particular area of study within the single cycle model. (Pegg, Panizzon, Arthur, Scott, & Aylmer, 2011, p. 24)

It was found in their responses that

> an earlier cycle of levels (i.e., a new unistructural-multistructural-relational cycle) was discerned. Interestingly, the responses coded at these levels still shared characteristics of the same mode [... and] were particularly relevant to primary and secondary education in the concrete symbolic and formal modes (p. 24)

The two-cycle model was subsequently validated by psychometric modelling involving Rasch analyses, and the results of three studies to that end were reported in a paper by Panizzon and Bond (2007). The theory underpinning the SOLO model was originally shaped by Piaget's thinking about developmental stages. However, Panizzon and Bond (2006) refer to Vygotsky's (1978) socio-cognitive theories, and they suggest that a teacher can work with students in ways that set up the social conditions that support the emergence of a new, higher mode of thinking in students.

2.7.3 The ESSA-VALID (EV) assessment framework

While the test was delivered as a pen and paper exercise (from 2005 to 2010), the assessment framework discussed in this subsection was being developed and validated.

Table 2.6 shows an extract of the framework. It shows how the syllabus outcomes (written and published for the 2003 science syllabus) were subsequently related to the six levels of the concrete symbolic mode of thinking in the SOLO model.

Two examples of tasks are provided in Appendix D. The first is a task related to heating ice from the 2005 EV pilot test; the second is a task about magnets from the 2008 test. The tasks are mapped to the shaded outcomes and related SOLO levels as shown in Table 2.6.

Table 2.6 and related explanatory material detailing the links made between syllabus and SOLO are reproduced as example three in Appendix D.

Table 2.6

Selected outcomes and related SOLO levels in the 2011 EV assessment framework

	LEVEL 1	LEVEL 2	LEVEL 3	LEVEL 4	LEVEL 5	LEVEL 6
Outcomes 4.1 to 4.5 (2 of 7 rows)	Identify a scientific discovery	Compare scientific discovery to other types of discovery	Link a scientific discovery to its effect on humans	Describe a development in science that has led to new developments in technology	Compare the methods of the scientist to the design model of the engineer and architect	Explain the role of scientific thinking on society
	Identify a possible career path in science	Identify a science context in a career	Link a career in science to knowledge and skills required	Identify science as a human activity	Discuss why society should support scientific research	
Outcomes 4.6 to 4.9 (3 of 16 rows)	Identify materials attracted by a magnet (example two)	Compare the observable effects when magnets are placed end to end	Link the observable effects when two magnets are placed end to end with their position	Describe a magnetic field as producing a force that attracts particular metals	Describe the poles of a magnet as the area/ends where the magnet's field is most intense	Explain the behaviour of magnetic poles using the term field
		Identify that objects / substances take up space and/or have mass/weight	Explain that materials are held together differently in solids, liquids and gases			Explain density in terms of a simple particle model
	Identify an observable feature in melting, freezing, condensation, evaporation or boiling (example one)	Describe observable features in melting, freezing, condensation, evaporation and boiling	Explain that, when substances melt, freeze, condense, evaporate and boil, they are still made of the same stuff	Identify that particles are continuously moving and interacting	Compare movement and interaction of particles in different states	Explain change of state in terms of rearrangements of particles
				Identify that as particles are heated they gain energy	Identify that as particles are heated they gain energy and move further apart	Relate changes of state to the motion of particles as energy is removed or added
No content for Outcomes 10 - 12 is included						
Outcomes 4.13 to 4.15* (1 of 8 rows)	Make a simple observation	Compare observations made by different people	Explain strategies to increase accuracy of observation	Correctly sequence steps in a scientific procedure	Accurately and systematically record observations and data	Discuss the relationship between accuracy and reliability
Outcomes 4.16, 4.17 a-d & 4.18** (1 of 8 rows)	Use a simple key or symbol to represent a concrete object or representation	Distinguish between different symbols	Complete diagrams and symbolic representations	Correctly sequence steps in a process described in a text	Distinguish between two related sets of data / information	Represent relationships using keys, symbols and flow chart
Outcomes 4.17e-g, 4.19-4.21*** (1 of 7 rows)	Identify a common unit of measurement (example one)	Identify the ratio of one unit to another	Complete a correct conversion of one unit to another	Create a simple scale	Compare the scale on two axes	Create an appropriate scale

Source: NSW Department of Education and Training DET, 2011. Shaded rows are referenced in the body text. * Planning and Conducting Investigations area / ** Communication area / and *** Critical thinking area

2.7.4 The EV test: “fit for purpose”?

The literature reviewed in this chapter describes three broad purposes for assessment: to improve learning (formative assessment); to assess progress in learning (summative assessment); and to monitor aspects of, and/or the overall effectiveness and efficiency of, the education system. The EV program provides feedback on all three purposes. Feedback (in the form of results from a one off external test) is provided to students; their parents and carers; their teachers; the schools they attend; the education system authorities; and governments.

Fensham (2013) described the EV test development processes as comparable to the PISA processes, which he said were exemplary. The international tests, he said, prioritised reliability, which in this case was about ensuring that the scores included the measure of statistical certainty related to the means scores. The discussion in the later part of this subsection will describe how the EV test development processes strive for both validity and reliability in an effort to be as fit for purpose as possible.

The results from the EV test are organised into a summative report of achievement at the end of Year 8. The report for students, parents and teachers provides the results for five areas or categories of outcomes. Examples and related discussion of this aspect of the EV program are provided as example four in Appendix D. The scores from items in the EV framework mapped to the *Critical Thinking* area (see Table 2.6) are distributed to the Working Scientifically and Communicating Scientifically categories, depending on whether the items had an investigating or communicating context. The student report provides individual feedback on every task and item in the test.

The formative intent of the EV program is signalled in the report to parents and students:

> Students, parents and teachers can use the [EV] levels to plan learning programs and activities so that students keep moving forward in their science knowledge and skills. (NSW DET, 2007, p. 3)

The levels referred to are the six levels linked to the SOLO model discussed above. Progress ("moving forward" in the EV report) in science learning is defined by the language used in each of the level descriptions for a particular reporting category.

Of interest in the early days of testing was the overall concern expressed by teachers that the test was too much about reading, which in their view was getting in the way of 'seeing the science' questions. The results from the student EV survey showed that students actually enjoyed the test and stimulus material and they did not think it distracted them (see questions in the last section of the survey). Articles that teachers saw as being 'too difficult', most students enjoyed doing.

One of the intentions of assessing this way was to put a high value on getting students to read science rich texts and to identify the science content. Students strongly agreed that "literacy is important in learning science" (third question in the survey). Detailed feedback from selected case study schools on some of the survey items is provided in Chapter 6.

According to Messick (1995), "Construct validity [in principle and practice] is based on an integration of any evidence that bears on the interpretation or meaning of the test scores" (p. 742). The processes used to develop items and tasks for the EV test provide a representative coverage of syllabus intentions (mapped to the EV assessment framework), and the responses items elicit from students are evaluated by experienced teachers for alignment with intended learning as described in the syllabus.

Current psychometric methods are used to monitor the consistency with which marking rubrics are applied during the actual marking process and in reviewing the results of pilot marking. The analysis of scoring of the extended response items "utilises the Rasch Unidimensional Measurement Model (RUMM)… and the Interactive Test Analysis System (QUEST)" (Pegg et al., 2011, p. 36). Items and tasks that do not meet the criteria for inclusion in the test are discarded or modified for piloting the following year.

Teachers who have had experience teaching Year 8 students (but are not currently doing so) were invited each year to express an interest in developing items and tasks for the tests. A group comprising teachers with prior experience and some who are new is selected, and after attending a one-day training workshop they are asked to write items and tasks, for which they are paid by the Department.

The workshop takes writers through the criteria for selecting appropriate stimulus material and writing related items related that address syllabus expectations (outcomes and related essential content) for Stage 4 students. Writers are also taken through the SOLO model and shown examples of items and tasks related to the two cycles within the concrete symbolic mode that are exemplars of items and tasks used in previous tests.

The items and tasks produced are collected, assessed and either discarded or edited by officers in the EV test development unit of the Department. Surviving stimulus materials and related sets of items are edited, mapped against the EV assessment framework and collated until more than enough for one test are available. These items are then reviewed by an expert panel of teachers drawn from a range of specialist areas within the Department including Assessment, Equity, Key Competencies, Aboriginal and Torres Straight Islanders, Language Backgrounds Other than English, and Literacy and Numeracy. Examples of test items, related stimulus materials and the student survey are included at example five in Appendix D.

Several different tests using a mix of items and tasks are compiled and sent off for piloting. In the early stages, piloting was done with students in their second year of secondary schooling in the various states of Australia. Now it is done early in term one of the new school year with students who did the test in the previous year. Piloting ensures that the items and tasks with poor test characteristics (discrimination, difficulty, ambiguity, construct validity) are identified and discarded from further consideration. Marking rubrics for the three extended response questions, developed by experienced science teachers with SOLO expertise, are refined during the pilot marking process.

Experienced science teachers are contracted to score online the three extended response questions. They are provided with up to four hours of training in the SOLO model and the consistent application of the marking rubrics before actual marking commences. The marking process is continuously monitored online to ensure consistency of rubric application. Every hour, all markers of a particular question are presented with the same student response and their scores are checked to ensure consistency. The check marking is done using student responses that highlight particular scoring issues that emerged during pilot marking.

The test includes multiple items targeting the same construct. This is to improve reliability of inferences about that construct. In the end, the interpretation of how many items are needed to achieve a reliable inference is a judgment call. In addition, items from previous years tests are included to enable equating of test results across the years of testing. The equating process uses samples of items distributed across the test taken by the whole cohort so that the risk of a school seeing items it has used before in its own testing is very unlikely.

As PISA, TIMSS and NAP-SL tests are considered high stakes testing, in the interest of fairness to all, equating items are not released. Test papers are retained at the end of the test sessions and sent back to the managing agency after the tests are completed. Online delivery makes security around items easier to ensure (as for NAP-SL testing in 2015). Examples of test items not retained for equating purposes were published in the reports some one to two years after the testing was completed. Fensham (2013) has expressed a view that more of the TIMSS and PISA items should be released to provide good assessment models for schools to use.

The next subsection examines how SOLO has been used in Australia and New Zealand.

2.7.5 SOLO and assessment in Australasia

SOLO theory has been used in the design of assessment frameworks for large scale testing in Australasia since the early 2000s. It has been used in science in Australia and for reading, writing and maths in New Zealand.

The 1991 version of the SOLO Taxonomy was used by the ACER to develop the *Scientific Literacy Progress Map* (SLPM) (ACER, 2004b). The SLPM was initially developed as a tool for categorising assessment items written for the *Science Education Assessment Resource* (SEAR) (ACER, 2004a,). Items and and tasks from this project are available online to science teachers (ESA, n.d./*Improve*).

The SOLO Taxonomy was subsequently used to develop one of the strands in the assessment framework for the national Year 6 science test (ACARA, 2014a). It provided the language for the scale used to describe the change in quality of student answers found in students' answers to the items and tasks in the Year 6 test.

The SOLO Taxonomy is utilised in the assessment and reporting framework for the New Zealand-based e-asTTle project that provides assessment items for reading, writing and maths. Items are classified against the New Zealand national curriculum and the five levels in the SOLO Taxonomy (Hattie & Brown, 2004).

SOLO was considered for inclusion in the PISA2015 assessment framework as disussed above. As far as I am aware, SOLO theory is not used anywhere else in the context of large-scale testing of science in Canada, New Zealand, the UK or the USA.

2.8 Themes from the literature review and their relevance to this thesis

A need to lift and broaden the level of skills students acquire in the first phase of their education.

The demands of the knowledge society and the related knowledge-based economy require a workforce able to adapt to changing opportunities. To do this, people need to keep learning as circumstances change over their lifetime. This realisation has led to the understanding that leaving school is the end of the first phase of preparation for a life that will require further episodes of formal learning or training at least to ensure ongoing access to employment.

Employers are telling governments that they need graduates from this first phase of school, training and university who have a broader range of skills (both cognitive and social) and higher levels of skill than before. Expertise is not just about knowing, it is about being able to use that knowing in the workplace and beyond to solve problems and explain those solutions to others, and to both quantify and qualify the risks involved in implementing different options. These issues have been dealt with in a number of OECD reports including (OECD, 1997);(CERI, 2008).

Education agencies charged by governments to produce the curriculum for schools in Australia have retained a core curriculum for all students up to the end of Year 10 broadly defined in eight learning areas, including science. The science curriculum at the time of interest for this project (up to the end of 2014) consisted of knowledge and understandings drawn from the models, theories and laws, structures, systems and interactions underpinning traditional disciplines of science and the skills of "working scientifically" (BOS, 2003, p. 21) in about equal measure (10 of the 22 outcomes are skills).

In recognition of falling student engagement and interest in science starting at school, but particularly so in the early years of secondary schooling, changes were made to the curriculum. The 2003 curriculum in NSW required science teachers to provide contexts for learning about science and in which to do science. The prescribed contexts in NSW were to do with the history of science, the nature and practice of science, the applications of science and implications of doing so including current examples and work involving science.

Teachers were also required by that curriculum to use science resources to provide students with the opportunity to acquire the Key Competences, develop skills in the use of ICT, work alone and in teams safely and inclusively (considering gender and cultural differences), acquire some understanding and appreciation of Aboriginal and Torres Strait Islander People world views, acquire some understanding and appreciation of how science impacts our civic life and the environment and to improve their general literacy and numeracy skills.

To the extent possible, given the breadth of expectations, the scope and depth of what was to be learned was described in bundles of learning framed as outcomes.

Outcomes were defined by a minimum number of actions and contexts for their acceptable performance. The outcomes described a hierarchy of learning (in a set of standards for two stages in the junior secondary curriculum) that students were expected to engage with and acquire. Years 7 and 8 comprised one stage and Years 9 and 10 the second stage. Twenty-two outcomes provided the scope and depth of expected learnings in science at the end of Year 8 and again at the end of Year 10. Teachers are expected to assess student achievement of these outcomes and report to parents on progress in their learning twice a year.

Assessment as an answer to higher expectations.

NSW had two external pen and paper tests as the primary means for satisfying stakeholders of the extent to which students had acquired the expected learnings, one at the end of Year 10 and the other at the end of Year 12. None of the other states and territories had a Year 10 science test. When NSW introduced the Year 8 science test from 2007 it was the only state to do so. Queensland introduced a science assessment program in 2009 for Years 4, 6 and 9, but abandoned it at the end of 2012 (QSA, 2012). Western Australia, introduced a science test for its students in Years 5, 7 and 9 from 2010 (SCSA, 2010) and abandoned it after 2013. Assessment in the junior years of secondary school in NSW was, and still remains, the responsibility of science teachers. They were supported in that task as discussed above in earlier sections of this chapter.

Goodrum et al, (2001) reported in their review that in secondary schools across Australia

> Traditional assessment practices remain as one of the most significant barriers to educational reform in secondary schools where teachers are required to cover too much content to prepare students for “the test”. Teachers indicate that tests are the most common form of assessment and, on average, represent 55% of the weighting of assessment […] Assessment is […] typically, summative, norm-referenced and focused on content. Students [report] that quizzes are frequently used to provide feedback to

> [them], however one-third of students indicate that their teacher never spoke to them about how they were going in science. (p. 155)

It is fair to say that NSW science teachers had a stronger tradition of external, summative testing embedded in their culture than other states and territories as is elaborated below.

Given the continuation with external testing for all students in NSW in Years 8 and 10 (up until the end of 2011) and continuing to this day at the end of Year 12, it is likely that the findings in 2000 might still apply in many secondary schools in NSW today. As the evidence from case study schools in this project shows, tests are a dominant form of assessment in science in Years 7 and 8 to this day. However, that assessment is now much more focused on the full range of outcomes and the shift toward the three bolded indicators of better assessment practice listed in Table 2.1 is well underway.

A role for SOLO to inform feedback about progress in learning.

As the discussion about the EV test indicated, the use of SOLO to provide an additional component of feedback about the level of understanding demonstrated by students in their answers was vindicated by PISA2015 testing that had items in it designed to provide feedback on the level of scientific literacy demonstrated by students. However, work in the US and elsewhere on how (to assess the full range of higher levels of cognitive functioning expected of students) has yet to be demonstrated as Ruiz-Primo reveals in her 2009 report to the US National Research Council.

Ruiz-Primo (2009) was asked by the US National Research Council to provide a paper that reconciled twenty-first century generic employment-related skills (NRC, 2008) and competencies at the core of science education (Duschl, Schweingruber, & Shouse, 2007). Her first comment was that expertise is located in a knowledge domain (science and technology in this case). She then goes on to elaborate that suitable science contexts need to be described to assess the extent to which students have acquired the following types of knowledge:

1. Declarative knowledge – knowing that
2. Procedural knowledge – knowing how
3. Schematic knowledge – knowing why
4. Strategic knowledge – knowing when, where, why and how to apply knowledge
5. Metacognitive knowledge – knowing about one's cognition and how to regulate one's cognition (with metacognitive strategies). (pp. 24-25)

Having reviewed the assessment frameworks for TIMSS, PISA and the US, Collegiate Learning Assessment (CLA) and National Assessment for Educational Progress (NAEP) science tests, she said that none of the current tests provide evidence for judging the degree of proficiency with all these forms of knowledge. However, she expresses the belief that access to appropriate computer-based technology (simulations) should enable tests that access all forms in the future. In the broad scheme of things, the inclusion of SOLO in the EV tests for NSW students (and Webb's DOK levels in the PISA2015 test) is a modest beginning to helping teachers support student acquisition of the highest levels of at least one of Ruiz-Primo's (2009) five types of knowledge, declarative knowledge.

The five types of knowledge described by Ruiz-Primo (2009) range well beyond cognitive functioning to include purposeful activity with other people and application of expertise to doing. Assessing performance in authentic settings is the preferred option here (Matters & Curtis, 2008). Choosing correct options from a battery of multiple choice items is not going to be seen as an authentic, valid or reliable demonstration of expertise needed in the 21st century by members of the ARG or researchers who hold situative or sociocultural perspective on learning (Billett, 1996; Cowie, 2013; Gipps, 1999; Lemke, 2001; Tobin, 2012), or by the wider community (Hattie, 2005). Nor is it a valid demonstration of the use of expressive language to construct a scientific report, explanation, or procedure, or for the justification of a course of action. Actual use of expressive language to represent knowledge and understanding in different learning domains has led to a view of science as a multi-literacy (Hackling, Peers, & Prain, 2007; Hand, Yore, Jagger, & Prain, 2010; Tytler & Prain, 2010; Waldrip, Prain, & Carolan, 2010). This view of science is discussed in the next chapter and used to justify the use of NAPLAN results as a valid predictor of scientific literacy as measured in the EV test.

The need to teach students how to learn so they can become independent learners

The research literature discussed in this chapter has identified that what teachers do with students in the name of science education accounts for 30% of the variability in achievement (Hattie, 2003b). What students bring to the classroom by way of natural ability, prior school experiences and family backgrounds accounts for 50% of the variability. The remaining 20% is attributable to how well the school environment (leadership) is managed to enhance the positive influences and minimise the negative influences on the overall learning of science in the school setting.

It follows that supporting teachers to do the best job they can is likely to have the most effect on student learning and engagement with science. Hattie (2005) has shown that teacher use of formative practices is one of the most effective ways to improve student achievement (as measured by large-scale test results). Other work looking at how to teach students to "learn how to learn" (LHTL) concludes that "emphasis should be placed on practices that have potential to promote autonomy in learning" (Black et al., 2006). One approach to doing this is to teach students how to learn by progressively giving them control and ownership of the strategies of formative assessment. Knowing how to learn and being motivated to do so (self-regulation) is probably the most important outcome for schooling.

The assumption that this capacity for self-regulation would show up in subsequent achievement in and engagement with science beyond Year 8 underpinned the third subsidiary research question in this research project. That question was: Is the use of formative practices by teachers linked to improvement in students' EV results and later achievement in and engagement with science?

Summary comments

This project is about the assessment-related work of science teachers in the early years of secondary education in a large government school system in one of the most advantaged and developed countries in the world (OECD, 2018; UNDP, 2018). It explores the impact of two assessment initiatives on teachers assessment-related work almost a decade after they were put in place. The constructs for five "dimensions of formative practice" are the windows through which that work can be examined.

Broadfoot (2009) implies that we are at a tipping point in our collective understanding and application of assessment:

> The purpose of assessment during the 20th century has been overwhelmingly the generation of summative data. The content addressed has concerned primarily cognitive tasks. The mode has been the largely traditional vehicle of paper-and-pencil tests and their organisation through large testing and assessment providers … Could it be, finally, that the grand narratives of intelligence and ability, which were regarded as the key to the determination of life chances, are beginning to yield to a more practical discourse of multiple experiences, skills, knowledge and dispositions? (pp. x-xi)

CHAPTER THREE: RESEARCH DESIGN, METHODOLOGY, METHODS

3.1 Introduction

Chapter Two provided an overview of the literature relating to assessment generally and formative practices in particular, along with the concept of a learning progression and SOLO theory. It also reviewed work being done with formative practices as a way of improving student capacity for self-regulated autonomous learning to equip them for lifelong learning, the latter being a highly sought-after outcome for education in the 21st century. The development of SOLO theory from Taxonomy to model and its use in the EV program in NSW schools and beyond was outlined as well.

This chapter describes and explains the research design and the methods used to collect data and information to answer the three research questions posed in Chapter One. The questions were:

1. What use are science teachers making of the EV program including SOLO and why is it used or not used?
2. What formative practices are evident in the work of science teachers and why are they used or not used?
3. Is the use of formative practices by teachers linked to improvement in students' EV results and later achievement in and engagement with science?

Section 3.2 provides a rationale for employing a mixed methods research design involving three phases. Phases one and two involved quantitative methods. The third phase employed quantitative and qualitative methods in the context of case studies.

Section 3.3 describes the first phase in which a quantitative method was used to deliver a sample of schools to work with. The quantitative method was a regression analysis using data provided by the Department for 394 government secondary schools in NSW. As will be explained in this section, the residuals from that regression analysis were used as a measure of the scientific literacy component of EV test results and as a measure of the effect size of science teaching. On the basis of the school residual, EV results for students at a school were designated as well above expectation (WAE), at expectation (AE) and well below expectation (WBE).

Section 3.4 describes the second phase that also involved using a quantitative methodology. Whilst the residuals were indicators of EV results above, at or below expectation and of the relative success of science teaching, the residuals say nothing about the teacher practices associated with those results. ANOVA was used to test for statistically significant relationships between aspects of the assessment-related practices used by teachers and EV results categorized as WAE, AE and WBE.

Section 3.5 explains the third phase involving case studies of assessment-related work practices in self-nominated government school science departments and of teachers working there. Quantitative data about student results and numbers of students completing Year 12 science courses were obtained from teachers at the case study schools, the state curriculum authority's website (the Board) and the *MySchool* website respectively. Qualitative data were also collected from teachers in the form of audio-recorded interviews and artifacts of assessment-related practice. Narratives describing the assessment-related work done by science teachers at each of the case study schools were constructed using interpretive methodology.

It was proposed at the end of Chapter Two that data collected to answer research question three could be used to test the proposition that students exposed to formative practices might be better self-regulated learners than those not so exposed. Section 3.6 discusses how data from the *MySchool* website was accessed and used to construct a basis for comparing schools in order to test three predictions designed to provide findings relevant to answering the second part of research question three. Statistical correlations were done to assesses the strength of association between achievement and engagement.

Section 3.7 discusses the limitations arising from the research design and methods used in this project.

3.2 Mixed method research, case studies and research design

Johnson, Onwuegbuzie, & Turner (2007) define mixed methods research as:

> the type of research in which a researcher or team of researchers combines elements of qualitative and quantitative research approaches (e.g., use of qualitative and quantitative viewpoints, data collection, analysis, inference techniques) for the broad purposes of breadth and depth of understanding and corroboration. (p. 123)

Following a review of many published studies, Creswell and Plano Clark (2011) proposed six mixed-methods research designs:

1. convergent parallel design
2. explanatory sequential design
3. exploratory sequential design
4. embedded design
5. transformative design
6. multiphase design.

Creswell (2012) characterises the first four of these designs as “basic” and the last two as “complex designs that are becoming increasingly popular” (p. 540). The explanatory sequential design method collects quantitative data first and then draws on qualitative data “to help explain or elaborate on the quantitative data” (p. 542). Creswell argues that explanatory sequential design (number 2 in the list above) can become a transformative design (number 5). The design becomes transformative when the explanatory sequential design is embedded within an overarching framework that

> informs the overall purpose of the study, the research questions, the data collection, and the outcome of the study. The intent of the framework is to

> address a social issue for a marginalized or underrepresented population and engage in research that brings about change. (p. 546).

This researcher intends at the conclusion of this study to provide feedback to all participant schools. The social purpose here is to assist schools, particularly regional schools where test results in science do not appear to be as strong overall as test results are in metropolitan schools. According to Flyvbjerg (2011), case studies provide the "concrete, context-dependent knowledge … necessary to allow people to develop from rule-based beginners to virtuoso experts" (p. 302). A case study conducted on a number of physically separated sites has been alternatively defined by other researchers as a multiple (Stake, 2005) or collective (Yin, 2003) case study. Given that one of the purposes for doing this study is to provide schools with feedback about practices, case studies provide a potentially powerful vehicle for doing so.

For Flyvbjerg (2011),

> (t)he decisive factor in defining a study as a case study is the choice of the individual unit of study and the setting of its boundaries … not so much making a methodological choice but a choice of what is to be studied. (p. 301)

The unit of study here is the set of assessment-related practices used by science teachers with their junior secondary science students in government schools and evidence of the impact of these practices on science achievement and engagement, both of which will be defined in the next section. The boundaries of the study were delineated by five constraints:

1. manageability of sample size for the researcher
2. purposive selection requirements
3. availability of volunteer participants
4. availability of relevant content
5. manageability for school participants.

First, the research was constrained by the number of schools able to be engaged with by a solo researcher. While 18 schools were identified and considered manageable, in the end only 16 were visited due to time and other constraints.

Second, the schools visited were purposively selected on the basis of their residual ranking. The aim was to work with schools as close to the top, middle and bottom of the three school groups that could be attained given the next constraint. Residuals, residual ranking and purposive selection will be explained in the next section (section 3.3)

Third, each participating science teacher had to be a volunteer and have the support of their department head and school principal. Research findings, in the event of publication, had the potential to be confronting so consent to collect information was asked for on the condition of anonymity for schools, teachers and students.

Fourth, school data sets, audio-recorded interviews and teacher-provided assessment artifacts all had to provide content relevant to or produced in the period of interest (2011-2014) as explained in Chapter One.

Fifth, the data-gathering exercises had to be manageable for school-based participants and seen as worthwhile from their perspective. This entailed the researcher being flexible in relation to his expectations of participants.

To summarise, the three phases of the research design and methods delivered:

1. three groups of schools differentiated from each other by an unconventional measure of scientific literacy attainment (a quantitative Phase One)
2. findings about science teacher engagement with the EV program (including SOLO) and formative practices based on the analysis of their responses to a common online survey, initially sorted according to the group of schools the responses came from (a quantitative Phase Two),

3. data and information about student achievement in and engagement with science up to the end of Year 12 plus information about assessment-related practices in the science departments of the 16 case study schools purposively chosen from each of the three groups of schools (a quantitative and qualitative Phase Three).

3.3 Phase One: selecting the sample of schools to work with

Bryman (2012) identifies nine approaches to purposive sampling, one of which, maximum variation sampling, he describes as "sampling to ensure as wide a variation as possible in terms of the dimension of interest" (p. 419). Flyvbjerg (2011) argues that by choosing "maximum variation cases" (p. 306) a researcher has the best chance of identifying findings that are either consistent or inconsistent with prediction or theory. The dimension of interest in this phase of the study is the scientific literacy attainment of students at a school. The goal was to select a sample of schools comprised of three groups whose statistical means for the measure of scientific literacy attainment were as different as possible.

As will be explained later in this section, a student's EV test results are a function of their general literacy and numeracy skills and their disposition to apply them to learning science. While a student acquires scientific literacy from many sources, the EV test targets the scientific literacy expectations described in the science syllabus that science teachers are expected to teach students.

The quantitative method used in this phase of the study separates the contribution of science teaching to student attainment of scientific literacy from other contributions. As a crude generalisation:

Scientific literacy attainment	=	**EV test results**	−	**general literacy and numeracy skills contribution**

The methodology used to achieve that separation and the thinking behind it follows.

3.3.1 Selecting the sample of schools to work with

As explained in Chapter One, this researcher first approached the NSW Department of Education in 2012 to discuss their possible interest in a proposal to research the impact of the EV program on science teaching in NSW. The Department agreed to assist.

The first step involved the Department psychometricians checking the integrity of data sets held for students who had sat EV tests in the four years 2011 to 2014. This check established that at least 465 schools had Year 8 students who sat EV tests in this period. To be eligible for this study, a school had to have a minimum of 10 Year 8 students who had sat the EV test in 2011. Department psychometricians also checked whether those same students had sat the Year 7 NAPLAN tests in 2010 and Year 9 NAPLAN tests in 2012 at that school. Comparable data sets for the next three years (2012, 2013 and 2014) had then to be confirmed. When this was done, the number of schools with sufficient students to meet the eligibility requirements was 394.

The next step was to use NAPLAN results to generate a science predictor that could be tested in a regression with actual EV results for the same students. The aim was to find a predictor that produced the best "fit" between a graph of the predictor (as the independent variable) and actual EV result (dependent variable) for pairs of students. The measure of "fit" is called the "coefficient of determination" (Laerd Statistics, 2013, p. 1) and has the symbol R^2. The value as a percentage (in this context) is a measure of how well the predictor accounts for the variability in the EV score. The closer to 100%, the more the predictor is said to account for the score in the EV test. A 'line of best fit' going through the graph of paired student results at a school can be drawn.

Plotted results are scattered above and below that line as shown in Figure 3.1.

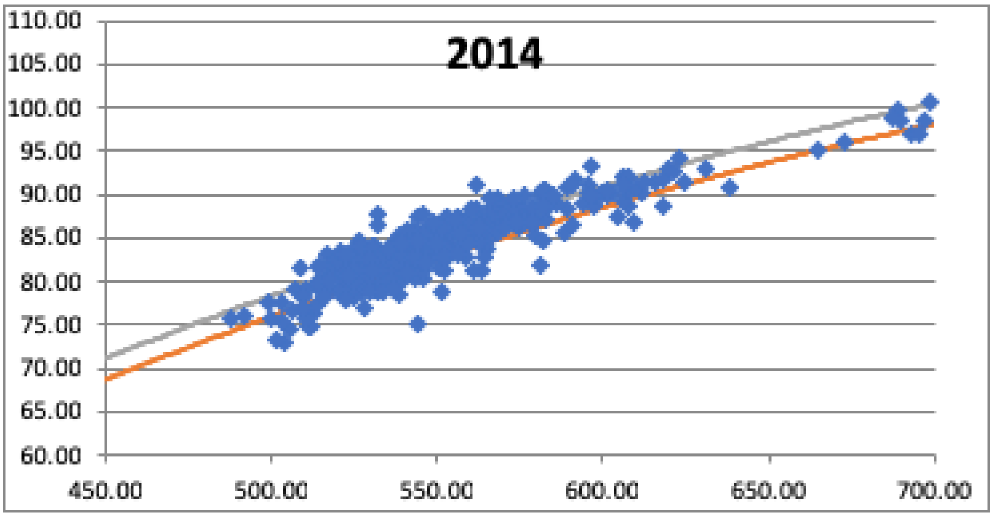


Figure 3.1 Regression of 2014 EV results over a NAPLAN-based predictor
Source: Department of Education, 2016

The two lines shown in figure 3.1 separate the top and bottom twenty percent of paired results. The statistical distance between the line of best fit and the plotted result is termed the residual. The residual size includes both measurement error and real differences between predictor and actual EV result. If the residual is above the line of best fit, then the EV result is positive and deemed for the purposes of this study, "better than expected"; if below, the result is negative and deemed "below expectation".

Four predictors of EV results were agreed to in discussions between researcher and the Department for testing. The predictors were: Years 7 and 9 literacy and numeracy results (combined and averaged); Years 7 and 9 literacy results only (combined and averaged); Year 7 literacy and numeracy results (combined and averaged); and Year 7 literacy results only. The Department performed separate regressions of EV results over the four different predictors and sets of residuals for the 394 schools for each of the years from 2011 to 2014 were calculated. A representation of the regression using 394 pairs of school results for the 2014 school year is provided in Figure 3.1. The blue diamonds are the paired school EV results (vertical axis) and predictor values (horizontal axis).

The slight curvature in the two lines delineating the 80th (top line) and 20th percentiles as drawn on the graph (see Figure 3.1) are the result of using first and third order factors (derived from the predictor) to provide 'lines of best fit'. Equivalent plots for years 2011, 2012 and 2013 were also produced.

The model of best fit turned out to be that the predictor based on the average of Years 7 and 9 literacy and numeracy results combined. The coefficient of determination (R^2) for that predictor, averaged over the four years of interest, was $R^2 = .892$. The four-year averages for R^2 for the other three predictors in the order listed above were .889, .887 and .870, respectively. The combined Year 7 and Year 9 literacy and numeracy predictor accounted for 89% of the explained variation in EV results across the state.

Residuals from the regression providing the line of best fit were used to generate three lists of schools from across NSW identified as having scientific literacy achievement well above expectation (WAE), as expected (AE), and well below expectation (WBE). The groups corresponded approximately with the top 20%, the middle 20% (spread evenly above and below the line of best fit) and lowest 20% of residuals respectively.

Science teachers from the three groups of schools with results identified as WAE, AE and WBE were invited to complete the same online survey (to be explained in the next section, section 3.4). The invitations identified a website for survey returns which was different for each of the three school groups. Chapter Four includes a statistical description of the sample and its constituent groups and analysis of those returns.

3.3.2 Regression residual as both measure of collective scientific literacy and 'effect size' of science teaching.

Six propositions provide the basis for using a regression residual as both a measure of scientific literacy and effect size of science teaching. The first proposition is that student responses to items and tasks in well-constructed pen-

and-paper tests (or online equivalents) provide valid evidence for making judgments about the level of achievement of many aspects of scientific literacy.

This proposition attracts support from Fensham (2013), who commends the EV test as an example of "a good model" (p. 18) in international comparisons. Rowe (2006) in preliminary commentary about relationships between PISA2003 reading, numeracy and scientific literacy results, says "*Reading Literacy* competence constitutes the foundational skill that underlies effective engagement with the school curriculum." (p. 9, Italics in the original)

The second proposition is that school science is a multiliteracy. Hackling, Peers, and Prain (2007) describe it this way.

> Science-specific as well as everyday literacies are required by students to effectively engage with science, construct science understandings and develop science processes, and to represent and communicate ideas and information about science. (p. 14)

While students acquire "science-specific" literacy from a number of sources, the EV test targets the "science-specific" literacy described in official curriculum documents. Science teachers are expected to teach that content and related vocabulary to students. As well, it is important to recognise that science teaching is expected to develop other science-related capabilities that are not directly assessable using pen-and-paper testing (such as those needed for managing practical investigations).

The third proposition is that according to the consensus of research reported by Hattie (2003b), only 30% of the accounted-for variation in achievement measured by tests is attributable to the experiences students have in the classroom; 50% is attributable to student factors such as ability and sociocultural background; and home, peer and school environment (physical, social and cultural) influences account for the remaining 20%.

The fourth proposition is that an empirically tested NAPLAN-based predictor of an individual's EV result provides the best independent measure of what is beyond the capacity of science teachers working in their science classrooms to influence. In other words, it is a measure of the factors Hattie (2003b) refers to in the previous proposition as responsible for 70% of the explained variation in achievement.

Of the four predictors tested for this project, the one based on an aggregation of Years 7 and 9 reading and numeracy scores, equally weighted, provided for 89% of the explained variability ($R^2_{av} = 0.892$) in the EV result over the four years of interest. The remaining 11% of explained variability is most likely attributable to the impact of science teaching. This is small in the overall scheme of things because, according Hattie (2003b), the teachers' contribution to achievement (in science in this case) is 30% overall.

This contribution of science teaching to science achievement (as measured in tests like the EV test) is so small that "maximum variation cases" (Flyvbjerg, 2011, p. 306) are sought to ensure the best chance of finding corroborating evidence that the residual is a measure of the effect of science teaching (such as better than expected scientific literacy achievement).

The fifth proposition is that the residual from a regression of actual student EV results over their predicted results is a valid measure of the impact of science teaching at the school level. It is a real effect (over and above the measurement error component) that contributes per student marks to both EV results and (to a lesser extent) to NAPLAN results. The residual is what you get when all the other contributions to EV results apart from science teaching are accounted for. This effect was designated for the purpose of this thesis as the collective scientific literacy score for the school.

The sixth proposition is that when standardised appropriately, the scientific literacy score is a valid measure for comparing schools. The standardised four-year average residual (actual marks) ranged from 2.68 marks per student per school above the state regression "line of best fit" and 2.50 marks per student per school below it. As a generalisation, as long as the same or equivalent sets of test results are used to generate the residuals, standardised scientific literacy scores (as represented by the residuals) provide a valid basis for comparing the effect of science teaching on individuals in a class; groups within a class; classes at a school; and schools in a district, state, nation or group of nations.

Given the above reasoning and conditions, NAPLAN-based predictors could also be used to assess the impact of teaching on achievement in other learning domains apart from science.

3.4 Phase two: online survey for science teachers

The main purpose of the online survey was to collect data from science teachers about their assessment-related work. Its other purposes were to collect direct evidence of teacher use of EV program resources and related understanding of the SOLO model embedded in the EV program. Both the EV program and embedded SOLO are specific NSW initiatives designed to support teacher adoption of formative practices. Findings from the analysis of survey return data were the primary source of evidence for answering research questions one and two.

3.4.1 Survey design

An initial set of items for the survey was constructed using ideas from a range of inputs that included the published work of researchers with a special interest in assessment, for example, Black, Harrison, Lee, Marshall, & Wiliam (2004), Hattie (2012), and Shute (2007). Another source of ideas for items was the NSW Board of Study's syllabus (BOS, 2003) and its sections on assessment for learning and the use of terms such as "practices" and "strategies" (pp. 71-75).

Other influences that impacted the content of survey items and the overall design of the survey were this researcher's previous experiences in the context of 'insider' work described in Chapter One. This work variously included critiquing, constructing and implementing surveys, collating and analysing the results, and providing feedback on proposed surveys. Two surveys that had a direct influence on the content and form of the final survey produced for this current research were:

- the telephone survey for secondary science teachers used to collect data for the Status and Quality of Science review (Goodrum, Rennie, & Hackling, 2001)
- a national survey on NAPLAN testing (Dulfer, Polesel, & Rice, 2012).

The former helped with the scope of the questions, the latter with the format of the questions and the decision to ask for personal information last of all.

To ensure face validity, items for the survey were refined in an iterative process involving several meetings with different groups of science teachers and one with education officers in the Department who had experience of survey design and expertise in assessment and SOLO. A draft version of the final survey was trialed online by five science teachers who volunteered to do so at the last meeting with science teachers. None of the trialing teachers were from schools subsequently invited to participate in the research.

It was this trialing that confirmed the 25-minute time allowance suggested for completing the survey online. Notwithstanding, the online version allowed teachers to stop, save and resume at will, and they were encouraged to keep a copy of their responses to check against the state results to be forwarded at a later date. Another purpose for an online trial was to ensure that the online platform holding the survey was working as anticipated. Following the trial, and after evaluating teacher feedback from meetings and one-on-one conversations with science teachers, it was decided that providing feedback to participants was an appropriate incentive.

The decision to use UTS's Survey Manager as the platform of choice for the online survey was based on:

- feedback from science teachers (convenience of online surveys and anonymity, if wanted)

- ease of distribution and management of returns
- support from experienced staff associated with the survey platform
- capacity for analysis using descriptive statistics of collected responses
- capacity to download to Excel and SPSS, if required, for more sophisticated analysis
- separate return of individual completed surveys with a date and time stamp to check the request for independent individual returns made in the survey itself.

The survey questions and related items are organised in four sections as shown in Table 3.1.

Table 3.1

Structure of online survey for science teachers

Section ONE: About ESSA/VALID

- Q1 a-i & Q2 a-m was about the EV program itself and included statements requiring yes/no responses about teacher engagement with test feedback data and components of EV program resources
- Q3 asked about their understanding of the purpose of the EV program (write a response)
- Q4 asked how well teachers understood the EV program (five point scale: very poor to very good)
- Q5 asked about intention to take up optional VALID 10 test (Yes / No / Unsure)

Section TWO: About SOLO

- Q6 a-j items here sought to discover the extent of teacher engagement with aspects of SOLO through a series of yes/no responses
- Q7 rate my understanding of SOLO (five point scale: very poor to very good)
- Q8 I learnt most about SOLO… (write a response)

Section THREE: About "Assessment for Learning"

- Q9 to Q15 were about formative practices. Questions and related items in this section were organised using the five dimensions of formative practices*
- Teachers were asked to choose between (not known-unsure about / never / seldom / sometimes / often) when responding to each of the survey items

Section FOUR: About your teaching experience / context

- Q16 to Q26 invited respondents to provide information about themselves, their experience and training and about their current school. The last two questions in the fourth section
- Q27 and Q28 asked teachers to participate in a follow-up case study and to identify themselves, their school and provide contact details to facilitate that if interested.

* Responses to survey items provided the opportunity to create individual teacher profiles in terms of the five dimensions of formative practices differentiated from each other by the relative strength of each dimension.

The complete set of survey questions is provided as Appendix F.

3.4.2 Analysis of survey responses

Cresswell (2012) describes a five step process for the conduct of hypothesis testing in the fourth edition of his handbook titled *Planning, Conducting, and Evaluating Quantitative and Qualitative Research.* The steps are:

1. *Identify your null and alternative hypothesis*
2. *Set the level of significance, or alpha (α) level, for rejecting the null hypothesis*
3. *Collect data*
4. *Compute the sample statistic*
5. *Make a decision about rejecting or failing to reject the null hypothesis.* (pp 188-195, italics in the original)

This procedure was generally followed in the conduct of analysis of quantitative data collected for this project and reported in terms consistent with current American Psychological Association (APA) protocols.

The design intention here was to characterise the assessment related-work of science teachers in terms related to their use of EV resources, SOLO and the five dimensions of formative practice. Further, the sampling methodology delivered the responses in three sets corresponding to the groups of schools with results labelled as WBE, AE and WAE. The three groups were in effect three separate populations. Survey returns constitute the samples representative of those populations. The separated survey returns presented the opportunity for testing the hypothesis that there were no differences in teachers' assessment-related work (the null hypothesis) despite the groups having EV results classified as WAE, AE and WBE.

The tools used to both manage and analyse the data collected from teacher responses to the survey were Microsoft's spreadsheet software, Excel and IBM's Statistical Package for the Social Sciences (SPSS) which was renamed IBM SPSS Statistics in 2014. SPSS software includes a range of statistical tools that can be applied to provide descriptive statistics and a range of inferential statistical analyses. Inferential statistics provide a method for “generalizing from a sample to a population.” (Lane, n.d.).

It was decided to use one-way, between-subjects ANOVA to test the null hypothesis (that teacher assessment related work was the same across the population of schools in each of the three groups of schools). If the analysis produced statistically significant differences in aspects of teacher practice between the populations and it was reasonable to reject the null hypothesis and consider an alternative hypothesis. The alternative hypothesis was that aspects of assessment-related work and student levels of scientific literacy are positively associated.

1The default assumption in SPSS for ANOVA calculations is the null hypothesis. Two errors are discussed in the statistics literature related to rejecting a true null hypothesis or failing to reject a false null hypothesis. Sample testing may return means differences that at first glance suggest population differences in the variable of interest when in reality the differences do not exist (a false positive result). As a consequence, rejecting the null hypothesis would be an error. This error is identified as a “type I error” (Lane, n.d., p. 377). SPSS software provides a printout of the target statistic and the level of statistical significance (designated by the letter p) related to that statistic. By convention. in social research, a p value below 0.05 (or .01 in some situations) is considered a reasonable basis for rejecting the null hypothesis (Bryman, 2012 and Cresswell, 2012).

In the event that there are actual differences between population means but the sample testing was not sensitive enough to revel the differences (a false negative result), it is possible to decide that the null hypothesis should be retained rather than rejected. This error (failing to reject the null hypothesis) is called a “type II error” (Lane, n.d., p. 378). The probability of making that mistake can be reduced by good experimental design and appropriate choice of statistical tools. The concept of statistical power is used in this context; it is a measure of “the probability of rejecting a false null hypothesis.” (Lane, n.d.). The greater the power the better.

The limitations related to using and interpreting the results of inferential statistics in this project will be provided in section 3.7.2 and 3.7.3.

3.5 Phase three: case studies and science department assessment related narratives

Both quantitative and qualitative data were collected for case studies of science departments in 16 schools. The methods used were audio-recorded semi-structured interviews, teacher-selected artifacts of assessment-related practices, and a proforma completed by teachers and populated with official school data about achievement in and engagement with science.

3.5.1 Audio-recorded semi-structured interviews: purpose and development

The purpose of the interviews was to collect qualitative data that could be interpreted to provide contextual information about the school and its science department's culture and practices, and from this to construct school-specific narratives about assessment-related work in the science department. Substantive content would be used to inform answers to the research questions.

The interview was semi-structured (after Bryman, 2012) using a set of key and follow up questions (to test silences in relation to options possibly forgotten). Given the demands being made of case study participants, a one-hour interview was considered sufficient for these purposes, and this proved to be the case. Because the interview was a one-off event, the questions sought responses to relatively specific aspects of assessment and related practices in the context of science teaching, many of which had been first raised in section three of the online survey that teachers had completed some months earlier.

A final set of questions was trialed at a school not involved in the research. The purposes for the trial were to assess the best place at the school to conduct the interview so that participants felt at ease; to test the language related to the conduct of the interviews; to check on the wording of questions; and to determine how best to describe the artifacts of interest for collection. The goal was to ensure that the interviewees felt as comfortable as possible as quickly as possible. A precis of the questions and purposes for asking them follows.

Questions 1 - 3 asked what prompted participants to join the case study. This was followed by two questions about their use of the EV test and related resources. The hoped-for responses were insights into what impact (if any) the EV program had on assessment-related work of teachers (section one of the teacher questionnaire).

Questions 4 - 8 asked about the collection and use of evidence of learning, and more specifically about peer and self-assessment opportunities given to students (again, seeking insights as to the extent to which these two key aspects of formative practice were a priority in teacher thinking at this school).

Questions 9 – 10 asked about school and science department priorities in an attempt to gain some insight into their alignment. Based on this researcher's experience, there was a likelihood of school priorities being formative assessment and/or the development of student literacy and numeracy skills, the latter being an attempt to understand whether there is an emphasis on 'writing to learn' and, if so, to what extent has it been take up by teachers in the science department.

Questions 11 – 12 were about resources used to teach science and a question about how knowing whether what one is doing works (as a test of their commitment to assessment). This was also related to surfacing understandings about using the same resource for both teaching and assessment.

An opportunity was provided in relation to the online survey teachers had completed some months earlier for interviewees to explain how they decided what were the appropriate response options from among the choices: not known/unsure about; never; seldom; sometimes: and often. The purpose here being to check that the basis for choosing was similar for all respondents.

A question was asked about what interviewees understood progression in learning science means (given that SOLO provides one and the syllabus outcomes in a standards framework another). The concept of a progression in learning is a strong theme in the research on formative assessment (see Chapter Two).

A question was asked about the regularity of science department meetings was, as was one about the nature and extent of discussions about assessment at those meetings. It was hoped that discussion here might provide insights into practices around the setting and assessing of student tasks; how issues about reliability and validity are dealt with; and whether the meetings provided opportunities for teachers to display good learning behaviours with each other.

An opportunity was provided for interviewees to discuss what, if anything, had surprised them about aspects of their school EV results or student survey feedback, Year 10 or Year 12 data put into the proforma. This question was exploratory, and hoped-for responses included references to how the science department was responding to student perceptions of their science experience or the extent to which this exercise in result analysis was more or less than what is currently the norm.

Interpretive analysis by the researcher of teacher responses to the interview questions was an iterative process. The process involved the production of comprehensive, holistic, qualitative descriptions (Sandelowski, 2000) of practice framed by the interview questions. The purpose of the analysis was to generate narratives including examples or contexts to support and illuminate answers to the research questions.

All 16 recorded interviews were listened to at least three times. No more than four interviews were listened to and analysed in any one day. The elapsed recording times to uniquely descriptive instances of practice in the context of that school was noted (to enable efficient return to then at a later time for additional replaying). Notes were created during the first replay to summarise responses. Replay was stopped and rewound over some sections to check that the record was a clear and accurate summary of what had been said.

While the second replay was in progress, the first set of notes was checked to ensure key activities, strategies, examples or insights related to formative practices already noted were consistent with what was being said. At the third listening, prior notes were compared with what was being heard to ensure all key insights and examples were appropriately referenced, and further additions/corrections were made when considered appropriate.

The ten interviews with the case study schools reported on in Chapter Five were then listened to again before writing the assessment narratives using the following scaffold. The components of the scaffold were derived from the teacher interview questions (A and B), the teacher survey questions including the five dimensions of formative practice (C to G). The last component (H) was an opportunity to provide summative comments identifying unique practices or commonalities with other schools.

A. Engagement with EV feedback, resources and SOLO

Any references to the EV program, how it was valued compared to NAPLAN, issues with doing the tests (students, staff supervision or access to computers), feedback used (or ignored), and impact on science assessment generally were reported here. Any references to SOLO or its uses were also reported here.

B. Grouping for instruction

The sources of assessment data used to establish Year 7 classes, who did it and how it was used to allocate students to groups for instruction are reported here. Classes so formed were variously labeled as mixed ability, graded, streamed, or parallel. The timing and basis for changing student allocations to classes as they progressed from Year 7 to Year 9 were also reported.

C. Use of learning intentions and success criteria

In this section, school and science department teaching and learning priorities and their sources were recorded. The form of teaching and learning program components that communicated learning intentions to teachers were noted. Also recorded were details of assessment tasks, priorities as revealed in the related rubrics, and alignment with syllabus intentions. The links between success criteria, mark allocation and subsequent conversion to grades for the purpose of reporting to parents was also examined. The researcher also listened for evidence of student involvement in devising or choosing either learning intentions or success criteria.

D. Classroom discourse and evidence of learning

Teaching science involves engaging students in a range of activities, including using equipment to measure and record observations; accessing second-hand sources of data and information; and designing and carrying out investigations to solve problems and answer questions. It involves working alone and with others and it may take place in a regular classroom, a dedicated space with special fittings (such as school laboratories) and access to a range of specialist equipment (including ICT based tools), or it may take place beyond the school walls. Of interest here was the extent to which teachers made use of the diversity of options in these settings to observe evidence of learning and how they managed the discourse so that evidence of learning was made explicit.

E. Feedback

This section records who did what with the evidence of learning produced from teaching and learning activities (such as those described in the previous section). In particular, it was useful to record whether the feedback provided sought to progress learning for both the student/s and their teachers, and whether it was about what form the completed task would take, the skills to be improved, metacognition, or praise for the learner (such as a tick or comment). Of interest too were the referents for criteria used in feedback. Referents of interest here were syllabus intentions (scope of responses and/or depth), misconceptions, SOLO levels of thinking, or some other referent such as the Board's *Common Grade Scale*. How accumulated marks are converted to grades for reporting purposes was also of interest here.

F. Activating students as instructional resources for others

Here the emphasis was on recording the opportunities students were given to provide peer feedback and the guidance to ensure that it was a productive process for both the provider and recipient. Examples might include structured group work where students are assigned roles or given opportunities to demonstrate to or instruct others; teacher use of strategies such as predict-observe-explain (POE); think-pair-share-report; jig-saw; or joint construction of student responses to phases in an investigation.

G. Activating students (and teachers) as learners

In this section the focus was on reporting examples of good learning behaviours modeled by either or both students and teachers. To be worth noting, the opportunities had to be explicitly provided (such as keeping reflective journals, choosing items for a portfolio, defending choices, or making links to previous learning in science and/or other subjects). For teachers, opportunities may include working collaboratively with each other to mark assessment tasks; annotating work samples to use when converting marks to grades; identifying mark cut-offs for converting to grades; developing further understanding about what a progression of learning in science looks like; developing a "scope and sequence" for a unit of work; or developing an assessment rubric that includes criteria for rewarding different levels of student response to an item or task.

H. Comparative summative comments

Summative statements relating comparative achievement and engagement to aspects of formative practice revealed in interviews and artifacts, along with commentary about the extent of confirmation for the predictions (or otherwise), completed the reports.

3.5.2 Artifacts of assessment practice: purpose

Schools identified for participation in the case studies were advised in an email to collect any documentation, models (or images of same) used to inform or support assessment-related work in science at the time of his visit. Artifacts sought were examples of things teachers considered to be 'best practice'. The purpose was to use the artifacts to confirm interview and survey responses and to provide examples to illustrate assessment-related narratives developed for specific case study schools. The artifacts asked for included:

- teacher-devised assessment policies to guide assessment-related work of science teachers
- formal reports of achievement or progress by students (the ones sent home)
- examples of assessment tasks
- learning programs where specific references to assessment were made
- lesson plans or student 'worksheets' where assessment-related activities were the main focus
- annotated exemplars of quality work at different levels produced by students
- rubrics used to assess activities and to provide feedback to students.

Analysis of the collected artifacts of assessment practice was performed after listening to and summarising the interviews. The focus was to look for confirmatory/contradictory/additional information to illustrate the narratives for each school.

3.5.3 Case study school data: purpose

Participating teachers at case study schools were asked in advance of a school visit to provide school-level data about EV achievement, Year 10 results and Year 12 science course completion data relevant to the years of interest (2010 to 2015). Participants were sent a proforma (in both hard copy and as an excel spreadsheet) to assist them prepare for a planned visit. The school-specific information was sought to provide data about later achievement and engagement (explained in subsection 3.5.5), both of which were relevant to answering research question three and for assessing predictions related to self-regulated learning. The proforma sent to schools is attached as Appendix E.

The EV data requested of case study schools was for the years from 2007-2015. It transpired that in most cases respondents were only able to access data in SMART for the years 2011 to 2015. SMART is the acronym for School Measurement Assessment and Reporting Toolkit. It is sophisticated software tool provided online to schools by the Department and it can be used to perform limited forms of analysis on test results from external testing.

Data for the years before that were apparently unavailable to the respondents, except for three schools where the data had been retained in science department records. Other data relating to Year 10 results and numbers for Year 12 completions of senior secondary science courses were available to schools in the Board-provided Results Analysis Package (RAP). Most schools did not retain the Year 10 data as part of their science department records. Year 12 results were generally retained at the science department level and was provided to the researcher in all cases. Most schools had to ask the Head Teacher English for their numbers in order to calculate the proportions (as a percentage of the English candidature) of students doing the various science subjects.

It is for the principal to decide who at a school has access to SMART and RAP. The purpose for asking schools about their results was to collect information during interviews about how that information was used to inform assessment-related work in the school and in its science department. Only three schools brought completed proformas to interviews. The remainder provided them after the interviews. In a few case study schools, this information was not immediately accessible to science teachers other than the HT.

The researcher had Departmental approval to access and use aggregated school-level results. However, access to the pattern of school results was at the discretion of school principals. Access to the results was provided by the Principal in all but two schools who withheld the Year 10 data requested.

Feedback to schools, students and parents from the Department about EV test results is provided in SMART. The proforma provided to schools included tabulated spaces for school-level data for four of the five reporting categories relating to EV test results. Student achievement data for the school and state are both reported in SMART against three achievement bands.

School-provided and other data from case study schools were collected from schools and recorded in an Excel spreadsheet which was later transferred to SPSS in order to perform statistical processes with the data. As will be discussed in more detail in Chapter Five, six items of the 21 in the survey will be reported on in this thesis. Analysis of student survey responses was designed to provide patterns of difference in strength of agreement /disagreement on each of the items within and between the paired schools. This analysis was straightforward. The mean scores were printed out as tables and different coloured hi-lighters were used to identify each school's difference with the state population rating (above, below, the same each had different colours).

How later achievement and engagement in science were assessed for the purposes of this research is explained next.

3.5.4 Defining later achievement in science

The measure of students' later achievement in science was the pattern of grades awarded to students at the end of Year 10 (two years after the EV test) based on school processes and endorsed by the Board. An option would have been to include end-of-Year 12 results in science as well. This was not done for two reasons: first, because the data collected about assessment practices was specifically focused on the first three years of secondary schooling; and second, to reduce the amount of time required of participating teachers.

The issue of assessing improvement in achievement over the years within a school is not straightforward because the basis for both assessing and reporting achievement is different at each of the two chosen points of interest. The key differences in the reporting of achievement are outlined next.

Results for the EV test are a one-off summative assessment reported in levels from 1 to 6 referenced to a scale based on SOLO levels of learning. In SMART, a second way of providing feedback on results is to report it as the proportion of students in achievement bands (three bands are used: band 1, band 2 and band 3, the latter being the proportion of students at the school attaining levels 5 and 6, the highest two levels).

Student achievement at the end of Year 10 is reported as grades (A to E, A is highest). The grades are referenced to the Board's (BOS, 2013). The Scale describes five standards of achievement. In all of the 16 schools interviewed the grades awarded by teachers are based on their judgment of the standard implied by particular mark ranges within the range of aggregated marks for all tasks completed in that year. For example, marks ranging from 60 to 70 (out of say 100) might be indicative of work consistent with that described for a B grade on the Board's Scale. The Board's *Common Grade Scale* is in not related to SOLO levels.

Case study schools provided the proportions of students obtaining the top, middle and bottom achievement bands in Year 8 for their school, and comparable data for the state in the years of interest. Schools also provided the proportions attaining grades A to E in Year 10 for their students. The relative proportions of students obtaining A to E in science in the state in the years of interest were obtained from the NSW curriculum and assessment authority's website (NESA, 2017).

Thus, changes in intra-school proportions relative to the state at both Year 8 and Year 10 provide a good basis for monitoring later achievement (Year 10 compared to Year 8).

3.5.5 Defining engagement with science

In the context of this project, a simple operational view of engagement was chosen to assess the extent of later engagement with science (see research question three). It was chosen for pragmatic reasons relating to data availability and the sense with which education minister Tebbutt used it when announcing the EV program in 2005 (see section 2.2). As students are free to choose whether or not they take up science courses after Year 10, comparing the proportions of students completing science courses at the end of Year 12 was chosen as the measure of later engagement (see research question three).

Because science is a compulsory course for the first four years of secondary schooling a different way of assessing engagement was needed. Student responses to items from the student survey accompanying the EV test provide an alternate way of measuring engagement at the end of Year 8. Items in the EV survey asked students to rate on a four-point scale their agreement (or disagreement) with a series of statements related to science and their school science experience of it. Selected survey items (six of 21) were chosen as the basis for measuring engagement.

The items chosen covered interest in science, enjoyment of science in primary and secondary school, perceived difficulty of science relative to other subjects, perceived success in learning it and whether it was one of their favourite subjects. These (and other) aspects of affect appear in research papers attempting to define engagement with science (including its retention when free to drop it). See for example the UK's, National Foundation for Educational Research (2011) report titled *Exploring young people's views on science education* where some of these aspects are discussed. At this time, there does not appear to be an evidence-based consensus about how best to define engagement. Student feedback on aspects of affect addressed by the items provide data for evaluating their usefulness as markers for student self-regulation which is explored in Chapter Five.

Later engagement

The data provided by schools was used to generate an operational definition of later engagement. English is a mandatory course for all students wanting the Higher School Certificate (HSC). The HSC is the school exit credential provided to students from NSW schools who want it as support for entry to post school options including work and higher or further education. Science teachers were asked to convert their science course completion numbers to a percentage relative to English numbers at the school.

English and senior science courses each year across the state for the purposes of the HSC were obtained by the researcher from the NSW Education and Standards Authority website (NESA, 2017). Statewide proportions relative to English were calculated for the state. These two sets of numbers, school and state proportions, provide an objective basis for making inter-school comparisons related to engagement as defined above. Students make a choice to continue with or drop science after Year 10. Thus, the measure of engagement based on proportions completing science courses (relative to English) at the end of Year 12 would appear also to be a strong measure of the collective valuing of science by students at a school half-way through Year 10 when they make their choices for subjects to study in the senior years of schooling.

The student survey component of the Year 8 EV test provides a way of measuring students' level of satisfaction with their experience of science in the first two years of secondary school. When aggregated and averaged over the years of interest and compared to statewide data, the survey provides an objective measure for engagement that can be compared over time both within the school and between schools (when referenced to state proportions).

The data used to produce a Year 8 measure of engagement were collated from school records by teachers in case study schools. The data asked for was a subset of the EV feedback provided in SMART. Given that one third of the teachers had said in survey responses that they had not looked at survey data (see section 5.6.2), responses to only six identified items (of 22) were asked for. SMART includes the state proportions of students at each achievement band level for each item.

While it is true to say that cross-school comparisons for both achievement and engagement can be made using objective measures, these would almost certainly not be valid unless other factors contributing to the scores are made explicit. Those doing the comparison are then able to make an informed judgment about differences after considering the likely impact of these factors. This issue is dealt with in the next section.

3.6 Comparable schools and three predictions

The point of making intra-school comparisons for achievement in and engagement with science is to assess whether, over time, successive cohorts of students are doing better at key points in the journey through secondary schooling, such as at the end of Years 8, 10 and 12. In other words, are the refinements being made at the school level to teaching and learning programs in the light of feedback resulting in better overall achievement for successive cohorts passing through those points? Changes to achievement and engagement patterns that reveal growing proportions of students at higher levels/grades/numbers taking senior science courses would no doubt be welcomed as evidence of improvement and be entirely consistent with the use of formative practices by science teachers.

In the context of this research, inter-school comparisons provide a means for independently testing the validity of a claim made in Section 3.3. The claim there was that the size and sign of the regression residual is a direct measure of the scientific literacy component of EV results…the more positive the residual, the bigger was the contribution to the EV result overall (making it better than expected). Also, it was claimed there that the scientific literacy effect is directly related to the impact of science teaching. If that is a valid claim, then for a pair of comparable schools (one in the WAE group of schools and one in the WBE group, say) the actual EV results in the WAE school should be better than the results in the WBE school. The meaning of 'comparable' is explained below.

Comparable schools were defined by the researcher as schools having the same or very similar SEA scores. SEA is the acronym for "socio-educational advantage"(ACARA, 2014b, p. 3), which is a measure published for schools on the *MySchool* website. It is an independent measure of the capacity for learning each student brings to school. This measure of student educational disadvantage/advantage is determined from parents' levels of education, occupation and post-school qualifications. A fourth category of current employment status of parents was added to SEA determinations from 2013 onwards (ACARA, 2014b) because it was found helpful in improving the correlation between the SEA score, ICSEA and subsequent NAPLAN results.

The School Profile page for each school on the *MySchool* website provides the SEA data as a quartile profile showing the proportions of students at that school in the four quarters from the most educationally disadvantaged to the most educationally advantaged. In order to protect the identity of the school, the SEA profile data for each school was used to produce what the researcher called the SEA score. The profile quartiles were converted to a single score on a scale of 0 – 10 using a simple linear transformation. The lower the number, the larger the proportion of educationally disadvantaged students at the school; the higher the number, the larger the proportion of educationally advantaged students at the school. The SEA score for each school is the four-year average of the SEA scores for the Year 7 intakes in 2010 to 2013, inclusive.

The reasoning behind the decision to use the SEA as the control follows. In Hattie's (2003b) terms ACARA's SEA is equivalent to the student factors that he says provide 50% of the accounted-for variability in test results. The measure of regional remoteness and percentage of Indigenous student enrolment, which ACARA refers to as school factors, are equivalent to the factors Hattie (2003b) says contribute up to 20% of the accounted-for variability in test results. What the teacher does in the classroom contributes the rest, he says.

The ANOVA performed on teacher responses to the survey questions about formative practices provided a profile of science assessment-related work for the sample of teachers from each of the school groups. If the sample means related to the dimensions of formative practice in, say, the WBE and WAE sample were shown to be significantly different, the difference in practice associated with that mean was then generalized to apply to all the schools in that group. The group profile is described in terms of the five dimensions of formative practice. If the EV results for comparable (that is, having the same SEA scores) schools are statistically significantly different in the way predicted by the residual, then it would be reasonable to attribute that difference to the formative practice profile of science teachers in that group of schools. This is because the residual assigning the school to a particular group is also an imputed measure of the 'effect size' of science teaching.

The strength of the relationships between school group, EV results, and formative practice profiles can be tested using correlation statistics which SPSS has the capacity to perform. As well, according to the research evidence discussed in Chapter Two, if formative practices are more frequent in WAE schools then we could reasonably expect that students in the WAE school are, collectively, more skilled at learning and more motivated and engaged than students in the WBE school. If that is the situation at the end of Year 8, it could reasonably be expected that students in the WAE school would apply those skills, motivation and engagement going forward, with the same relative effects on achievement at, say, the end of Year 10.

With the above in mind, three predictions were made:

1. Overall EV results for students in comparable schools will be better in WAE schools than AE schools, and AE school results will be better than WBE schools.
2. Overall Year 10 science result patterns for students in comparable schools will be better in WAE schools than AE schools, and AE school result patterns will be better than WBE school patterns.
3. The proportion (relative to English) of students completing Year 12 science courses in comparable schools will be highest in WAE schools, and AE schools will have a higher proportion of completions than WBE schools.

Verification of the predictions and related discussion drawing on the assessment-related narratives particular to the case study schools will be provided in Chapter Five.

Findings (Chapter Four) and assessment-related narratives (Chapter Five) provided the data and information used to inform discussion reported in Chapter

Six about the impact of formative practices on student learning of science in the early years of secondary education in NSW government schools.

3.7 Limitations

Specific factors that impact the trustworthiness and the validity of findings in both qualitative and quantitative research generally and in this research specifically follow.

3.7.1 Trustworthiness of qualitative research

To ensure the persuasiveness of the answers relating to the "why", "how" and "impact" components of the research questions, this researcher took steps to ensure that the evidence used to construct answers satisfies the four criteria for a "trustworthy study" (Shenton, 2004, p. 64): credibility, transferability, dependability, and confirmability. Potential concerns that the researcher in this project should have been positioned as a participant researcher / observer (Denzin & Lincoln, 2011 and Hammersley, 2008) are addressed.

Originally proposed by Guba (1981), Shenton has used the above four criteria in his own work, claiming that the criteria have been "accepted by many" (Shenton, 2004, p. 64). Shenton (2004) argues that these criteria are analogous to four criteria used by positivists to defend their work. Credibility is the qualitative research analog for internal validity; transferability is the analog for external validity/generalisability; dependability replaces reliability, and confirmability replaces objectivity.

Credibility

Credibility is about congruence of findings with reality (Merriam, 1998). Transferability is about providing enough contextual detail for a person to make a judgment that "findings can justifiably be applied to [a different] setting" (Shenton, 2004, p. 63). Dependability is difficult to achieve in a qualitative study, but a goal should be to have sufficient detail to enable "a future investigator to repeat the study" (Shenton, 2004, p. 63). Confirmability is about "researchers [taking steps] to demonstrate that findings emerge from the data and not their own predispositions" (Shenton, 2004, p. 63).

In relation to credibility, Shenton (2004) advocates 14 "strategies" (p. 64) that may be used to achieve credibility. These include using well-established methods in qualitative research "in general and in [education] in particular" (p. 64). Interpretive analysis of interviews and artifacts of practice within the constraints of a case study is a well-accepted methodology in qualitative research. In this project, interpretive analysis of semi-structured interviews and artifacts of assessment practice that were selected by teachers as representative of their 'best practices' produced data and findings about context relevant to understanding results obtained quantitatively.

Another strategy for ensuring credibility is researcher "familiarity with the culture of participating organisations" (Shenton, 2004, p. 65). This researcher's direct and continuous involvement with science education since the late 1960s was an important factor in his decisions about what to ask of participants in the case study components and, as mentioned earlier in Section 3.4, his choosing and devising items for the online survey. Other strategies mentioned in relation to credibility include tactics to ensure respondent honesty including iterative questioning. These were explicit considerations at various points in the research reported here.

Using multiple sources and multiple data collection strategies is another way to promote credibility in research. In this research project, some of the interview questions sought to corroborate the extent of shared understanding between interviewee and interviewer (this researcher) when it came to items listed in the online survey. Examples from the online survey include item 10e about the use of think-pair-share-report strategy; item 11c about the use of grades as a form of feedback; item 15e about how staff develop a shared understanding of what progression in learning science looks like; and a direct question asking teachers how they decided between often, sometimes and seldom when considering how frequently they employed the activities/strategies described in the online survey items.

Transferability

Transferability is the second criterion used to establish trustworthiness. In relation to qualitative research, this is contentious because of the limitations imposed by the boundaries of case study work. Shenton (2004) says:

> Ultimately, the results of a qualitative study must be understood within the context of the particular characteristics of the organisation or organisations and, perhaps, geographical area in which the fieldwork was carried out. (p. 70)

Shenton (2004) cautions that when inconsistencies are found, this may not reflect on the trustworthiness of the research but may be an indicator of multiple social realities. In this research, every attempt was made to provide sufficient contextual information for people to make a judgment about the contention that formative practices have a demonstrable impact on science learning and related attitudes to science.

Dependability

Dependability is the third criterion. The detail provided about the conduct of the research reported here should enable a person to repeat the process at another place or in a future time period. Their intention might be to confirm findings, but equally, it might be about whether a different reality is a better fit for the findings.

Confirmability

The fourth criterion of trustworthiness (confirmability) can be provided by triangulation to check investigator bias (or to assess participant researcher / observer bias); making explicit the researcher’s beliefs and assumptions; drawing attention to limitations of the methods used and their potential impact on findings; and describing explicitly and in detail the methods that enable scrutiny of results. The details provided in this thesis relating to the methods and sample sizes should enable a reader to verify for themselves the findings, inferences and conclusions.

In this project, interpretive analysis of semi-structured interviews and artifacts of assessment practice that were selected by teachers as representative of their 'best practices' produced data and findings about context relevant to understanding results obtained quantitatively.

3.7.2 Validity and reliability of quantitative data

Quantitative research criteria relating to validity, reliability and objectivity have long been touchstones for assessing the worth of research findings (Bryman, 2012). In the application of statistical methods to provide an objective basis for reporting findings, a distinction is made between descriptive statistics and inferential statistics. Descriptive statistics include concepts such as sum, average, mean, measures of frequency, measures of distribution. Inferential statistics involve the use of concepts such as correlation, probability, statistical significance, power and confidence levels in discussing test results.

SPSS software provides tools to analyse quantitative data and produce a range of descriptive statistics characterizing the data. Features of the data can then be evaluated for impact on the inferential statistic of interest. Data may be judged as being either parametric or non-parametric and the appropriate tool can be chosen for the proposed test, such as ANOVA. The accuracy of the calculated ANOVA statistic may be compromised (and in extreme situations, invalidated) by using data that does not fully comply with all the data assumptions for parametric analysis which, according to a Laerd Statistics (2018) tutorial and Lane (n.d.), are:

1. The dependent variable should be measured at the interval or ratio level
2. The independent variable should consist of two or more categorical, independent groups
3. Independent observations (no relationships between the groups; no subject in more than one group)
4. No significant outlier data values
5. Dependent variable data should be approximately normally distributed for each category of the independent variable
6. Data displays homogeneity of variation

7. Sample numbers in the different groups are approximately equal.

According to Lane (n.d) and Rennie (1998) the power of the statistic being calculated using samples is enhanced (reducing the chance of failing to reject the null hypothesis) when the:

1. sample size is large
2. standard deviation is small
3. difference between the hypothesized and actual means being compared are large
4. significance level is less stringent
5. a test is one tailed (and the hypothesized direction is correctly specified).

In the event that the parametric statistic and related statistical significance figure based on an assumption of parametric data is inconclusive, post hoc tests based on the assumption that the data were, in effect, nonparametric may be more powerful or robust and provide a reasonable basis for rejecting (or retaining) the null hypothesis.

Teacher survey responses and the school-level data sets for EV test results, Year 10 assessments and Year 12 science course completion numbers were processed using both descriptive and inferential statistics. Findings from the applications of statistical processes will be provided in Chapters Four and Five.

3.7.3 Summary of limitations affecting this study's findings

Qualitative data

Two limitations in relation to the artifacts collected for this project are worth mentioning. The first was that, for the most part, artifacts reflected current practice and with few exceptions had been produced in the two years preceding this research in response to the introduction of a new syllabus that was being implemented from 2014. The years of interest for this project predated 2015. The second was the extent to which the artifacts were representative of the diversity of teacher practice.

In the end, the samples provided were assessed for alignment between aspects of the provided assessment rubric and syllabus intentions (as expressed by outcomes, access to related content prescribed by the syllabus, and the context in which the activity was embedded). The syllabus then, as now, intended teachers to provide contextualised activities to engage student interest.

When considering the characterisations of formative practices produced from teacher survey responses, it was important to remember that the profiles drawn were only in relation to practices in Years 7, 8 and 9. This is relevant to any discussion about the extrapolation of findings in relation to the three predictions described in Section 3.6.

Great care when trying to interpret teacher responses to interview questions about assessment-related practices had to be taken for two reasons. The first was that this researcher (who conducted the interviews) knew only one of the participating case study school science teachers before the interviews. He had attended a two-day workshop presented by this researcher more than ten years earlier. Initial natural reserve when it came to disclosure of practices was evident in most cases.

However, an hour is a generous time for a one-on-one discussion and most participants seemed to appreciate the opportunity to discuss their practice with an interviewer who understood their situation and to whom they could make frank disclosures about their work. No interview was terminated before the assigned time; most went longer.

The second reason was that the interviews were being conducted in 2016 about assessment practices related to a syllabus that schools were no longer working with (it was replaced after 2014). The new syllabus was sent to schools in 2012 and science teachers were encouraged then to begin planning for its implementation into Years 7 and 9 from 2014 and Years 8 and 10 from 2015. The new syllabus became the basis for EV testing in Years 8 and 10 from 2015, the year after the period of interest for this thesis.

This meant that HTs in case study schools were managing syllabus implementation processes that had been in progress for at least two years after the period of interest relating to achievement. These processes included reviewing and adjusting the set of summative assessment tasks to reflect new syllabus learning intentions. In practice this meant very little change in the subject matter and the weighting between knowledge and understanding and skills was the same (50:50).

A number of the case study schools had changed the assessment modes used to collect assessment data. Some replaced formal pen-and-paper tests with research projects, practical tasks and oral presentations. The issue was to work out whether what was being provided in the discussion and artifacts were recent innovations (i.e. had been introduced after 2014 or were in place before that).

Artifacts of assessment-related practice provided by teachers needed to be considered in the light of recency as well. The main issue was to work out which part of the school narratives about assessment for learning applied before or after 2014. Questions from the interviewer were used to assist with that where necessary.

Quantitative data

The criterion of data independence was provided by an experimental design that asked for and delivered responses from either WAE or AE or WBE designated schools to three different websites. The instructions with the online survey were explicit in asking for individual responses. A check on the timing of survey returns supported the assessment that returns were from individuals even when multiple returns from (teacher) identified schools were received. One school that identified itself said it had provided a consensus return from the five teachers comprising the science department. It was treated as an individual return for the purposes of this exercise.

The data normality requirement was tested using the Shapiro-Wilk test in SPSS. The SPSS tutorial advice was that the Shapiro-Wilk test is more appropriate for sample sizes less than 50 (Laerd Statistics, 2017).

The requirement for homogeneity of variance was tested using both the parametric Levene test and Welch (nonparametric) test for 'robustness of means equality' and the most appropriate test result was reported. Both of these tests are readily available in SPSS.

If the ANOVA statistic for the between-group means was statistically significant, the nonparametric Games-Howell Multiple Comparisons Test was used to identify the groups with statistically significant means. The Games-Howell test is recommended where the group sizes were relatively small and unequal in number, and, as in some cases, data sets were borderline in terms of homogeneity of variance and normal distribution (Laerd Statistics, 2017). The Tukey HSD test is a parametric test and was not an appropriate test in most cases. These two tests (and more) were readily accessible within the SPSS software used.

Because the survey was voluntary and anonymous, it was not possible to predetermine the total number of responses or how the individual response numbers would be distributed across the three populations. As a consequence, the group sizes were unequal and the number of subjects relatively small. While there were 101 respondents in total, only complete or almost complete data sets for sections being analysed were used. The number of data sets remaining in each group were: n_{WBE} = 32, n_{AE} = 28, and n_{WAE} = 25, meaning that data from 16 (15%) of the respondents was not used. The impact of missing data within the data sets used was managed by the SPSS tools used to report the statistical significance of the statistic produced.

The nonparametric Kruskal-Wallis ANOVA was generally used where tests for homogeneity of variance and normality were not completely satisfied.

Year-on-year variability and school misfortunes can impact results in a one-off test. Examples might be the death of a teacher or a student, as well as individual student circumstances. The relative impact of individual or group misfortune on aggregated results is inversely proportional to the Year 8 population. For example, one student dropping an achievement level in a school's Year 8 population of 30 produces a 3% variation in the proportion of students at that grade level; in a Year

8 with 100 or more students, the impact is of the order of a 1% variation or less. Averaging results over four years reduces the impact of year-on-year variations, particularly for small schools. This was a factor taken into account when determining the tolerances for deciding differences in results or engagement patterns (see Chapter Five).

3.8 Research approvals

As a PhD candidate, this researcher sought and was granted UTS ethics approval (UTS HREC REF NO. 2015000453) in September 2015 to undertake the research described in this thesis.

An application to the NSW Department of Education to access its state-wide EV and NAPLAN results and to approach schools to participate in research was granted in November 2015 (SERAP 2015373).

CHAPTER FOUR: FINDINGS FROM PHASE TWO

This chapter reports findings from phase two of the research design, the analysis of survey returns from science teachers. The findings provide partial answers to the first two research questions:

1. What use are science teachers making of the EV program including SOLO, and why is it used or not used?
2. What formative practices are evident in the work of science teachers, and why are they used or not used?

Findings in relation to the why or why not components of the questions are provided in Chapter Five.

Section 4.1 reports the size of the groups comprising the sample of schools invited to participate in the research (from phase one of the research design). Also discussed here is the impact of using the regression residual (which is an imputed measure of the scientific literacy attained relative to a predictor) to rank schools instead of EV results. It is relevant to the transformative intent of doing this research as will be discussed further in Chapters Five and Six.

Section 4.2 provides the results and findings from analysis of the survey returns. They are reported in four sets relating to the sections in the survey.

Section 4.3 reports some additional findings that will be referred to in subsequent chapters.

Section 4.4 provides a summary of key findings grouped under the two research questions they provide answers to.

Section 4.5 provides a summary of findings in relation to the second research question.

4.1 Introduction

Phase one in the research design delivered the sample of schools to work with. The regression analysis of EV results over the chosen predictor produced residuals for 394 schools. The schools were then ordered according to their residuals (biggest positive residual at the top). The size of the residual was deemed for the purposes of this thesis (see subsection 3.3.2) to be a measure of the scientific literacy component of EV test results and a measure of the science teaching associated with it.

As shown in Table 4.1 the approximately 20% of schools with the biggest positive residuals were labelled as schools having EV results that were well above expectation (WAE); approximately 20% of schools with the largest negative residuals were labelled as having EV results well below expectation (WBE). A middle group of schools (approximately 20%) straddling the line of best fit line (zero residual) were labelled as having results at expectation (AE). The remaining schools were labelled as 'not defined'. Expectation was defined in terms of the difference between the actual EV result and NAPLAN-based predictor.

Table 4.1
Defining populations from which to invite research participants

Standardised residuals	Residual Rank	Quintile group	Group label	Number of schools
2.68 to 0.56	1—85	TOP	Well above expectation (WAE)	85
0.55 to 0.16	86—166	-	Not defined	81
0.15 to -0.20	167—254	MIDDLE	As expected (AE)	88
-0.21 to -0.56	255—309	-	Not defined	55
-0.57 to -2.50	310—394	BOTTOM	Well below expectation (WBE)	85

Note. A positive residual means that EV results were above expectation. A negative residual means that EV results were below expectation. Expectation is defined as relative to the "line of best fit" for the result pairs used in the regression model.

As will be demonstrated in Chapter Five (see Table 5.1), the three groups of schools are in effect three separate populations defined by the size of their group mean residuals and the fact that when measurement errors are taken into account, there is negligible overlap between the distributions of results associated with the WAE and AE and AE and WBE groups. There is no overlap between the WAE and WBE distributions. This last difference is important because it means that, in terms of statistical convention, findings of statistical significance between the sample means in each group can be generalised to the group population from which that sample was taken.

Also, the differences between WAE and WBE groups mean residuals are as far apart as could be managed within the constraints of the methodology used. The intention was to achieve Flyvbjerg's (2011) pre-condition of maximum difference between the group measure of the key variable (scientific literacy) we are interested in.

In the NSW government education system, schools are classified in a number of ways, including by proximity to major population centres (metropolitan, provincial, rural and remote), by gender (coeducational, boys or girls schools), and by student entry criteria (comprehensive, partially selective entry or fully selective entry). When schools are ranked using conventional measures of achievement, such as EV test results, the fully selective entry schools occupy the top 19 positions and provincial schools perform poorly relative to metropolitan schools. Only 9% of provincial schools were in the top 20% of schools based on EV results.

The use of the residual to rank schools (Table 4.1) produced the following findings. Scientific literacy scores, for the years from 2011 to 2014, were better than expected (a residual above zero) in 53% of the 394 schools meeting criteria for inclusion in the study. Whilst it is arguable that the difference is not statistically significant, the consistency of the slight positive bias over four years is interesting, if not real. When this result is looked at by government school category, 67% of all provincial schools, 68% of fully selective entry schools, and 23% of partially selective entry schools all achieved better than expected EV results (the residual was the four-year average of school residuals).

According to Thomson et al. (2017) approximately 25% of schools in Australia are classified as provincial (the next category after metropolitan, based on their size and distance from major population centres). Assuming this figure is relevant to NSW, around 115 schools would be in that category of school. When we count up the number of provincial schools in the top 20% of schools ranked according to their residual, 56% of the schools there are provincial schools. Also 25% of the schools in the bottom 20% of schools were provincial schools.

Thus, on the basis of residual rankings, provincial schools had more than double their expected presence in the top 20% group and were represented as expected in the bottom 20% group. It was argued in Chapter Three that the residual is a direct measure of the effect of science teaching. The justification for looking at EV results above, at and below expectation and their attribution to school type is provided in the next paragraph.

In Section 3.2 reference was made to the transformative intent of the mixed methods design employed in this research project. The researcher will provide the findings to the schools that participated and to the NSW Department of Education that supported it. If the unconventional measure of teaching success (residual value and polarity) is validated, then the schools really needing help to improve student achievement in and engagement with science can be specifically identified and targeted for support.

Leaving the category of school out of consideration in the first phase of the research, principals of schools with WAE, AE and WBE EV results were invited to support their science teachers' participation in the research. Of the 394 eligible schools, 258 principals received invitations (66% of eligible schools and 55% of all 465 government secondary schools in NSW with Year 8 student enrolments.

Of the 101 surveys returned by science teachers, 35 were from WBE schools and there were 33 each from AE schools and WAE schools. It is not possible to determine the response rate because the number of teachers who received notification about the survey is unknown. In their responses to the online survey, 42 respondents identified themselves and the 36 schools in which they taught. Not all the survey returns were complete and this shows up in the numbers counted for the purpose of statistical analysis.

The survey questions are available as Appendix F and a printout of descriptive statistics of teacher responses is provided as Appendix J.

4.2 Findings from analysis of the science teacher survey returns.

The residual used to create school groups from which to sample contains no information about the characteristics of the teaching experienced by students in the schools that provided responses. Phase two of the research sought to establish the relationship, if any, between the three school groups and the extent to which teachers use EV resources, including SOLO and formative practices in their work. The survey undertaken by all responding teachers was identical. However, their returns were collated according to the group their school had been assigned to.

A series of ANOVAs were performed to establish the strength (in the statistical sense) of any associations between the school group and aspects of assessment-related work done by teachers in those groups. The survey had four parts and analysis of the set of results from each part is reported separately in subsections 4.2.1 to 4.2.4.

Subsection 4.2.1 describes the extent of teacher engagement with and use of EV resources, their understanding of the EV program, and their involvement with it at and beyond school.

Subsection 4.2.2 describes the extent of science teacher engagement with and understanding of SOLO. These two sets of results and related findings detail the extent and depth of the impact of the EV program, including SOLO, on the assessment-related work of the sampled junior secondary science teachers from 2011 to 2014.

The findings in these two subsections are the main inputs for addressing research question one.

Analysis of teacher responses and items in the third section of the survey provided data relevant to characterising teachers' assessment related work in terms of the five dimensions of formative practice. The analysis was also aimed at establishing the generality of the finding from the sample to the group population. The findings from that analysis are reported in Subsection 4.2.3 and were used to inform answers to the second and third research questions.

International research discussed in chapter two shows that better learning outcomes are strongly associated with teacher use of formative practices (see, for example, CERI, 2005). As explained in Chapter One, it was for this reason that syllabus advice supporting the use of assessment for learning (underpinning formative practices) was included in official syllabus documents in NSW. The findings reported in Subsection 4.2.3 are also the basis for discussion in Chapter Six on the extent to which the findings here make a contribution (through replication) to the growing body of international research on the power of formative practices and on learning how to learn.

The fourth set of findings, reported in Subsection 4.2.4, are about the participating teachers and their schools. Findings in this section provide background information used to inform assessment narratives and conclusions reported in Chapters Five and Six respectively.

Subsection 4.2.5 reports other findings from the survey analysis used to contextualise discussion in Chapters Five and Six.

4.2.1 Set one results: Teacher engagement with EV resources (survey questions 1 to 5)

Question one (Q1) items in the teacher survey addressed the scope of recent (past 12 months) teacher engagement with EV results. Teachers responded yes or no to a total of nine items. Items were grouped into the following categories of actions:

- accessing results (items 1a to 1d)
- discussing results with colleagues (items 1e, 1g & 1h)
- discussing results with students (items 1f & 1i).

Question two (Q2) items sought to find out the extent of teacher engagement with and use of EV related activities and resources over the past two years. Teachers responded yes or no to a total of 13 items. Categories of actions were:

- accessing EV resources and materials (items 2a, 2b & 2d)
- using EV resources in the classroom (items 2c & 2g)
- using EV questions and other resources in or as models for school assessments (items 2e, 2f & 2h)
- changing faculty programs (item 2i)
- engaging beyond school in EV related activities (2j to 2m).

Analysis of data from Questions one and two

The hypothesis was that teachers in schools where EV results were deemed to be WAE would make greater use of EV resources than their colleagues in schools where results were deemed as WBE. The decision was made to include AE schools in the testing to assess the consistency with which the measures of teacher activity associated with AE schools was lower than in WAE schools, but higher than in WBE schools. On balance, EV results in AE schools should be below WAE schools and above WBE schools EV results. If this pattern is found, it adds weight to the credibility of the residual as a measure of science teaching effectiveness.

ANOVA proceeds on the assumption of the null hypothesis (that there are no statistically significant differences in the level of EV resource use by teachers in the three groups). Subsection 3.7.2 discussed general considerations relating to the features of data sets and the appropriate choice of tool from the suite of tools available in SPSS. Subsection 3.7.3 particularised that discussion to this project. Consequently, data sets were analysed for normality and homogeneity of variation and appropriate statistical tools chosen to perform ANOVA and related significance testing. Indicative findings from ANOVA were assessed against a significance level

(p) of .05. The decision to accept or reject the null hypothesis was made by reference to the conventional standard.

The descriptive statistics for Q1 & Q2 (combined) are presented in Tables 4.2 and the related means plots in Figure 4.1.

Table 4.2
Descriptive statistics for Q1 & 2 (n = 85)

Result group		$\bar{x}$	s	$\sigma\bar{x}$	n
Q1 & Q2 (/ 22)	WBE	7.63	4.85	.86	32
	AE	11.82	3.98	.75	28
	WAE	11.48	4.55	.91	25
	Total	10.14	4.86	.53	85

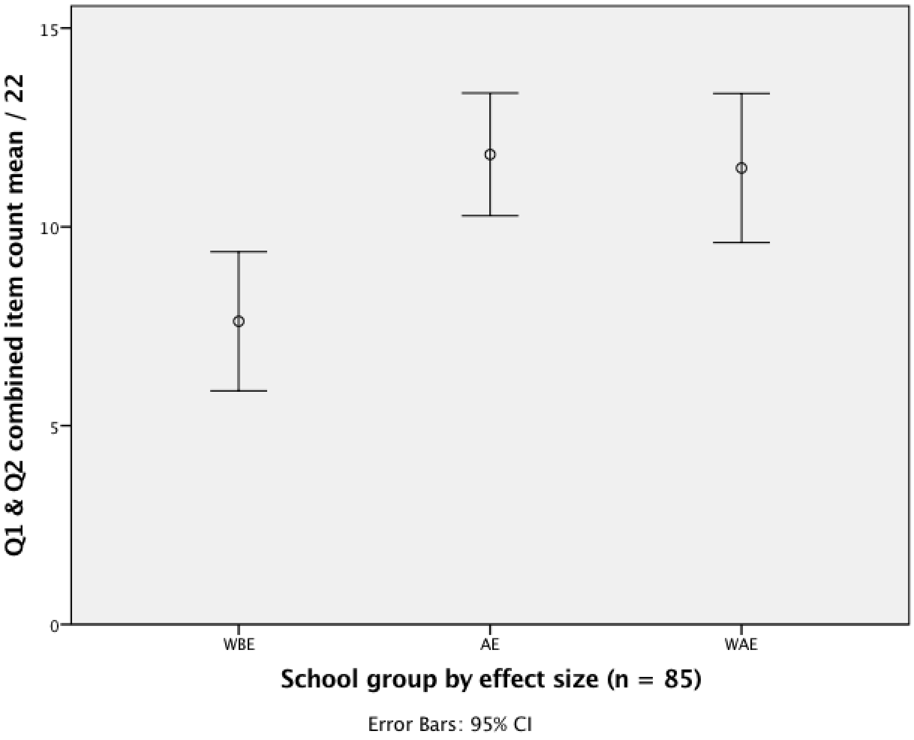


Figure 4.1 Means plots for Q1 & Q 2 combined

The Q1 & Q2 (combined) data sets (n = 85) passed both the normality and homogeneity of variance tests (p > .05). The Shapiro-Wilk statistic (W) for the

WBE group W = .965, p = .38; the AE group W = .982, p = .90, and the WAE group W = .964, p = .49. The Levene variance statistic was $F_{2,82}$ = .821, p = .44.

The parametric ANOVA statistic for Q1 & Q2 combined ($F_{2,82}$ = 8.093, p = .001) supported the rejection of the null hypothesis (p < .05). This means that there was a statistically significant difference between one or more of the groups means.

The Games-Howell multiple comparisons analysis indicated that the $\bar{x}_{WAE}$ - $\bar{x}_{WBE}$ (difference = 3.86, p =.009) and $\bar{x}_{AE}$ – $\bar{x}_{WBE}$ (difference = 4.20, p = .001) were statistically significant, but that the $\bar{x}_{AE}$ – $\bar{x}_{WAE}$ (difference = .34, p = .994) was not.

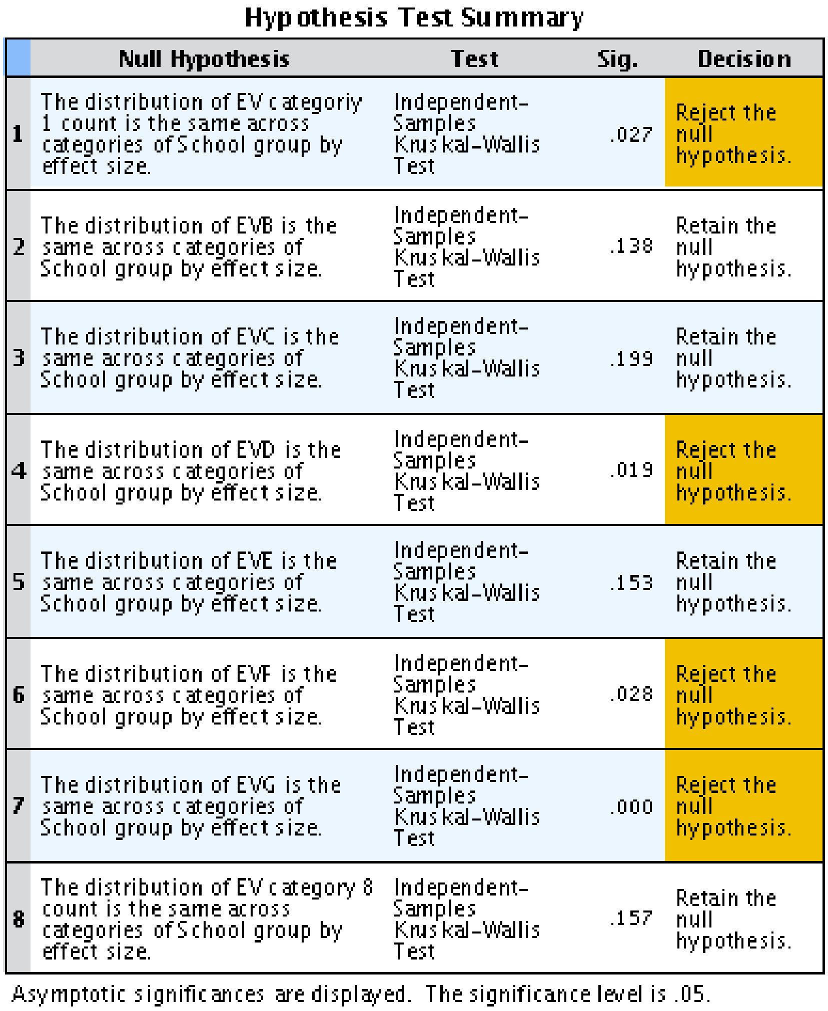
Based on the data analysis for Q1 & Q2 combined, it can be reasonably concluded that, as a group, teachers at schools where results were deemed to be WBE make less use overall of EV results and resources to support their assessment-related work than do their colleagues at schools where results are deemed to be AE or WAE.

A supplementary analysis was then performed on the combined data but this time disaggregated against the eight categories identified above to differentiate particular similarities and differences between group practices.

All eight category-separated data sets failed the Shapiro-Wilk test for normality (p < .05) and all but one (category F) failed the Levene test as well. In the light of that failure, the nonparametric Kruskal-Wallis ANOVA was applied. It demonstrated statistically significant differences between four of the eight category means, as shown in Table 4.3.

Table 4.3

Results of nonparametric ANOVA for eight EV categories



Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of EV categoriy 1 count is the same across categories of School group by effect size.	Independent-Samples Kruskal-Wallis Test	.027	Reject the null hypothesis.
2	The distribution of EVB is the same across categories of School group by effect size.	Independent-Samples Kruskal-Wallis Test	.138	Retain the null hypothesis.
3	The distribution of EVC is the same across categories of School group by effect size.	Independent-Samples Kruskal-Wallis Test	.199	Retain the null hypothesis.
4	The distribution of EVD is the same across categories of School group by effect size.	Independent-Samples Kruskal-Wallis Test	.019	Reject the null hypothesis.
5	The distribution of EVE is the same across categories of School group by effect size.	Independent-Samples Kruskal-Wallis Test	.153	Retain the null hypothesis.
6	The distribution of EVF is the same across categories of School group by effect size.	Independent-Samples Kruskal-Wallis Test	.028	Reject the null hypothesis.
7	The distribution of EVG is the same across categories of School group by effect size.	Independent-Samples Kruskal-Wallis Test	.000	Reject the null hypothesis.
8	The distribution of EV category 8 count is the same across categories of School group by effect size.	Independent-Samples Kruskal-Wallis Test	.157	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Figure 4.2 provides a visual representation of the four categories means that were statistically significantly different. Read vertical bars L to R (matched with EVA to EVG down the RHS labels).

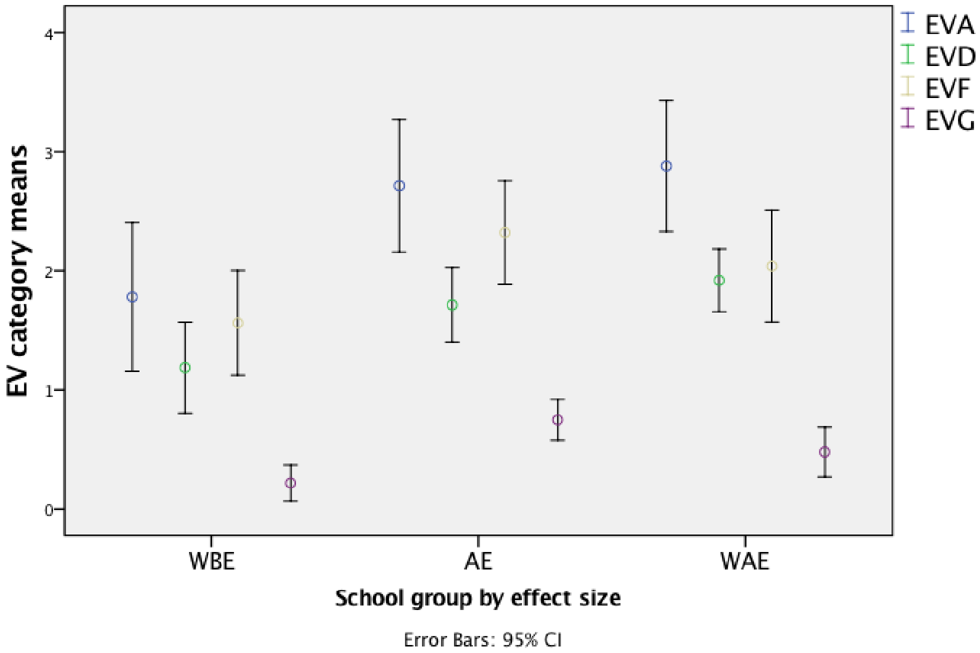


Figure 4.2 EV category means shown to be statistically significantly different

Taking into account the 95% confidence intervals for the means, visual inspection shows that WBE means spreads for categories EVA and EVD (first and second plots from the left) appear to be below the WAE means spreads for the same categories. WBE means spreads for categories EVF and EVG (third and fourth plots) appear to be lower than the AE means spreads for those categories. The AE and WAE means spreads for all four categories appear to overlap each other.

The post hoc Games Howell multiple comparisons tests confirmed that the statistically significant differences in means (p < .05) were, as observed, between the WBE and WAE means for categories EVA and EVD (difference = 1.1, p = .024 and difference = .73, p = .006 respectively) and between the WBE and AE means for categories EVF and EVG (difference = .76, p = .039 and difference = .53, p. = 000 respectively). The Games-Howell means comparison process showed that for the eight data sets, the AE and WAE means were not statistically significantly different.

Based on the above, it would be reasonable to conclude the following about teachers use of resources in the light of the eight categories.

The second (EVB), third (EVC), fifth (EVE) and eighth (EFH) category responses were not statistically significantly different from each other. Thus figures discussed for these four categories of actions are based on the combined total of teachers responding from each of the three groups (n = 85).

In relation to EVB which was about discussing results with colleagues, 66% had discussed the test item and task analysis, 49% had discussed the results of the student survey, and 33% had discussed the student profile information.

EVC was about discussion with students. 22% had discussed the item or task analysis with students and 18% had discussed the results of the student survey.

EVE was about using EV resources in the classroom. 45% had used the teaching strategies provided in the SMART package and 68% had used items and tasks from EV tests in their school assessments.

EVF was about engagement beyond school. Two teachers from the AE group had written items for the EV test; two teachers each from the AE and WAE group had evaluated items for the test; 39% had marked extended response tasks; and 30% had attended workshops about the EV program (different to training for marking).

The following findings can reasonably be made for the four categories where statistically significant differences between teacher use of EV resources were demonstrated.

The first category (EVA) asked teachers to say whether they had, in the previous twelve months, looked at EV results for the student survey (for their class), the analysis of answers to the extended response tasks, and individual student profile results. Teachers in WBE schools had not accessed (viewed) this information as much as their colleagues in WAE schools.

The fourth category (EVD) asked teachers whether they had in the previous two years accessed EV related materials in TaLE (the Department's internal teacher support website), SMART provided feedback on EV results as well as advice about teaching strategies to address science misconceptions and the separately produced marking manuals for extended response tasks. Again, teachers in WBE schools had not accessed these resources as much as their colleagues in WAE schools.

The sixth category (EVF) asked whether teachers in the previous two years had used EV test items and tasks in their own tests or as models to work with. Teachers in WBE schools had done so less than their colleagues in AE schools.

The seventh category (EVG) asked whether schools had used EV results to inform changes to faculty (teaching and learning) programs in the previous two years, Teachers in WBE schools made less use of EV results in that process than had teachers in AE schools.

Survey question three (Q3 or EV3) asked teachers to self-report their level of understanding of the EV program.

The descriptive statistics for the combined data and related plots are shown in Table 4.4 and Figure 4. 3

Table 4.4

Descriptive statistics for Q3 (n = 85)

Result group		$\bar{x}$	s	$\sigma\bar{x}$	n
	WBE	2.97	1.15	.20	32
Q3 (/ 5)	AE	4.04	.79	.15	28
	WAE	3.84	.90	.18	25
	Total	3.58	1.07	.12	85

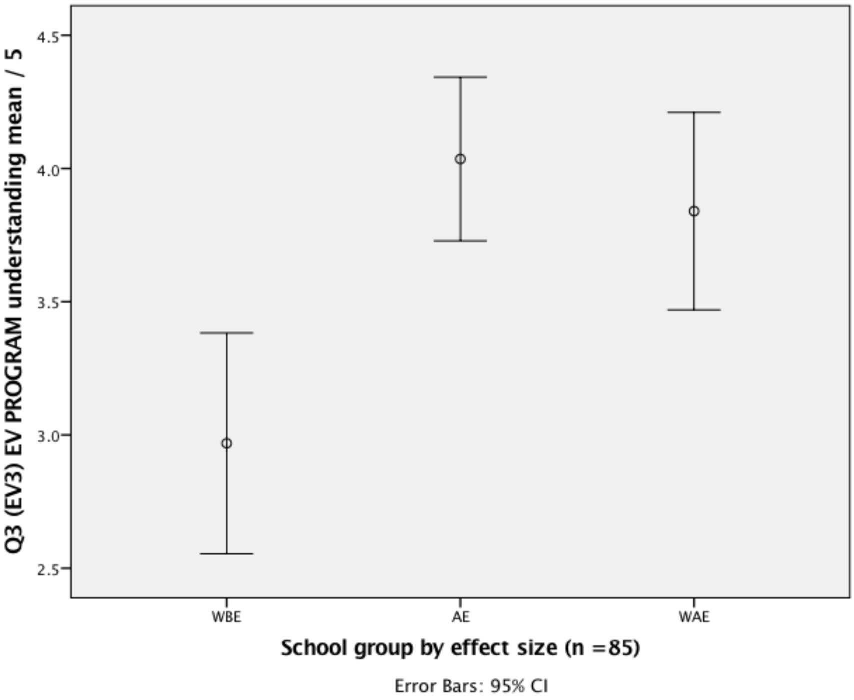


Figure 4.3 Teacher self-rating for their understanding of the EV program (n = 85)

Responses to Q3 were analysed to discover whether teacher rated understanding of the EV program was different between the three groups of schools.

The three data sets for Q3 failed the normality tests ($p < .05$) but did pass the homogeneity of variance tests ($p > .05$).

Given the failure on the normality test, it was decided to apply the Welch robust test of equality of means (Welch's $F_{2, 53.19} = 9.162$, $p = .000$). As the p value was $< .05$, the results were taken as showing a real difference between one or more of the group means.

The Games-Howell multiple comparisons analysis attributed the differences to $\bar{x}_{WAE}$ - $\bar{x}_{WBE}$ (difference = .871, p =.006) and $\bar{x}_{AE}$ – $\bar{x}_{WBE}$ (difference = 1.067, p = .000) which were statistically significant ($p < .05$). The the $\bar{x}_{AE}$ – $\bar{x}_{WAE}$ (difference = .196, p = .682) was not statistically significantly different.

Based on the data analysis for Q3, it can be reasonably concluded that teachers in schools with results deemed to be WBE had a lower self-rated understanding of SOLO than their colleagues in schools where results were deemed to be AE and WAE.

Q4 asked teachers to write what they thought was the most important purpose for the EV test. Table 4.5 shows their collated and categorised responses.

Table 4.5
Summary of EV purposes

Response numbers per group Category of response	WBE n = 32	AE n = 38	WAE n = 25
For students			
Opportunity to demonstrate their learning	3	0	0
Provide students with feedback to improve their learning	0	3	0
Opportunity to improve test taking skills	0	1	2
Provide challenge for higher achievers	0	0	1
For teachers			
Opportunity for professional learning about assessment	0	1	0
Provide feedback on student performance / achievement relative to others	10	7	6
Provide feedback on student performance / achievement relative to standards	1	1 1	2
Provide feedback on student learning	12	3	4
Provide feedback on learning progress	2	7	1
Provide feedback on teaching	3	7	8
Provide feedback on teaching programs	5	5	3
Other responses			
No idea of EV purpose	2	0	0
An unwelcome imposition	1	0	0
Jobs for head office workers	2	0	0
No response or left blank	6	3	4

Note. Some respondents mentioned more than one purpose thus the group sample numbers (n) do not match the comment totals.

Examples of typical responses include:

> *Understand how well our students perform relative to the rest of the state.* (WBE teacher)
>
> *The tracking of students as they progress through high school.* (WBE teacher)

> *Understand your students and amend teaching and learning strategies for students.* (WBE teacher)
>
> *Provide feedback to teachers on the effectiveness of their teaching the stage 4 Science syllabus.* (AE teacher)
>
> *To get a snapshot of how Stage 4 students have progressed specifically in Science since primary school. The extended responses are particularly useful in identifying the students' ability or lack of ability in communicating and/or understanding scientific concepts in different scenarios. It is also very useful to identify misconceptions – so influences our teaching approaches.* (AE teacher)
>
> *Provide feedback to students on their knowledge and understanding of scientific concepts and their scientific literacy. Provide information to teachers on areas that need improvement.* (AE teacher)
>
> *Record of student growth, strengths and weaknesses of programs/areas of teaching.* (WAE teacher)
>
> *To assess students' scientific literacy comparative to their peers in the state.* (WAE teacher)
>
> *Identify areas where we need to improve our teaching of particular concepts or skills.* (WAE teacher).

To summarise, all three groups of teachers most frequently identified the purpose of the EV program as being to provide:

- feedback to teachers about student learning/learning progress
- comparative information about achievement/performance relative to other schools
- feedback about teaching
- feedback on teaching and learning programs.

Survey question five (Q5) asked teachers whether their school was taking up the invitation to participate in VALID10, which is an acronym for Validation of Assessment for Learning and Individual Development. VALID had been introduced on a voluntary basis for Year 10 students for the first time in 2015. It is a new test designed to provide data about achievement in science at the end of Year 10. It is a Year 10 equivalent test to the Year 8 EV test.

Intended participation in VALID 10 in 2016 was lower for WBE schools (n = 3) than either AE (n = 6) or WAE (n = 6) schools. The numbers are based on a count from identified schools in each group to avoid double-counting the same school.

4.2.2 Set two results: SOLO and extent of teacher engagement with it (survey questions 6 to 8)

SOLO is the theoretical model that informs feedback to schools about the level of science thinking exhibited by students as revealed in their selected responses to items and written responses to the extended response tasks (see Chapter Two for a full explanation).

Survey questions six (Q6 a-j) and seven (Q7) were about teacher engagement with and use of SOLO at school and their understanding of SOLO respectively. Q6 a-j asked teachers to respond yes or no to 10 items describing actions taken over the previous two years. The Q6 a-j means for teachers at the schools in each school group at the time of interest are provided in Table 4.6 and Figure 4.4.

Table 4.6
Descriptive statistics for Q6 (n = 85)

Result group		$\bar{x}$	s	$\sigma_{\bar{x}}$	n
Q6 (out of 10)	WBE	2.00	1.87	.330	32
	AE	2.21	2.41	.455	28
	WAE	2.96	3.18	.636	25
	Total	2.35	2.49	.270	85

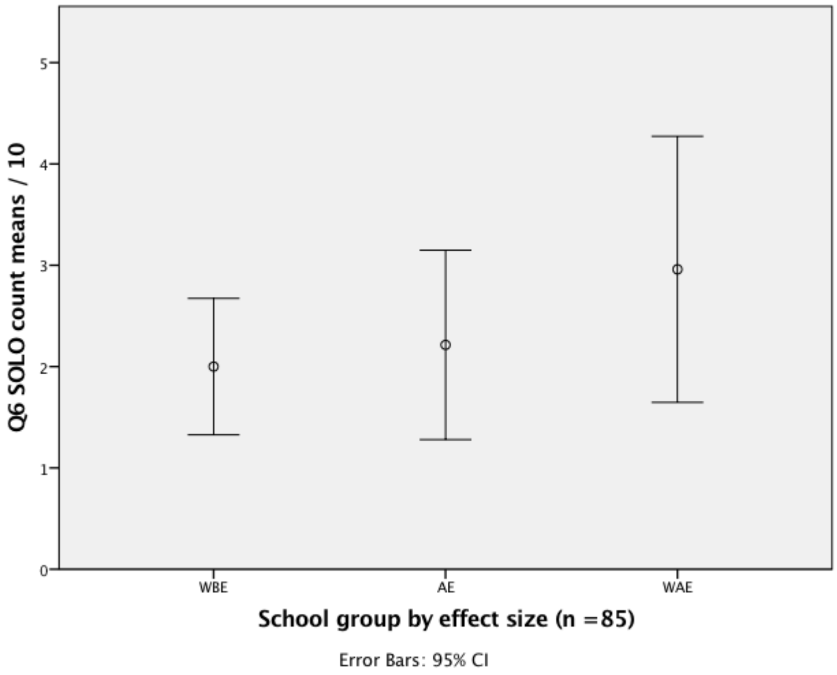


Figure 4.4 Means plots for Q6

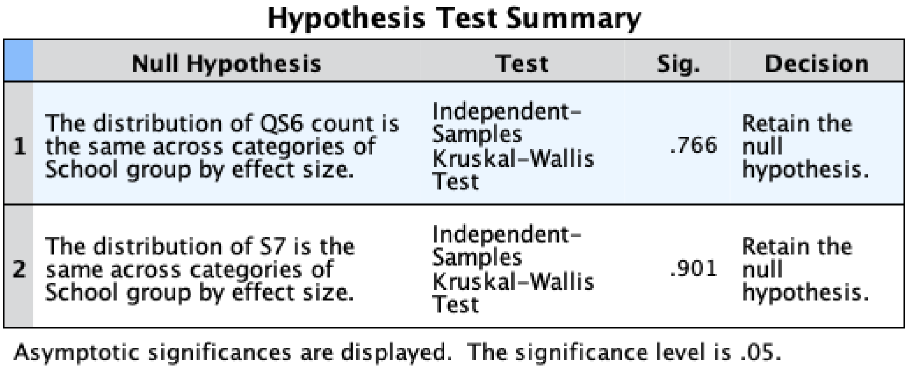
Looking at Q6 a-j means (n = 85), the first observation is that all three group means are low. The second is that when confidence levels are taken into account, the visual representation of the means all overlap and do not appear to be statistically significantly different to each other.

To confirm that result, the Shapiro-Wilks test and Levene tests for data set normality and variance of homogeneity respectively were not satisfied ($p < .05$ for both and thus below the accepted p value of .05) and thus the nonparametric Kruskal-Wallis ANOVA was used.

The nonparametric ANOVA (Table 4.7) includes results for both Q6 and Q7. Row 1 in that table supports the above finding that the means differences between the three categories are not statistically significantly different for Q6.

Table 4.7

Nonparametric ANOVA (n = 85) for SOLO questions (Q6 & 7)



Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of QS6 count is the same across categories of School group by effect size.	Independent-Samples Kruskal-Wallis Test	.766	Retain the null hypothesis.
2	The distribution of S7 is the same across categories of School group by effect size.	Independent-Samples Kruskal-Wallis Test	.901	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Given that the means differences between the samples from the three school groups were not statistically significantly different, total sample responses are provided for Q6 (Table 4.8 and Figure 4.5). The frequencies recorded are for the totals of YES responses to the items in Q6. No teacher scored 9 or 10 out of 10.

Table 4.8

Q6 SOLO category counts (n =85)

Total	Frequency	Percent	Cumulative Percent
0	31	36.5	36.5
1	7	8.2	44.7
2	13	15.3	60.0
3	10	11.8	71.8
4	8	9.4	81.2
5	5	5.9	87.1
6	3	3.5	90.6
7	2	2.4	92.9
8	6	7.1	100.0
	85	100.0	

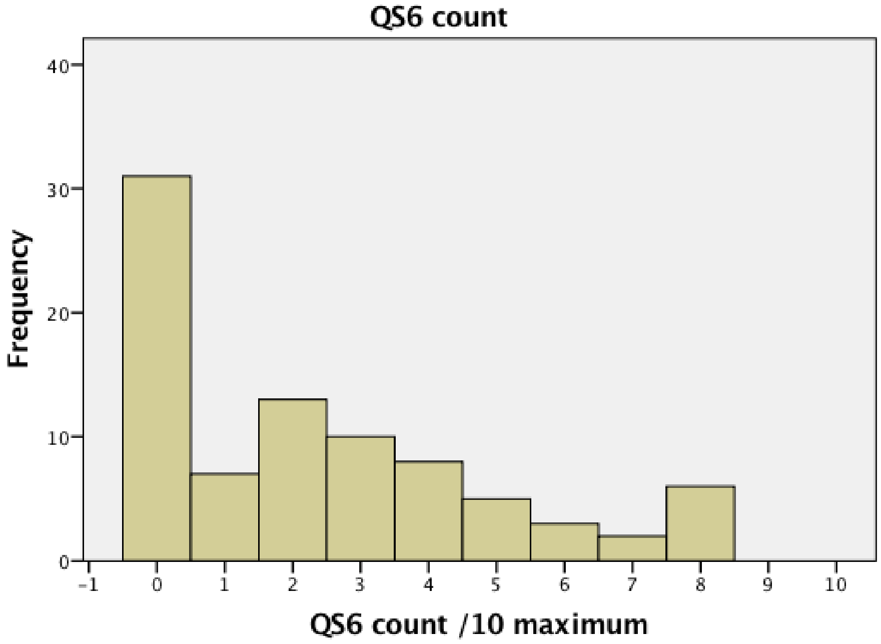


Figure 4.5 Frequency V level of engagement (zero to ten)

In light of the above analysis, when almost a third of the sample said no to all items and with half of the sample responding yes to from one to five of the ten questions, it is reasonable to conclude that in all probability, most teachers across the state have not engaged with SOLO to an extent where it greatly informs their assessment-related work.

Q7 asked for a self-rating by teachers of their understanding of SOLO (1 = very poor to 5 = very good). The means for all three groups are shown in Table 4.9. The level of self-reported understanding ranged from above poor to below acceptable and the means for all three groups responding to Q7 are not statistically significantly different to each other as shown in Figure 4.6 and confirmed above (see row 2, Table 4.7).

Table 4.9

Descriptive statistics for Q7 (n = 84)

Result group		$\overline{x}$	s	$\sigma\overline{x}$	n
Q7 (out of 5)	WBE	2.47	1.11	.196	32
	AE	2.61	1.10	.208	28
	WAE	2.58	1.47	.300	24
	Total	2.55	1.21	.132	84

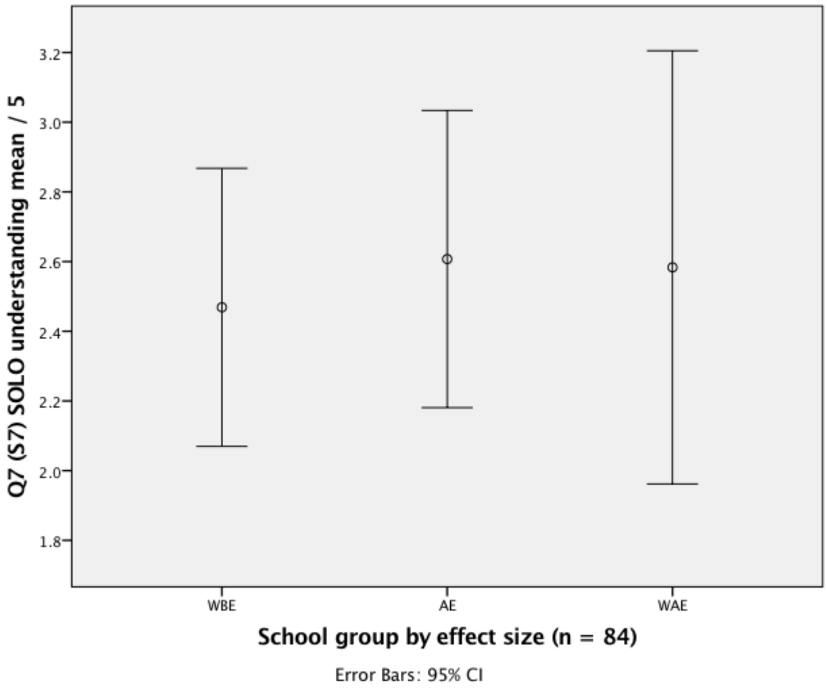


Figure 4.6 Means plots for Q7 self-reported understanding of SOLO

According to Figure 4.7, 45% of the teachers responding to the survey rated their understanding of SOLO as poor or very poor.

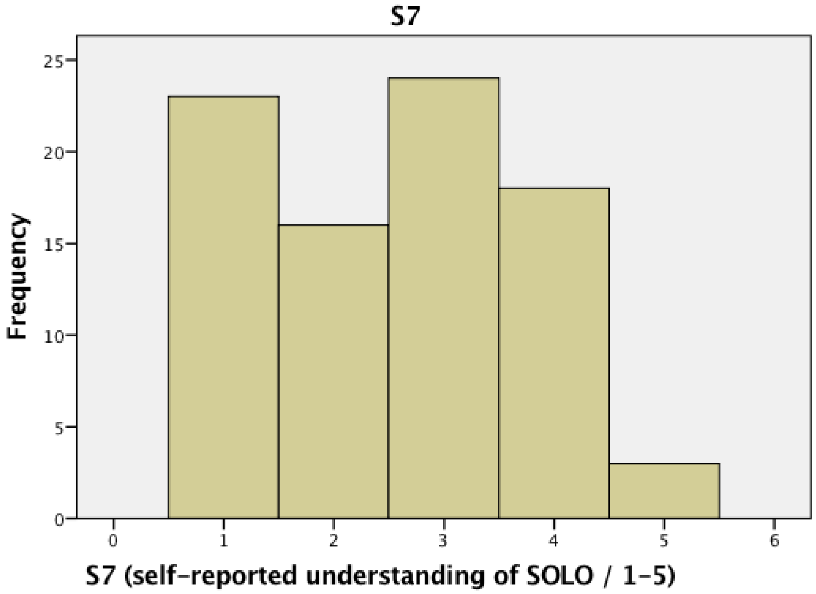


Figure 4.7 S7 Frequency (n = 85) verses level of understanding

The results from the analysis of Q6 and Q7 support the following findings that apply to science teachers in the three school groups sampled (n = 85):

- around 40% of respondents had "accessed material about SOLO" (survey wording)
- fewer than 30% had explained SOLO to anyone or used SOLO in the classroom
- 46% of teachers said they had a very poor or poor understanding of SOLO
- fewer than 10% reported that their school had used SOLO concepts or the SOLO model to inform faculty assessment policies or to provide feedback on student achievement to parents, and; the overall level of self-reported understanding of SOLO ranged from poor to acceptable at best.

Survey question eight (Q8) asked respondents where they learnt most about SOLO. Table 4.10 summarises collated responses from the three samples (WAE, AE and WBE).

Table 4.10

Q8 summary of sources for learning about SOLO

Category of response	WBE n = 32	AE n = 28	WAE n = 25
No response / left blank	9	11	8
Training for ESSA / VALID marking	2	3	4
Actually marking ESSA / VALID	10	4	7
Applying it to school assessment	2	-	2
ESSA / VALID workshop	2	3	3
Using it in class	2	1	-
Researched it	-	7	1
Explaining it to others	1	2	-
Talking to colleagues	-	4	-
Nothing helped	2	-	-
Never heard of SOLO / what is it?	2	-	4

Note. Total responses do not match total sample (n = 85) because some mentioned more than one source. Highlighted responses indicate the sources most commonly identified.

A range of responses from the three groups included:

> *Marking extended response questions for VALID10 this year.* (WBE teacher)
>
> *WTF is SOLO? I've never heard of this. I don't think I spend a huge amount of time under a rock, with my fingers in my ears, crouched in the foetal position whilst humming nursery rhymes, but I have not heard of this term.* (WBE teacher)
>
> *I attended a workshop run by the ESSA people.* (WBE teacher)
>
> *I read about it online to determine what it was. I don't remember it specifically from any training.* (AE teacher)
>
> *Explaining it to other staff.* (AE teacher)

> *Participated in a marking course for what ESSA is and how SOLO marking schemes work.* (AE teacher)
>
> *Marking ESSA.* (WAE teacher)
>
> *I attended the recent Meet the Markers seminar on SOLO and VALID. Our faculty then decided to implement specific SOLO based questions and marking schemes in our half yearly examination for all junior years. This all gave a clear perspective and good practice in the use of SOLO. The outcomes and marking schemes from these examinations have not yet been communicated to students or parents.* (WAE teacher)
>
> *Attended STANSW MTM (Meet the Markers) on ESSA (some years ago and regularly every few years since) and more recently by investigating the work of Pam Hook and others.* (WAE teacher)

Six respondents reported that they had never heard of SOLO or they wanted to know what SOLO was. Marking or training for marking and workshops were the most frequently mentioned sources for learning about SOLO.

By way of explanation, training for the Year 8 test was provided in workshops by a skilled trainer with understanding of SOLO; training for marking the Year 10 tests at the school level involved accessing online materials and may or may not have been done collaboratively with colleagues.

4.2.3 Set three results: Formative practices (Questions 9 to 15)

Questions nine to 15 (Q9 to Q15) sought to capture the extent of use by teachers of assessment for learning strategies/formative practices beyond those associated with the EV program. The EV program is about using assessment data for diagnostic purposes as discussed in earlier chapters. All items in the assessment for learning (AFL) / formative practices section of the survey are available in the survey itself, which is provided as Appendix F.

Q9 to Q15 included 47 separate items. Each item invited one of five responses from teachers: Not known or Unsure about (NKUA) / Never / Seldom / Sometimes / Often.

The analysis for the NKUA option across Q9 to Q15 for the three groups shown is presented in Table 4.11 (descriptive statistics) and their graphical representation in Figure 4.8.

Table 4.11
Means for NKUA option (n = 85)

School group	n	$\bar{x}$	s	$\sigma_{\bar{x}}$
WBE	32	.94	1.46	.26
AE	28	.79	2.06	.39
WAE	25	.72	1.02	.20
Total	85	.82	1.57	.17

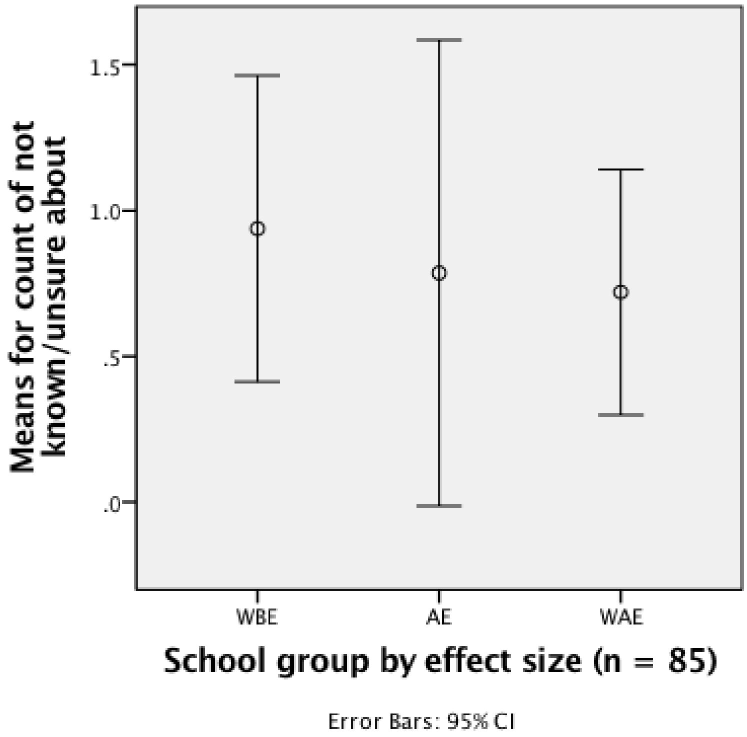


Figure 4.8 NKUA graphical representation of means

The means from each group overlapped when the confidence intervals were taken into account and were thus not statistically significantly different. On that basis the null hypothesis (comparable understanding of the items by teachers in the three groups) was retained.

The next set of data represented in Table 4.12 and Figure 4.9 summarises the data from all respondents (n = 84) for all items in Q9 to Q15. Calculations were based on assigning values to teacher decisions on the following basis. NKUA = 1; never = 2; seldom = 3; sometimes = 4 and often = 5.

Table 4.12
Descriptive stats for Q9 -15 (n = 84)

Result group		$\bar{x}$	s	$\sigma_{\bar{x}}$	n
Q9 - 15 (out of 5)	WBE	3.86	.32	.06	32
	AE	4.07	.41	.08	28
	WAE	4.10	.37	.08	24
	Total	4.00	.38	.04	84

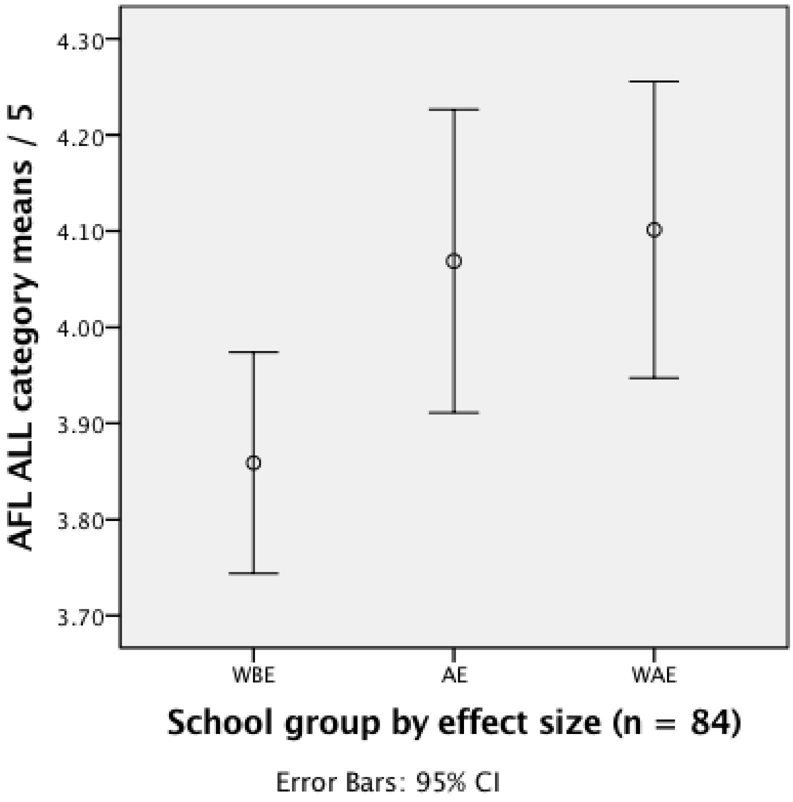


Figure 4.9 Means plots for Q9 – Q15

Visual inspection taking into account the confidence level spread for each group mean strongly suggests that the AE and WAE means were not statistically significantly different. Also, the confidence level spread for the WBE group mean overlaps somewhat the AE and WAE mean spreads.

To test whether some or all of the means were statistically significantly different (or not), the following tests were conducted on the all items data (Qs 9-15). Tests for data normality (Shapiro-Wilk) and homogeneity of variance (Levene) are provided in Table 4.13.

Table 4.13

Tests for normality and homogeneity of variance for all items Qs 9-15 (n = 84)

Shapiro-Wilk test		
	WBE	W_{32} = .956, p = .217
	AE	W_{28} = .978, p = .793
	WAE	W_{24} = .943, p = .191
Levene test		$F_{2,81}$ = .356, p = .702

The results satisfied the thresholds for data normality and homogeneity of variance (p > .05) in all three groups.

Despite there being unequal numbers in the three samples, the parametric ANOVA statistic ($F_{2,81}$ = 3.849, p = .025) and non-parametric Kruskal-Wallis ANOVA statistic ($\chi^2_{(2)}$ = 6.695, p = .035) both returned a significance figure < .05. (for the Kruskal-Wallis result, see row one in Table 4.17).

The Games-Howell multiple comparisons analysis indicated that the $\bar{x}_{WAE}$ - $\bar{x}_{WBE}$ difference (difference = .24, p = .033) was statistically significant (p < .05), but the $\bar{x}_{AE}$ – $\bar{x}_{WBE}$ difference (difference = .21, p = .081) and $\bar{x}_{AE}$ – $\bar{x}_{WAE}$ difference (difference = -.03, p = .950) were not (p > .05).

An on-balance decision was made to reject the null hypothesis based on the results of the above three tests.

A reasonable conclusion was that, in all probability, teachers in schools where results were deemed WBE were less frequent users of formative practices than were their colleagues at schools where results were deemed to be WAE.

As explained in Chapter Two, formative practices were categorized into five dimensions. Survey item returns were subsequently grouped to provide data relating to each of the five dimensions and then disaggregated to identify whether the activity was teacher focused or student focused.

Figure 4.10 represents that organization. It provides a summary descriptor for each of the five dimensions and a unique acronym in parenthesis after it; teacher focused and student focused items related to each dimension are identified and grouped below the descriptor.

1. Clarifying and sharing learning intentions and success criteria (LISC):
   Teacher focus: 9a, 9c & 9e
   Student focus: 9b, 9d & 9f
2. Engineering effective classroom discourse and using learning tasks that elicit evidence of student learning (CDEL):
   Teacher focus: 10a, 10b, 10c & 10f, 10g & 10h
   Student focus: 10d
3. Providing feedback that moves learners forward (FTAL):
   Teacher focus: 9h, 11a – e, 12a – g, 14b & 14e
   Student focus: 14a
4. Activating students as instructional resources for one another (and the teacher) including peer assessment (ASIR):
   Teacher focus: 15a, 15b & 15c
   Student focus: 9g, 10e, 13a, 13b & 13c
5. Activating students (and teachers) as the owners of their own learning including self-assessment (ASTL):
   Teacher focus: 14c, 14d, 14f, 14g, 14h & 15d, 15e
   Student focus: 13d, 13e & 13f

Figure 4.10 Survey questions sorted to show teacher or student as the lead actor

Examples from the survey to illustrate the distinction between teacher and student focus are provided in Table 4.14.

Note that strategies further down the list are about helping students to exercise greater control over their learning (Mitchell et al., 2009). This is relevant to the discussion in Chapter Five about the degree to which self-regulation was evident.

Table 4.14

Sample items from the online survey with a teacher or student focus

Teacher focus	Student focus
Q9c explain to students the indicators or success criteria I will be looking for in their work	Q9d allow students some input in deciding what success criteria are to be applied
Q10h I explain my responses / thinking	Q9f ask students why they think they are being asked to do the proposed activities
Q10f I use test or assignment items and tasks as stimulus for discussion (in class)	Q9g encourage peer feedback based on success criteria
Q11e (provide feedback) advice about how to improve	Q10d ask students to explain their thinking
Q12c (feedback) refers to misconceptions	Q10e use the “think-pair-share-report” strategy
Q14c I evaluate lessons and record ideas for change next time	Q13d (students) self-assess by redoing work to a higher standard
Q14f, g & h access and use information in class…about assessment for learning	Q13e (student self-) selection of items for a portfolio
Q15a collaborate with my science teacher colleagues to develop a shared understanding of what progression in science learning looks like	Q13f self-assess by getting students to keep a journal of their reflections in their own words (on what they have learned in science lessons)
	Q14a students give feedback on my teaching

The descriptive statistics for the separate teacher focused and student focused subsets of items for Qs 9 -15 are provided in Table 4.15. The graphical representations of the teacher focused and student focused means are provided in Figure 4.11 (second and third vertical bars in each group).

Table 4.15

Descriptive statistics TAFL and SFAL (n = 84)

School group		n	x̄	s	σx̄
AFL for teachers	WBE	32	3.84	.35	.06
	AE	28	4.04	.33	.06
	WAE	24	4.05	.33	.07
	Total	84	3.97	.35	.04
AFL for students	WBE	32	3.40	.43	.08
	AE	28	3.56	.64	.12
	WAE	24	3.59	.54	.11
	Total	84	3.51	.54	.06

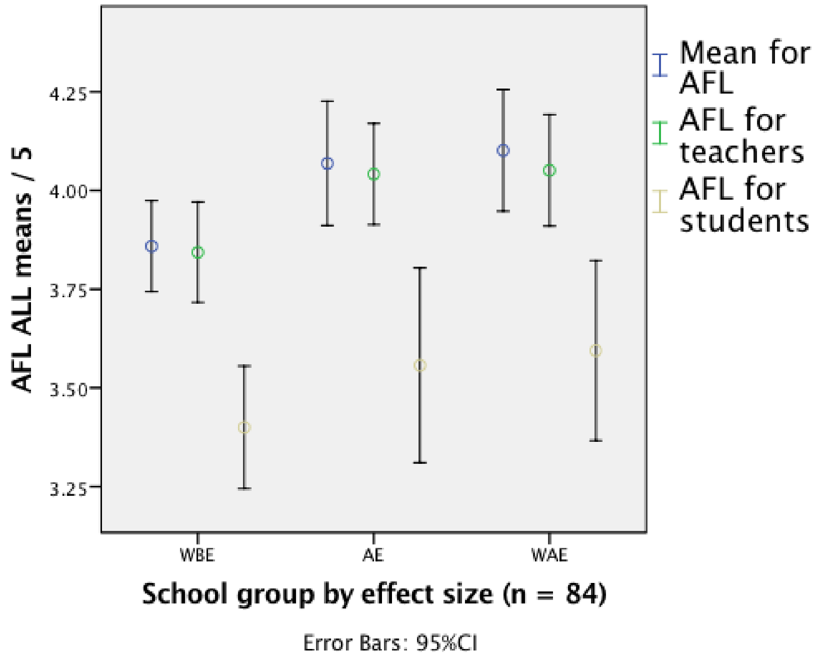


Figure 4.11 Formative practice means for all items, teacher items and student items (n = 84)

Based on the above table and means plots, it would appear that the sample mean for teacher focused items in schools designated as WBE was lower than the sample means for their colleagues in both AE and WAE schools, but that the differences are borderline statistically significant. The means for student focused items did not appear to be statistically significantly different when the confidence level spreads were taken into account.

Normality and homogeneity of variance tests were performed on the data subsets related to teacher focused and student focused items within Qs 9-15. Table 4.16 presents the results of that analysis.

Table 4.16

Tests for normality and homogeneity of variance on assessment for learning (AFL) responses data sets (n = 84)

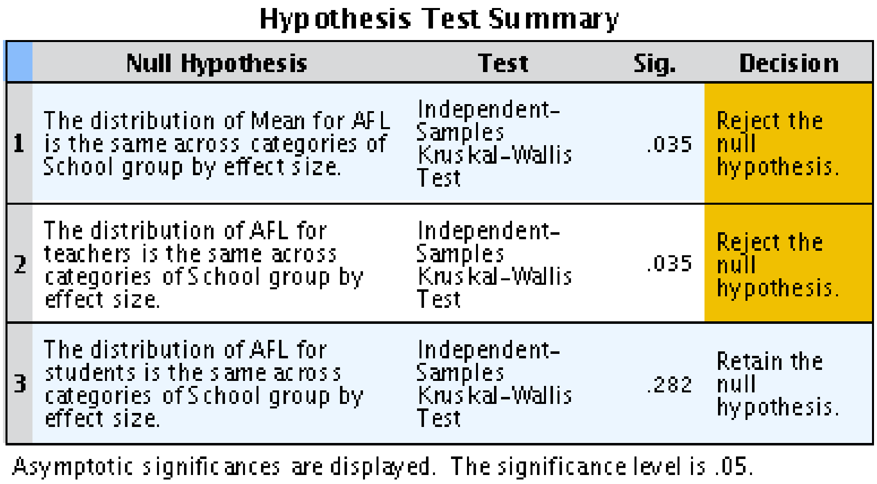
Shapiro-Wilk tests		
AFL for teachers	WBE	W_{32} = .987, p = .963
	AE	W_{28} = .914, p = .025*
	WAE	W_{24} = .961, p = .453
AFL for students	WBE	W_{32} = .917, p = .017*
	AE	W_{28} = .960, p = .353
	WAE	W_{24} = .958, p = .399
Levene test	AFL (teachers)	$F_{2,81}$ = .421, p = .658
	AFL (students)	$F_{2,81}$ = 1.796, p = .173

*sample failed the Shapiro-Wilk normality test (p < .05)

The test results (Table 4.16) did not support the use of parametric tests for comparing means (small and unequal sample numbers in all three groups and in the teacher focused and student focused data sets, one data set in each failed the tests for normality).

Based on that assessment, the nonparametric Kruskal Wallis ANOVA test was applied to the data sets. The results are provided in Table 4.17.

Table 4.17
Nonparametric ANOVA on AFL ALL, AFL teacher and AFL student means (n = 84)



Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Mean for AFL is the same across categories of School group by effect size.	Independent-Samples Kruskal-Wallis Test	.035	Reject the null hypothesis.
2	The distribution of AFL for teachers is the same across categories of School group by effect size.	Independent-Samples Kruskal-Wallis Test	.035	Reject the null hypothesis.
3	The distribution of AFL for students is the same across categories of School group by effect size.	Independent-Samples Kruskal-Wallis Test	.282	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Statistically significant group means differences (p < .05) were found for the all items data ($\chi^2_{(2)}$ = 6.695, p = .035) and the means for the teacher focused items ($\chi^2_{(2)}$ = 6.704, p = .035). There were no statistically significant differences (p >.05) between the means for student focused items ($\chi^2_{(2)}$ = 2.529, p = .282).

Welch robust tests of means equality produced statistically significant results (p < .05) for the all item (Qs 9-15) data (Welch $F_{2,\,50.737}$ = 4.236, p = .020) and the teacher focused data (Welch $F_{2,\,52.620}$ = 3.365, p = .042) but not for the student focus data (Welch $F_{2,\,49.209}$ = 1.283, p = .286) where p > .05.

The Games-Howell multiple comparisons analysis returned a statistically significant difference between the all items (Q9 to Q15) mean for the WBE group of schools ($\bar{x}_{WBE}$ = 3.86) and the all items mean for the WAE group of schools ($\bar{x}_{WAE}$ = 4.10). The means difference was 0.24, p =. 033 which is less than the .05 threshold for statistical significance. No statistically significant differences were shown for the teacher focused items or student focused item means.

On the basis of the parametric and nonparametric ANOVA on the all AFL item data set and subsequent post hoc analysis (Welch test and Games-Howell multiple comparisons tests), it is reasonable reject the null hypothesis and to conclude that teachers in schools where results are deemed WBE make less frequent use of teacher focused formative practices than their colleagues in schools where results were deemed WAE but that no distinction between the groups could be made on the basis of differences in student focus.

Given that there was a statistically significant more frequent use of teacher focused formative practices by WAE teachers than their WBE colleagues, the next step was to test for statistically significant differences between the means for items related to each of the five dimensions of formative practice in each school group.

In order to determine which of the dimensions might present group means that were statistically significantly different, Welch tests for means equality were performed on the five subsets of sample data for each of the dimensions. The results of those tests are presented in Table 4.18.

Table 4.18

Welch statistics for robust equality of means

Dimension	Welch $F_{df1, df2}$	Statistic	Significance
LISC All	$F_{2, 51.823}$	.460	.634
CDEL All	$F_{2, 52.205}$	3.684	.032
FTAL All	$F_{2, 50.494}$	4.522	.016
ASIR All	$F_{2, 51.2.6}$	1.714	.190
ASTL All	$F_{2, 50.650}$	3.475	.039

Shading indicates dimensions where statistically significant means differences were found

The results of the tests on the first and fourth dimensions of formative practice revealed no statistically significant differences between the group means. Statistically significant differences were found between group means for the second, third and fifth dimension. For those dimensions the null hypothsis was rejected and the attribution of those differences is reported below.

4.2.3.1 Learning intentions and success criteria (LISC)

This dimension of formative practice is about learning intentions and success criteria being made explicit by teachers for (or by) students. The items were about who determined what was to be taught and learned and why and how it would be assessed. Means data and plots are provided in Table 4.19 and Figure 4.12 respectively.

Given that there were no statistically significant differences between the group means, only the descriptive statistics for this dimension will be provided here.

Table 4.19

LISC combined means

School group		n	$\bar{x}$	s	$\sigma_{\bar{x}}$
Mean for LISC	WBE	32	4.10	.46	.08
	AE	28	4.20	.42	.08
	WAE	24	4.12	.48	.10
	Total	84	4.14	.45	.05

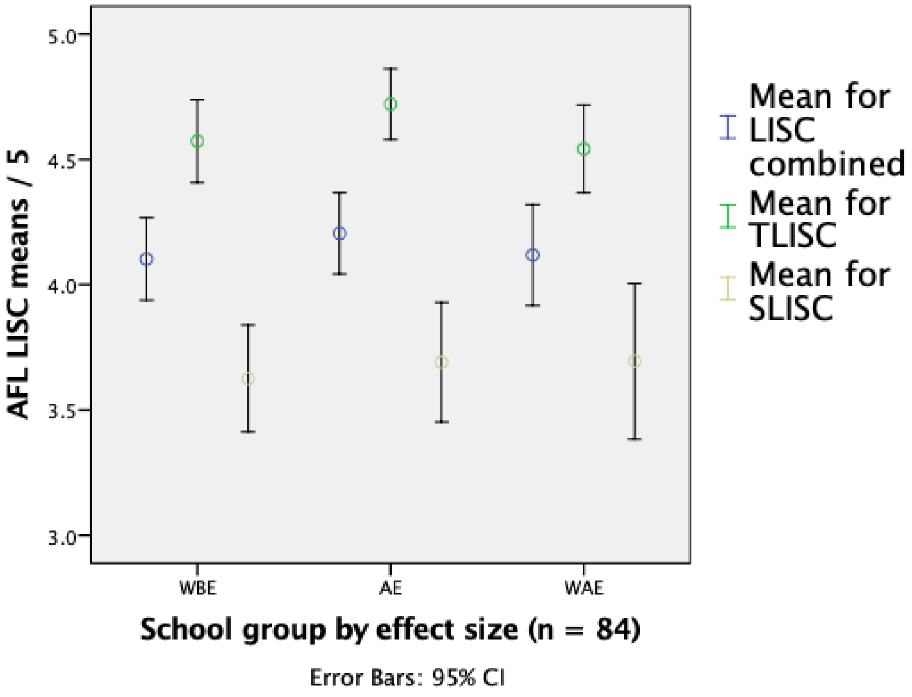


Figure 4.12 LISC means plots

The means spreads for the three group samples show that teacher led activity compared to student opportunities to set learning intentions and choose (or formulate) success criteria do not overlap and are thus statistically significantly different.

Findings from the LISC subsection

From the above it is reasonable to conclude that teachers in all three school groups more often take the lead when it comes to establishing learning intentions and success criteria. They do this at self-reported frequencies between sometimes and often. Teachers report that they involve students between seldom and sometimes in negotiating learning intentions or success criteria.

4.2.3.2 Classroom Discourse that produces evidence of learning (CDEL)

This dimension of formative practice is about classroom discourse eliciting evidence of learning for both the teacher and students. The items associated with this dimension were about questioning and discussion in class and the use of assignments and assessment items as the stimulus for that discussion.

The Welch statistic reported above in Table 4.18 for this second dimension shows there were statistically significant differences between one or more pairs of sample means. The means and mean plots for the teacher and student focused combined and separated data for items related to CDEL are shown in Table 4.20 and Figure 4.13 respectively. This data were examined to see whether the teacher focused (TCDEL) or student focused (SCDEL) data means or both were statistically significantly different.

An examination of the means spreads in Figure 4.13 suggests that the means for both teacher focused and student focused data relating to CDEL in at least the WBE and WAE schools may be statistically significantly different. Subsequent testing for normality and homogeneity of variance in the data is reported in Table 4.21 (note that the student focused data is based on only one item, 10d. That item was about the frequency of opportunity given to students to explain their thinking.

Table 4.20

CDEL combined, TCDEL & SCDEL means

School group		n	$\bar{x}$	s	$\sigma\bar{x}$
Mean for CDEL combined	WBE	31	4.00	.41	.07
	AE	28	4.16	.32	.06
	WAE	24	4.28	.35	.07
	Total	83	4.14	.38	.04
Mean for TCDEL	WBE	31	3.91	.40	.07
	AE	28	4.09	.31	.06
	WAE	24	4.20	.37	.08
	Total	83	4.06	.38	.04
Mean for SCDEL	WBE	31	4.52	.68	.12
	AE	28	4.57	.63	.12
	WAE	24	4.75	.44	.09
	Total	83	4.60	.60	.07

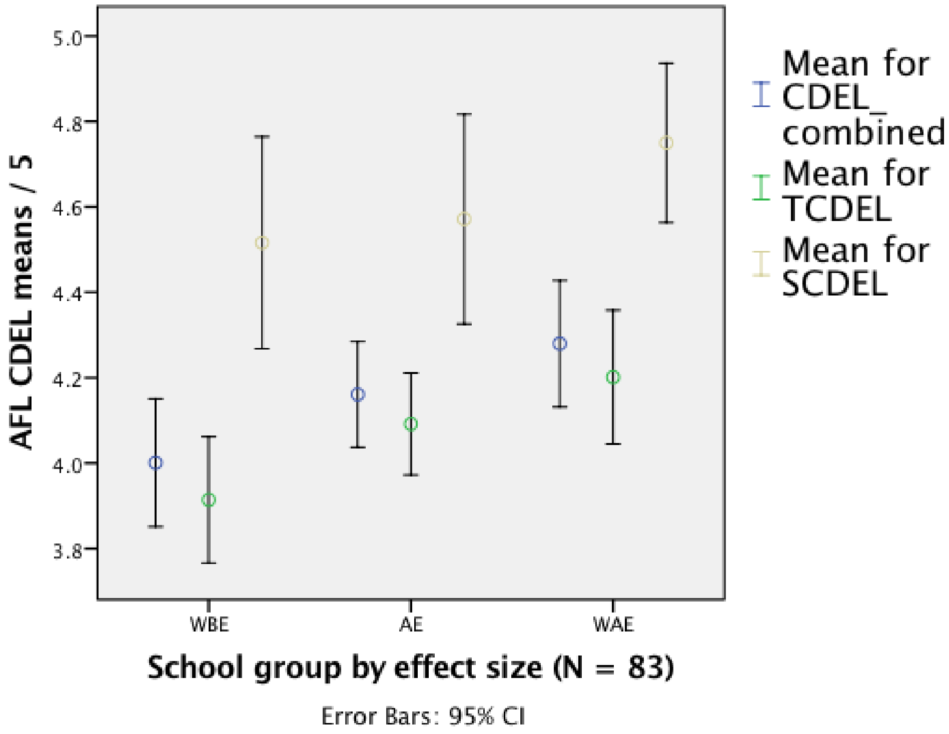


Figure 4.13 CDEL combined, TCDEL, SCDEL means

Table 4.21

Tests for normality and homogeneity of variance on CDEL data sets (n = 84)

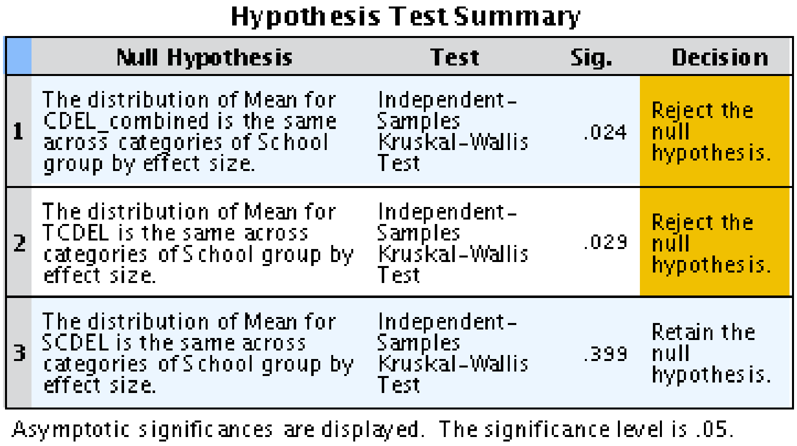
Shapiro-Wilk tests		
CDEL combined	WBE	W = .928, p = .038*
	AE	W = .949, p = .185
	WAE	W = .931, p = .104
CDEL teachers	WBE	W = .936, p = .062
	AE	W = .902, p = .013*
	WAE	W = .944, p = .204
CDEL students	WBE	W = .656, p = .000*
	AE	W = .675, p = .000*
	WAE	W = .542, p = .000*
Levene tests	CDEL (ALL)	$F_{2,81}$ = .011, p = .989
	CDEL (teachers)	$F_{2,81}$ = .123, p = .884
	CDEL (students)	$F_{2,81}$ = .128, p = .053

*sample failed the Shapiro-Wilk normality test (p < .05)

The nonparametric ANOVA (Table 4.22) did show statistically significant differences between at least one pair of means (p < .05) and that that difference was related to the teacher focused data (TCDEL).

Table 4.22

Nonparametric ANOVA: ALLCDEL, CDEL teacher and CDEL student means (n = 84)



Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Mean for CDEL_combined is the same across categories of School group by effect size.	Independent-Samples Kruskal-Wallis Test	.024	Reject the null hypothesis.
2	The distribution of Mean for TCDEL is the same across categories of School group by effect size.	Independent-Samples Kruskal-Wallis Test	.029	Reject the null hypothesis.
3	The distribution of Mean for SCDEL is the same across categories of School group by effect size.	Independent-Samples Kruskal-Wallis Test	.399	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

The Games-Howell multiple comparisons test results (Table 4.23) follow.

Table 4.23

TCDEL & SCDEL Games-Howell multiple comparisons test

						95% CI	
Dependent Variable	(I) School group by ES	(J) School group by ES	Mean Diff (I-J)	SE	Sig.	Lower Bound	Upper Bound
TCDEL	WBE	AE	-.17769	.09295	.145	-.4015	.0462
		WAE	-.28741*	.10480	.022	-.5403	-.0345
	AE	WBE	.17769	.09295	.145	-.0462	.4015
		WAE	-.10972	.09553	.490	-.3413	.1218
	WAE	WBE	.28741*	.10480	.022	.0345	.5403
		AE	.10972	.09553	.490	-.1218	.3413
SCDEL	WBE	AE	-.05530	.17070	.944	-.4661	.3555
		WAE	-.23387	.15142	.279	-.5993	.1315
	AE	WBE	.05530	.17070	.944	-.3555	.4661
		WAE	-.17857	.15004	.465	-.5414	.1843
	WAE	WBE	.23387	.15142	.279	-.1315	.5993
		AE	.17857	.15004	.465	-.1843	.5414

* Grey shading indicates significantly different means

The Games Howell analysis for the teacher focused (TCDEL) data, revealed that the the $\bar{x}_{WAE}$ - $\bar{x}_{WBE}$ pair difference (difference = .29, p = .022) was statistically significant but the $\bar{x}_{AE}$ – $\bar{x}_{WBE}$ pair difference (difference = .18, p = .145) and the $\bar{x}_{WAE}$ - $\bar{x}_{AE}$ pair difference (difference = .11, p = .490) were not. For the student focused (SCDEL) means the test showed no statistically significant difference between the group means.

Findings from the CDEL data analysis

From the above analysis it was reasonable to conclude that teachers in schools where results were deemed to be WBE, compared to their colleagues in schools where results were deemed to be WAE, were more likely to ask closed questions, less likely to use open-ended questions or allow wait-time before answers, or use assignments and assessment tasks as stimulus for discussion. Teachers in the sample of WBE schools were less likely (39%) to ask their students to explain their thinking than their colleagues in WAE schools (75%).

4.2.3.3 Feedback that advances learning (FTAL)

This dimension of formative practice is about feedback that takes learning forward.

The Welch statistic for robust equality of means reported in Table 4.18 for this dimension ($F_{2, 50.494}$ = 4.522, p = .016) indicated that there are statistically significant differences between one or more of the group means. The following analysis will show which of those means pairs are statistically significantly different.

The means and means plots are shown in Table 4.24 and Figure 4.14 respectively. The data for SFTAL is based on one item (14a) which asks how often students are given the opportunity to provide feedback on the teaching they receive.

Table 4.24

FTAL combined, TFTAL & SFTAL means

School group		n	$\bar{x}$	s	$\sigma\bar{x}$
Mean for FTAL combined	WBE	32	3.38	.29	.05
	AE	28	3.59	.36	.07
	WAE	24	3.60	.35	.07
	Total	84	3.51	.34	.04
Mean for TFTAL	WBE	32	3.42	.31	.05
	AE	28	3.64	.39	.07
	WAE	24	3.66	.36	.07
	Total	84	3.56	.37	.04
Mean for SFTAL	WBE	30	3.30	.75	.14
	AE	28	3.82	.82	.16
	WAE	23	4.00	.85	.18
	Total	81	3.68	.85	.09

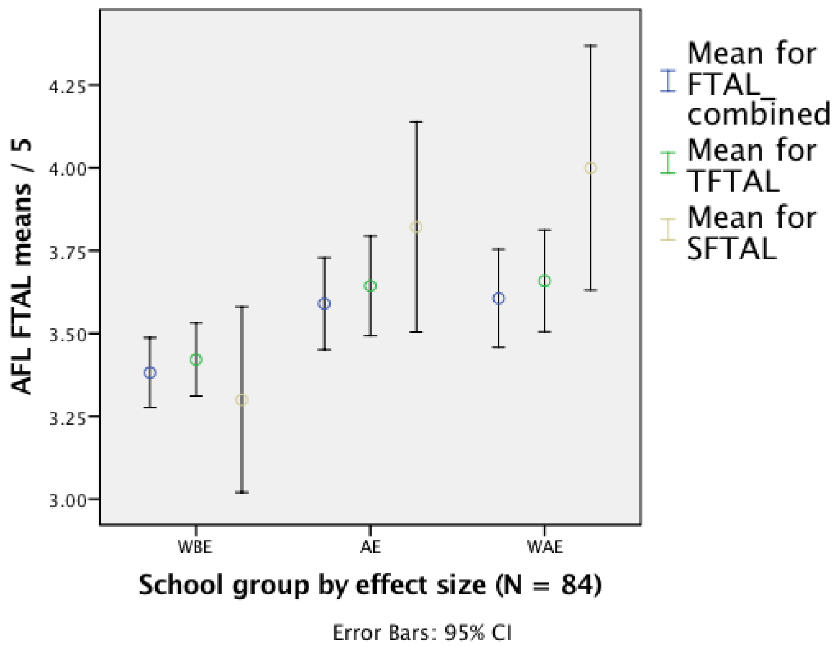


Figure 4.14 FTAL combined, TFTAL, SFTAL means

From observation of the means and related confidence interval spreads in Figure 4.14, it would appear that statistically significant means differences might be found in both the teacher focus and student focus data.

The FTAL teacher data set satisfied the normality tests (p > .05) but the three student data sets all failed (p < .05); all three data sets passed the Levene homogeneity of variance tests (see Table 4.25).

Table 4.25

Tests for normality and homogeneity of variance FTAL responses data sets (n = 84)

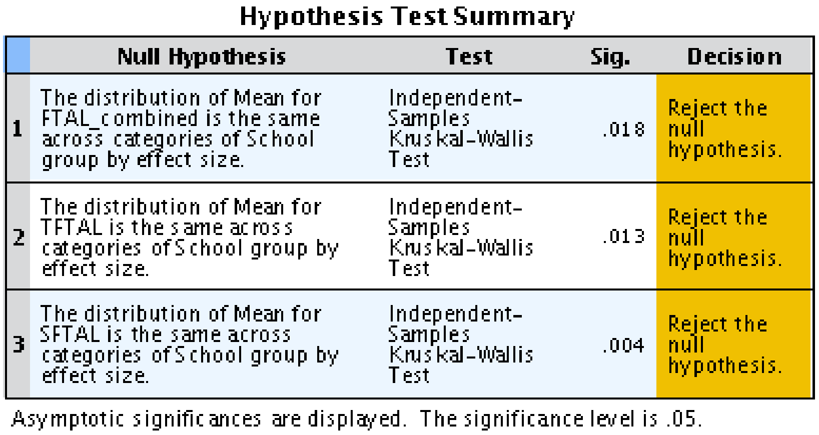
Shapiro-Wilk tests		
FTAL for teachers	WBE	W = .968, p = .439
	AE	W = .948, p = .174
	WAE	W = .953, p = .310
FTAL for students	WBE	W = .830, p = .000*
	AE	W = .848, p = .001*
	WAE	W = .856, p = .003*
Levene tests	FTAL (ALL)	$F_{2,81}$ = .560, p = .574
	FTAL (teachers)	$F_{2,81}$ = .411, p = .664
	FTAL (students)	$F_{2,81}$ = .004, p = .996

*sample failed the Shapiro-Wilk normality test (p < .05)

The nonparametric ANOVA (Table 4.26) indicated that at least one of the pairs of means for both the teacher and student data sets were statistically significantly different ($\chi^2_{\mathrm{TFTAL}\ (2)}$ = 8.713, p = .013 and ($\chi^2_{\mathrm{SFTAL}\ (2)}$ = 11.100, p = .004).

Table 4.26

Nonparametric ANOVA on FTAL ALL, FTAL teacher and FTAL student means (n = 84)



Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Mean for FTAL_combined is the same across categories of School group by effect size.	Independent-Samples Kruskal-Wallis Test	.018	Reject the null hypothesis.
2	The distribution of Mean for TFTAL is the same across categories of School group by effect size.	Independent-Samples Kruskal-Wallis Test	.013	Reject the null hypothesis.
3	The distribution of Mean for SFTAL is the same across categories of School group by effect size.	Independent-Samples Kruskal-Wallis Test	.004	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

The Games-Howell multiple comparisons results are included in Table 4.27.

Table 4.27

TFTAL & SFTAL Games-Howell multiple comparisons (n = 84)

						95% CI	
Dependent Variable	(I) School group by ES	(J) School group by ES	Mean Diff (I-J)	SE	Sig.	Lwr Bound	Upr Bound
TFTAL	WBE	AE	-.22228*	.09138	.048	-.4428	-.0017
		WAE	-.23736*	.09200	.035	-.4604	-.0144
	AE	WBE	.22228*	.09138	.048	.0017	.4428
		WAE	-.01508	.10432	.989	-.2671	.2370
	WAE	WBE	.23736*	.09200	.035	.0144	.4604
		AE	.01508	.10432	.989	-.2370	.2671
SFTAL	WBE	AE	-.52143*	.20661	.038	-1.0192	-.0237
		WAE	-.70000*	.22440	.009	-1.2443	-.1557
	AE	WBE	.52143*	.20661	.038	.0237	1.0192
		WAE	-.17857	.23574	.731	-.7494	.3922
	WAE	WBE	.70000*	.22440	.009	.1557	1.2443
		AE	.17857	.23574	.731	-.3922	.7494

* The grey shading indicates a statistically significant difference

For the TFTAL data, the $\bar{x}_{WAE}$ - $\bar{x}_{WBE}$ pair difference (difference = .24, p = .035) and the $\bar{x}_{AE}$ – $\bar{x}_{WBE}$ pair difference (difference = .22, p = .048) were statistically significant but the $\bar{x}_{WAE}$ - $\bar{x}_{AE}$ pair difference (difference = .02, p = .989) was not.

For the SFTAL data, the $\bar{x}_{WAE}$ - $\bar{x}_{WBE}$ pair difference (difference = .70, p = .009) and the $\bar{x}_{AE}$ – $\bar{x}_{WBE}$ pair difference (difference = .52, p = .038) were statistically significant but the $\bar{x}_{WAE}$ - $\bar{x}_{AE}$ pair difference (difference = .18, p = .731) was not.

Thus, statistically significant means differences were both identified and confirmed.

Findings from the FTAL subsection

From the above analysis teachers at schools where EV results were WBE (compared to their colleagues at WAE and AE schools) were more limited in the range of options used to provide feedback to their students and did so less frequently. WBE teachers were less likely to seek student feedback on their teaching, less responsive to student feedback on their teaching, and less inclined to change the next step in a lesson in response to feedback from students.

It was also appropriate to conclude from this analysis that on the one SFTAL item asking about the opportunity for students to provide feedback about the teaching they experience, teachers in WBE schools were less likely to invite it (closer to seldom than sometimes) compared with their colleagues at WAE schools who said they invited it sometimes.

On one item (Q9h) which asked about the use of digital technology to provide feedback during a lesson, teachers in the three group samples had a similar low response rate, with most saying (53%) they didn't know about it or were unsure about it or never used it for feedback.

4.2.3.4 Activating students as instructional resources (ASIR)

This dimension explores the opportunities that might be provided for students to work collaboratively with peers as a teacher would work with their colleagues.

The Welch statistic for robust equality of means reported in Table 4.18 for this dimension ($F_{2, 51.2.6}$ = 1.714, p = .190) indicated that there are no statistically significant (p > .05) differences between one or more of the group samples means. Only the means and means plots will be provided as shown in Table 4.28 and Figure 4.15 respectively.

Table 4.28

ASIR combined, TASIR & SASIR means

School group		n	$\overline{x}$	s	$\sigma\overline{x}$
Mean for ASIR combined	WBE	32	3.88	.40	.07
	AE	28	4.08	.47	.09
	WAE	24	4.02	.45	.09
	Total	84	3.99	.44	.05
Mean for TASIR	WBE	31	4.63	.50	.09
	AE	28	4.74	.49	.09
	WAE	24	4.60	.49	.10
	Total	83	4.66	.49	.05
Mean for SASIR	WBE	32	3.43	.52	.09
	AE	28	3.68	.69	.13
	WAE	24	3.65	.57	.12
	Total	84	3.58	.60	.07

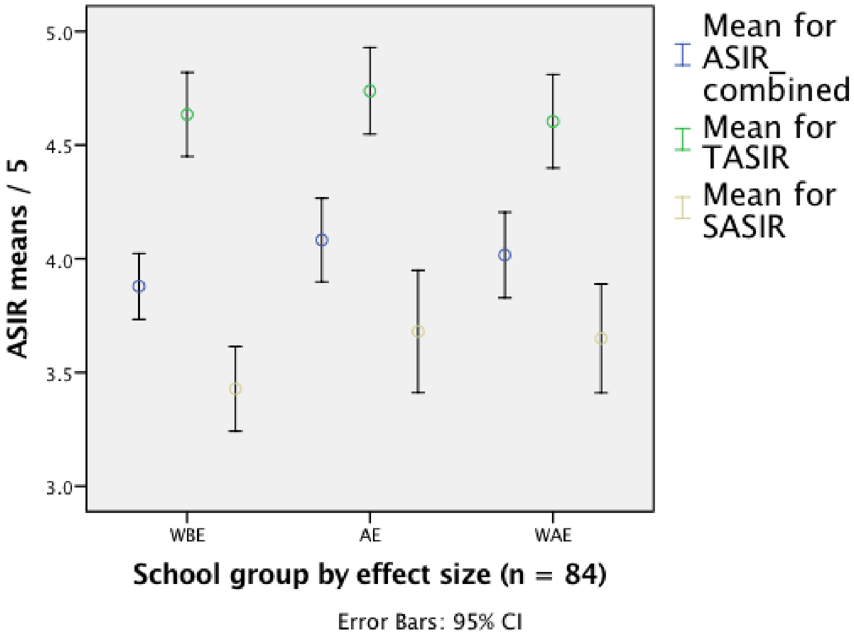


Figure 4.15 ASIR combined, TASIR, SASIR means

Observation of the relative difference between the TASIR and SASIR sample means shows that the differences in each group pair were statistically significantly different.

Findings from the ASIR subsection

The main finding here is that across the three groups of schools combined, teachers in each sample said they work collaboratively more often than sometimes with colleagues on assessment related tasks. However, they only provide their students with opportunities to work collaboratively or provide feedback to each other seldom to sometimes in about equal measure.

4.2.3.5 Activating students (and teachers) as owners of their learning (ASTL)

Items relating to this dimension of formative practices canvass a range of activities for teachers and students designed to promote self-assessment leading to meaningful learning (a fact or concept and its connection/s to other aspects of a particular context that is understood by the learner at the very least).

Table 4.29 and Figure 4.16 provide the descriptive statistics and graphs of the means for this dimension.

Table 4.29

ASTL combined, TASTL & SASTL means

School group		n	$\bar{x}$	s	$\sigma\bar{x}$
Mean for ASTL combined	WBE	31	3.49	.47	.08
	AE	28	3.74	.62	.12
	WAE	24	3.84	.52	.11
	Total	83	3.68	.55	.06
Mean for TASTL	WBE	31	3.66	.52	.09
	AE	28	4.00	.61	.12
	WAE	24	4.07	.60	.12
	Total	83	3.89	.60	.07
Mean for SASTL	WBE	31	3.11	.57	.10
	AE	28	3.16	.91	.17
	WAE	23	3.26	.69	.14
	Total	82	3.17	.73	.08

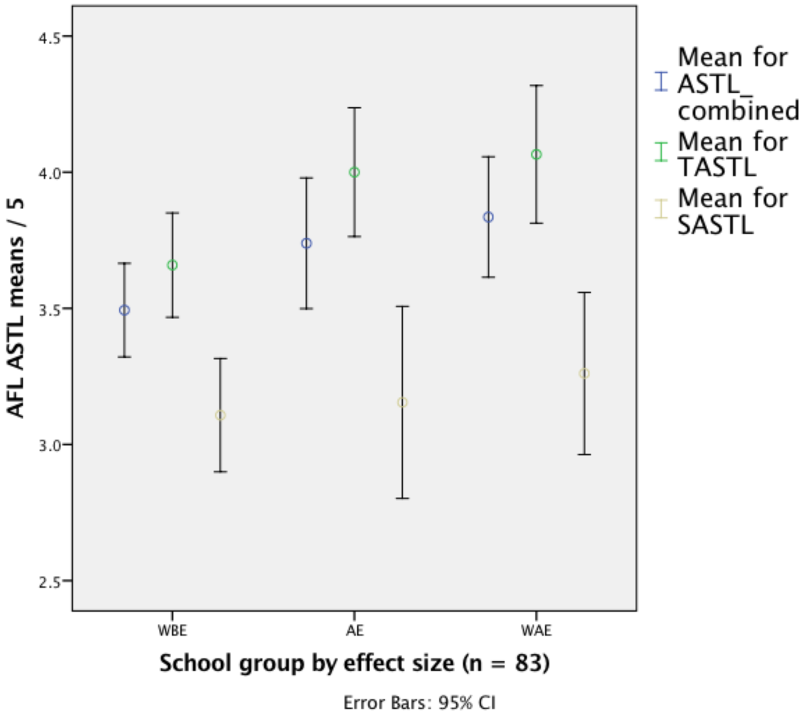


Figure 4.16 ASTL combined, TASTL, SASTL means

Across the three groups collectively, the means for the teacher and student data are statistically significantly different. That said, the data sets were analysed to locate which of the means pairs were statistically significantly different. All data sets passed normality and homogeneity of variance tests (p > .05), as shown in Table 4.30.

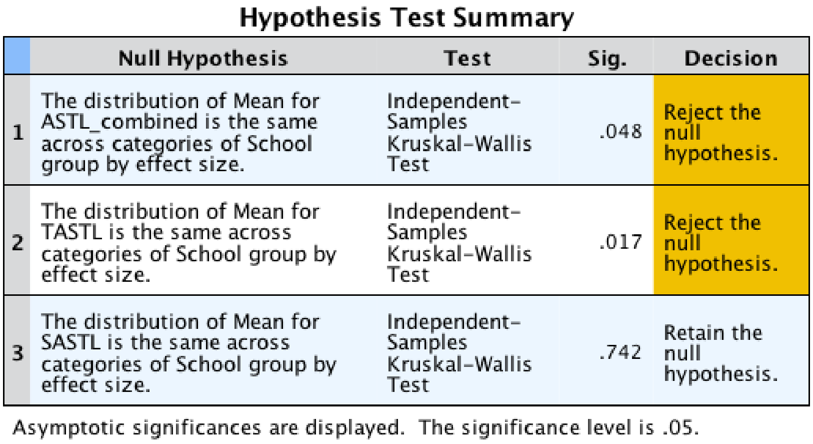
Table 4.30
Tests for normality and homogeneity of variance on ASTL data sets (n = 83)

Shapiro-Wilk tests		
ASTL	WBE	W = .951, p = .169
	AE	W = .954, p = .252
	WAE	W = .948, p = .248
ASTL for teachers	WBE	W = .961, p = .317
	AE	W = .948, p = .172
	WAE	W = .929, p = .093
ASTL for students	WBE	W = .933, p = .055
	AE	W = .951, p = .211
	WAE	W = .961, p = .476
Levene tests	ASTL (ALL)	$F_{2,80}$ = 1.451, p = .240
	ASTL (teachers)	$F_{2,80}$ = .372, p = .690
	ASTL (students)	$F_{2,79}$ = 2.984, p = .056

The nonparametric ANOVA indicates that there were statistically significant differences between at least one pair of the means for the combined scores and that that difference is located with the teacher component (TASTL) as shown in Table 4.31.

Table 4.31

Nonparametric ANOVA on ASTL ALL, ASTL teacher and ASTL student means (n = 83)



Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Mean for ASTL_combined is the same across categories of School group by effect size.	Independent-Samples Kruskal-Wallis Test	.048	Reject the null hypothesis.
2	The distribution of Mean for TASTL is the same across categories of School group by effect size.	Independent-Samples Kruskal-Wallis Test	.017	Reject the null hypothesis.
3	The distribution of Mean for SASTL is the same across categories of School group by effect size.	Independent-Samples Kruskal-Wallis Test	.742	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

The Games-Howell multiple comparisons process (Table 4.32) confirmed that the mean difference between the WAE and WBE for TASTL was statistically significant (difference = .41, p = .030) because p < .05.

Table 4.32

TASTL & SASTL Games-Howell multiple comparisons

						95% CI	
Dependent Variable	(I) School group by ES	(J) School group by ES	Mean Diff (I-J)	SE	Sig.	Lwr Bound	Upr Bound
TASTL	WBE	AE	-.34101	.14862	.065	-.6993	.0173
		WAE	-.40649*	.15409	.030	-.7797	-.0333
	AE	WBE	.34101	.14862	.065	-.0173	.6993
		WAE	-.06548	.16798	.920	-.4715	.3405
	WAE	WBE	.40649*	.15409	.030	.0333	.7797
		AE	.06548	.16798	.920	-.3405	.4715
SASTL	WBE	AE	-.04724	.19976	.970	-.5316	.4371
		WAE	-.15334	.17604	.661	-.5811	.2744
	AE	WBE	.04724	.19976	.970	-.4371	.5316
		WAE	-.10611	.22396	.884	-.6475	.4353
	WAE	WBE	.15334	.17604	.661	-.2744	.5811
		AE	.10611	.22396	.884	-.4353	.6475

* Means are statistically significantly different (p < .05)

Findings from the ASTL dimension

The relevant findings were that teachers in WAE schools, compared to their colleagues in WBE schools, more frequently self-monitor their teaching, use a greater variety of resources to inform their assessment-related work and engage more in professional discussions about syllabus intentions and what is meant by progression in science learning.

All three group samples of teachers indicated they seldom provide students with opportunities to acquire learning how to learn skills such as redoing work to a higher standard, self-selecting items for a portfolio (or explaining their choices for inclusion) or keeping a reflective journal.

4.2.4 Set four results: Respondent Data

Data and information about teachers and their schools were sought in the final section of the online survey. Table 4.33 presents the aggregated data provided by teachers from all three groups of schools.

Table 4.33

Data about respondents and their schools

Question	Response/s			
16. Gender:	**n_F = 54 (63%)***		n_M = 24	
17. Years teaching:	0-5 yrs: n = 10	6-10 yrs: n = 12	11-15 yrs: n = 12	**15+ yrs: n = 44 (56%)**
18. Science teacher by training /qualifications:	**Yes: n = 76 (95%)**		No:n = 4	
Other qualifications:	4 listed, only one not obviously science related			
19. Head teacher:	Yes: n = 39 (48%)		**No: n = 42 (52%)**	
20. Highest science teaching qualification (n =)				
BA + Dip Ed	**55 (70%)**			
BTeach (4 yrs)	12			
MTeach (5 yrs)	7			
Doctorate or PhD	4			
Other	3			
21. Year training completed:				
earliest:	1973			
latest:	2015			
22. Where trained (n =)				
completely overseas:	8			
overseas and in Australia:	5			
completely in Australia:	**65 (76%)**			
23. I teach / have taught Y7-9 classes (n =)				
this year	**69 (87%)**			
last year	2			
the year before last	3			
more than three years ago	5			

Note. Numbers in bold show the **mode**. Because most respondents did not identify themselves or their school, it is not possible to provide a meaningful summary of the figures for Q 24-27 inclusive. Q 24 asked for the number of Y8 classes at your school; Q 25 asked for the number of full time teachers at your school; Q 26 asked for the number of part-time teachers at your school and Q 27 asked about part-time science teachers; it seems that almost all schools had part-time science teachers (from 1-3) in 2016.

* DE employment figures for 2015 show that 61.7% of permanent secondary teachers are male. (**n_F** = number of females; **n_M** = number of males)

A higher proportion of female science teachers responded than males (two to one) even though the proportions of science teachers in Department schools is three males to two females. More than half the respondents were in the most experienced category. Around 1 in 20 science teachers in the sample here have more than the basic qualification to teach science. All but one had a qualification that was mostly science based. Half the respondents were head teachers.

4.3 Other findings

Attention is drawn here to findings that will be referred to in the discussion of answers to the research questions (Chapter Six).

The first is a breakdown of respondents to the survey in terms of teaching experience in each of the three school groups (WAE, AE or WBA).

4.3.1 Teacher experience and student achievement

The proportion of teachers with 15 or more years teaching experience in each of the three groups was: 44% (WBE); 57% (AE) and 56% (WAE). However, an ANOVA to compare the between group means for teaching experience ($F_{2,80}$ = 2.567, p = .083) showed that there were no statistically significant differences (p > .05) when it came to comparing respondent experience.

4.3.2 Teacher use of EV student survey feedback

The survey was designed to provide feedback to teachers about their students' experiences of science at school, including what students thought of the test itself, about science lessons, about science, intentions to study science later in school, which school subjects they liked most (three to choose from of fifteen provided), and which subject they thought they learnt most in (three to choose from of fifteen provided).

Three items in the online survey asked teachers whether in the previous 12 months they had:

- looked at the results from the student survey in the last year (Q1a)
- discussed the results with colleagues (Q1g)
- discussed those results with students (Q1i).

The relevant between groups ANOVA statistic ($F_{2,82}$ = 2.563, p = .083) for the cluster of three items revealed that the between group sample means were not statistically significantly different (p > .05). Thus, descriptive statistics for all survey respondents (n = 85) are presented below in Table 4.34 and Figure 4.17.

Table 4.34

YES counts for student survey items

Total	Frequency	Percent	Cumulative percent
0	26	30.5	30.6
1	19	22.4	52.9
2	26	30.6	83.5
3	14	16.5	100.0
Total	85	100.0	

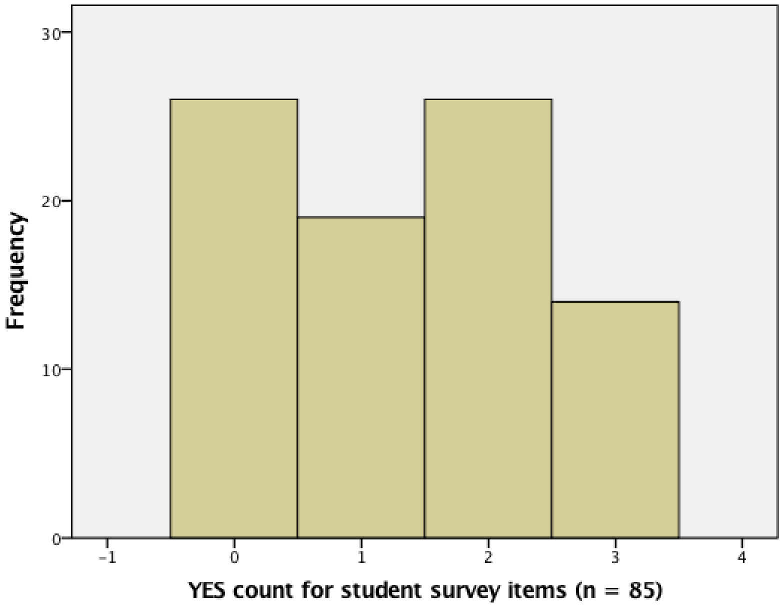


Figure 4.17 Frequency verses item sets for student survey (none to three yes responses)

Just over 30% of teachers had not engaged with the student feedback at all. Fewer than one in five (16%) teachers had looked at and discussed the results with colleagues and students.

4.4 Key findings from the survey analysis

This section summarises the survey findings as they relate to the first two research questions. The survey did not address the issue of why (or why not) teachers made use of the EV program resources. Data and information to answer that part of the two questions is provided in Chapter Five.

Where findings were described as statistically significant the sample findings generalise to the relevant population from which the samples were taken. The expression WAE teachers is shorthand for saying teachers at schools where EV results were WAE (well above expectation). AE or WBE teachers have comparable meanings except that the reference is to the relevant expectation.

> ***Research question one: What use are science teachers making of the EV program including SOLO and why is it used or not used?***

1. Just over 70% of survey respondents had looked at the feedback from the student survey.
2. Teachers at schools where results were deemed to be WBE make less use overall of EV results and resources to support their assessment-related work than do their colleagues at schools where results are deemed to be AE or WAE.
3. There were no statistically significant differences between AE and WAE teachers' engagement with the EV program.
4. In relation to EVB which was about discussing results with colleagues, 66% of the total teacher sample had discussed the test item and task analysis, 49% had discussed the results of the student survey, and 33% had discussed the student profile information.
5. EVC was about discussion with students. 22% of the total sample had discussed the item or task analysis with students and 18% had discussed the results of the survey with students.
6. EVE was about using EV resources in the classroom. 45% of the total sample had used the teaching strategies provided in the SMART package and 68% had used items and tasks from EV tests in their school assessments.

7. EVF was about engagement beyond school. Two teachers from the AE sample had written items for the EV test; two teachers each from the AE and WAE sample had evaluated items for the test; 39% of the total sample had marked extended response tasks; and 30% of the total sample had attended workshops about the EV program (different to training for marking).
8. The first category (EVA) asked teachers to say whether they had, in the previous twelve months, looked at EV results for the student survey (for their class), the analysis of answers to the extended response tasks, and individual student profile results. Teachers in WBE schools had not accessed (viewed) this information as much as their colleagues in WAE schools.
9. The fourth category (EVD) asked teachers whether they had in the previous two years accessed EV related materials in TaLE (the Department's internal teacher support website), SMART provided feedback on EV results as well as advice about teaching strategies to address science misconceptions and the separately produced marking manuals for extended response tasks. Again, teachers in WBE schools had not accessed these resources as much as their colleagues in WAE schools.
10. The sixth category (EVF) asked whether teachers in the previous two years had used EV test items and tasks in their own tests or as models to work with. Teachers in WBE schools had done so less than their colleagues in AE schools.
11. The seventh category (EVG) asked whether schools had used EV results to inform changes to faculty (teaching and learning) programs in the previous two years, Teachers in WBE schools made less use of EV results in that process than had teachers in AE schools.
12. All three groups of teachers rated their understanding of the EV program as acceptable or higher. Teacher self-ratings of EV program understanding in the AE and WAE groups was higher than in the WBE group (good compared to acceptable).
13. Most teachers in the three groups identified that the purpose for the EV program was to provide feedback to teachers about teaching, progress in learning and/or their teaching and learning programs.

14. Fewer WBE schools (three schools) indicated that they would take up the VALID 10 test compared to AE or WAE schools (six schools each).
15. Fewer than 20% of respondents had 'accessed' SOLO; fewer than 10% reported using it to inform faculty policy or as a basis for reporting to parents.
16. The most commonly mentioned source of learning about SOLO was reported by respondents as either EV marking or workshop attendance, and these made up around one third of all responses to the question about where they had learnt most about SOLO.
17. Seven percent of respondents said they had not heard of SOLO.
18. The overall level of self-reported understanding of SOLO by respondents ranged from poor to acceptable.

Research question two: What formative practices are evident in the work of science teachers and why are they used or not used?

19. In relation to the use of formative practices overall, there were statistically significant differences between WBE and WAE teachers. Teachers in WBE schools used formative practices less frequently in their teaching than did their colleagues in WAE schools. Teachers in all three groups more often decided the formative practices to be used rather than share decision making with students on what tasks were to be done and why and how tasks were to be done and assessed.
20. Overall, AE teachers had more in common with their WAE colleagues than WBE colleagues when it came to frequency of use of formative practices.
21. When it came to learning intentions and success criteria (LISC), which was the first dimension of formative practice, teachers in all three samples provided students with the learning intentions and success criteria (between sometimes and often) more than students were asked to identify or choose them (between seldom and sometimes).
22. The second dimension involving classroom discourse eliciting evidence of learning (CDEL) revealed that WBE teachers were more likely to use closed questions; less likely to use open-ended questions; less likely to allow wait-time before answering and less likely to use assignments and assessment tasks

as stimuli for discussion than were their WAE colleagues. Teachers in the WAE sample of schools were most likely to ask students to explain their thinking (more often than sometimes) when compared to either their colleagues in the WBE or AE samples.

23. In relation to feedback (the third dimension of formative practice), WAE teachers compared to WBE teachers were more likely to: use grades linked to syllabus expectations, provide feedback to students addressing misconceptions, refer to success criteria or syllabus intentions and were more responsive to student feedback on their teaching. WAE teachers were more inclined to change the next step in a lesson in response to feedback from students and were the most likely to ask students to provide them with feedback on their teaching.
24. The most frequent response from all three samples of teachers to the item asking about the use of digital technology to monitor learning progress during a lesson was never.
25. In terms of working collaboratively with peers (dimension four) there were no statistically significant differences between practices across the samples of respondents. Teachers collectively said they work collaboratively more often than sometimes with colleagues on assessment related tasks. However, they only provide their students with opportunities to work collaboratively seldom to sometimes in about equal measure.
26. The fifth dimension of formative practice is about taking responsibility for their own learning. WAE teachers model learning-how-to-learn strategies with students and colleagues more frequently (sometimes) than their WBE colleagues (seldom-sometimes equally).
27. Overall, teachers in the three samples indicated they seldom provide students with opportunities to acquire the skills needed to take control of their own learning.

4.5 Summary of findings in relation to science teacher use of formative practices

The analysis of the responses by science teachers to the online survey (phase two) produced statistically significant findings about the use of formative practices and EV results. In schools where EV results were well above expectation (WAE), compared to schools where EV results were well below expectation (WBE), science teachers were more frequent users of activities associated with the following three (of five) dimensions of formative practice:

- discourse eliciting evidence of learning (second dimension)
- the provision of feedback known to progress learning (third dimension)
- the use of and modeling (to peers and students alike) of good learning behaviours, including self-assessment (fifth dimension).

There were no statistically significant differences in the frequency of teacher practices related to the first and fourth dimension of formative practices for sampled teachers in each of the three school groups. As well, the frequency with which teachers engage students in collaborative work with each other and opportunities for peer assessment is comparable across all three samples and less frequent than they do with colleagues.

The next chapter provides additional context for these finding in specific assessment narratives generated for case study schools. It also explores the extent to which case study school data confirm or refute the three predictions made in Section 3.6. The predictions are designed to test two claims that are at the core of this research. The first is that the dual measure of scientific literacy and effect size of teaching vested in the regression residual is valid; and the second is that formative practices are associated strongly with higher achievement and engagement with science later in secondary school years. The confirmation (or otherwise) of the second prediction is an important contribution to answering research question three.

CHAPTER FIVE: PHASE THREE-COMPARING CASE STUDY SCHOOLS

This chapter reports findings with which to answer research question three. This question is

> Does the use of (and if so, how do) formative practices by teachers improve students' EV results and later achievement in and engagement with science?

Achievement data and evidence of engagement with science from five pairs of case study schools provided the basis for findings related to improvement (or otherwise) in Year 8 science results and again at the end of Year 10 (later achievement). Schools were paired on the basis of having comparable SEA scores and statistically significantly different residuals. Comparable SEA scores are scores that are not significantly different in the statistical sense.

Residuals are imputed to be a measure of the impact of science teaching on EV results; the bigger the positive residual, the greater the contribution of science teaching to that EV result in terms of the mark gain (the measure of improvement) above a predicted mark based on NAPLAN results, as explained in Chapter Three.

The bigger the residual difference the better because it improves the chance of identifying what might be causing the differences in those results. These differences, if they exist, are likely to be found in the case study school narratives of assessment-related work provided in Appendix H.

Evidence of engagement was provided in the form of:

- measures of student responses to six items in the EV student survey
- proportions of students completing senior science courses (relative to the state)
- information in case study school narratives about assessment-related work.

The findings related to three predictions linking residual differences to achievement and engagement provide the basis for answering the question. The predictions were:

1. At the end of Year 8 comparable schools with the biggest residuals will have better EV results and engagement figures than schools with smaller or negative residuals.
2. At the end of Year 10 comparable schools with the biggest residuals at the end of Year 8 will have better results than schools with smaller or negative residuals.
3. At the end of Year 12 comparable schools with the biggest residuals at the end of Year 8 will have a higher proportion of their students (relative to English) complete senior science courses than schools with smaller or negative residuals.

Findings related to the first prediction demonstrate the relationship between teacher use of formative practices (indicated by the size and polarity of the residual) and the size of EV result for a school. A highly positive correlation between the residual and EV result for comparable schools would be a strong indication that the use of formative practices was somehow involved.

Findings related to the second prediction may show an ongoing positive correlation between a high positive residual for a school at the end of Year 8 and continuing high achievement in science two years later. This researcher was speculating that later high achievement at this school would be associated with either continuing use by teachers of formative practices or a lasting effect on students from that use.

Findings related to the third prediction may show that later high engagement (Y12 science completions) is positively correlated with either high achievement at the end of Year 8 or high engagement (as measured by scores on the six items from the student survey completed with the EV test) or both. This researcher was speculating that high engagement would be associated with continuing use by teachers of formative practice or be a lasting effect on students from that use.

The lasting effect referred to in the context of predictions two and three is the acquisition of self-regulation and related learning skills by students as a result of their exposure to formative practices. This researcher's assumption in framing the predictions was that more students at WAE schools would become self-regulated, autonomous and skilled learners (of science) as a result of their relatively high exposure to those practices than at AE and WBE schools.

The credibility of this assumption is supported by the research into learning how to learn reported in Chapter Two. Purposely teaching students the five strategies of formative assessment has been demonstrated to produce students who use "good learning behaviours" (Boyle et al., 2001, p. 200). Evidence of the extent to which teachers had directed their efforts to helping students acquire these five skill sets was provided in the results of the teacher survey reported in Chapter Four.

Self-regulated students are also motivated to keep learning. The extent of student liking for their science experience at school is a possible indicator of the extent to which students had acquired the disposition for continued learning in science implied by self-regulation. A measure of student liking for science was available in the scores students returned on the six items of the student survey reported by case study teachers. Anecdotal evidence of student attitudes to science was also provided in the case study school narratives of assessment-related work practices.

The justification for some of the content in this chapter, particularly the identification of specific examples of assessment-related practices associated with successful case study schools, arises from the intention to report the findings to participating teachers and to the Department. A further intention is that the findings be used to support professional learning that leads to greater use of formative practices in science classrooms. This is consistent with the transformative intent of the research as outlined in Chapter Three.

Table 4.1 in Chapter Four showed the 394 participating schools sorted from 1 to 394 on the basis of their residual ranking and subsequent division into five groups. Schools in the top, middle and bottom groups were invited to participate in the research. As was reported there, of the 101 survey returns from teachers, 42 teachers identified themselves and the 36 schools they were working at.

5.1 The case study schools

Table 5.1 reports selected quantitative data for all 36 self-identified schools. That data were sourced from the Department and the *MySchool* website (the SEA score). School identities were protected by replacing the school name with an identifier code. The 16 case study schools engaged with are highlighted in the table.

Table 5.1

Schools that identified themselves including case study schools (shaded)

SCHOOL CODE	n =	SEAS	PEV	RPEV	AEV	RAEV	SR	RSR
PCWAE1	24	2.7	85.40	127	89.95	46	2.68	1
MCWAE1	19	2.8	78.89	374	82.14	286	1.85	3
PCWAE2	44	1.8	81.90	306	84.79	165	1.69	5
PCWAE7	30	2.3	83.19	237	85.81	129	1.59	8
MCWAE2	54	6.9	87.96	68	90.65	41	1.57	10
PCWAE3	55	2.0	81.26	325	83.64	221	1.43	12
MCFSWAE1	106	8.6	99.90	11	101.97	3	1.19	23
MCWAE3	150	6.2	87.45	77	89.47	54	1.17	24
PCWAE4	161	5.5	89.09	56	91.05	37	1.12	29
PCWAE5	49	2.3	82.57	273	84.44	175	1.08	36
MCWAE6	136	6.0	89.26	51	90.50	43	0.73	58
PCWAE6	28	0.9	82.31	289	83.34	235	0.60	78
n = 12								
MGFSAE1	113	9.1	100.76	7	100.23	7	0.12	174
MGAE1	108	3.0	81.65	316	81.86	298	0.12	176
PCAE1	108	3.7	85.40	129	85.55	136	0.08	186
MCAE8	70	2.6	78.62	377	78.82	373	0.06	192
MCAE2	88	3.9	84.94	147	84.85	161	0.03	201
MCAE3	204	3.8	84.30	179	84.28	185	0.01	207
MCAE4	93	2.2	82.19	292	82.16	285	-0.01	213
MCAE5	146	4.1	85.39	128	85.38	141	-0.02	214
MGFSAE2	141	8.3	101.32	5	101.00	6	-0.09	232
MCAE6	89	1.5	79.19	368	79.01	370	-0.01	235
MCAE7	141	2.4	83.42	227	81.91	284	-0.16	244
n = 11								
MBFSWBE2	133	8.2	98.99	14	97.99	17	-0.58	313
MGWBE1	142	7.1	89.60	48	88.34	67	-0.75	330
MCWBE7	153	8.2	91.70	31	90.47	44	-0.76	331
PCWBE2	68	2.1	83.01	248	81.42	316	-0.81	335
MCPSWBE3	123	6.9	92.33	26	90.59	42	-1.03	360
PCWBE6	97	2.9	84.16	184	82.14	287	-1.20	368
MGFSWBE1	135	8.9	101.69	3	99.28	14	-1.42	376
PCWBE1	51	1.7	82.97	253	80.61	340	-1.44	377
MCWBE5	79	2.1	85.09	140	82.54	275	-1.48	378
MCWBE4	47	0.7	76.30	392	73.63	394	-1.58	382
MCWBE3	148	4.0	85.70	118	82.85	256	-1.69	383
MCPSWBE2	144	5.4	90.92	37	87.61	78	-1.91	388
MCPSWBE1	34	6.3	92.93	23	89.63	51	-1.93	389
n = 13								

Note. School code: First letter is (P)rovincial or (M)etropolitan (ACARA defined). Second letter is (C)oeducational, (G)irls or (B)oys. FS = fully selective entry / PS = partially selective entry. Residual group WAE – AE – WBE then a final number to differentiate schools. Columns: n = number of students whose results were used to perform the regression / SEAS = socio-educational advantage score / PEV = predicted EV result / RPEV = rank out of 394 based on predicted EV result / AEV = actual EV result / RAEV = actual EV rank out of 394 / SR = standardised residual used to designate schools as WAE – AE – WBE / RSR = school rank order based on residual (N = 394).

From these codes one can identify the category of school (described in Chapter One). At least one fully selective entry school (FS) was found in each group of schools (WAE, AE & WBE). One of the FS schools was coeducational (C) and the other two were girls (G) schools. Provincial (P) schools were represented in all three groups and there were three in the WAE group. Provincial schools were all coeducational (C) schools. The WBE group included two partially selective entry (PS) coeducational (C) schools as well as one fully selective girls (G) school. There were metropolitan (M) schools in all three groups.

The schools in Table 5.1 are ranked according to standardised residuals (RSR) shown in the far right-hand column. Lines separate WAE from AE and AE from WBE schools. Note that generally speaking actual EV results (AEV) higher than predicted EV results (PEV) are associated with positive school residuals (second column from the left); AEV results lower than PEV results are associated with negative residuals.

The School Profile page for each school on the *MySchool* website shows the proportions of students at that school in four quarters from the most educationally disadvantaged to the most educationally advantaged (L to R). School profile data for Year 7 student entry from 2010 to 2013 was averaged over the four years. As explained in Section 3.6, the profile quarters were converted to a single SEA score (SEAS) using a linear transformation as a further measure to protect the school's identity. The SEA score is an independent measure of the collective learning potential of students at a school. The column headed SEAS shows the four quarters of the socioeducational profile for students at that school as a single score.

The aim was to have among the case studies the six highest-ranked schools from the WAE category, the six schools closest to a zero residual (AE category) and the six lowest-ranked schools (in the WBE category). To this end, teacher-identified schools in each residual category were invited to participate in order of their residual size.

Table 5.2 provides descriptive statistics for the three groups of schools chosen on the basis of their residuals.

Table 5.2

Mean standardised residuals and SEA scores

	Residual means for the three populations	Residual means for self-identified schools	Mean SEA scores for self-identified schools	Residual means for case study schools	Mean SEA scores for case study schools
WAE	$\mu = 1.02$	$\bar{x} = 1.42$	$\bar{x} = 3.86$	$\bar{x} = 1.8$	$\bar{x} = 3.85$
	$\sigma = 0.39$	$s = 0.55$	$s = 2.24$	$s = 0.45$	$s = 2.4$
		$\sigma_{\bar{x}} = 0.16$	$\sigma_{\bar{x}} = 0.65$	$\sigma_{\bar{x}} = 0.19$	$\sigma_{\bar{x}} = 0.94$
	$N = 85$	$n = 12$	$n = 12$	$n = 6$	$n = 6$
AE	$\mu = -0.01$	$\bar{x} = 0.01$	$\bar{x} = 4.06$	$\bar{x} = -0.02$	$\bar{x} = 4.4$
	$\sigma = 0.10$	$s = 0.09$	$s = 2.44$	$s = 0.05$	$s = 2.84$
		$\sigma_{\bar{x}} = 0.03$	$\sigma_{\bar{x}} = 0.74$	$\sigma_{\bar{x}} = 0.03$	$\sigma_{\bar{x}} = 1.42$
	$N = 88$	$n = 11$	$n =11$	$n = 4$	$n = 4$
WBE	$\mu = -1.08$	$\bar{x} = -1.28$	$\bar{x} = 4.96$	$\bar{x} = -1.67$	$\bar{x} = 4.6$
	$\sigma = 0.44$	$s = 0.46$	$s = 2.85$	$s = 0.22$	$s = 3.0$
		$\sigma_{\bar{x}} = 0.13$	$\sigma_{\bar{x}} = 0.79$	$\sigma_{\bar{x}} = 0.09$	$\sigma_{\bar{x}} = 1.21$
	$N = 85$	$n = 13$	$n = 13$	$n = 6$	$n = 6$

Note. μ = population mean / $\bar{x}$ = sample mean / σ = standard deviation (population) / s = standard deviation (sample) / $\sigma_{\bar{x}}$ = standard error (sample) / N = population number / n = sample number

The second column in Table 5.2 (reading left to right) shows the residual means for all the schools in each of the three groups invited to participate. The third column has the residual means for self-identified schools including the case study schools. The fourth column is the SEA score data for the self-identified schools. The fifth column from the right shows the residual means for the case study schools, and the sixth column shows the mean SEA scores for the case study schools.

The residual means for the three school groups (column 1 in the Table 5.2) are separated by almost three standard deviations, which effectively provides three different populations from a statistical perspective (the overlap at the extremes of the residual distributions is approaching one percent or less). This distinction is important, as was shown in Chapter Four when ANOVA findings based on sample data could be applied to all the schools in that population.

Figure 5.1 represents visually the means and related error bars (at the 95% confidence interval) for the data in columns two, four, three and five (reading L to R) in Table 5.2.

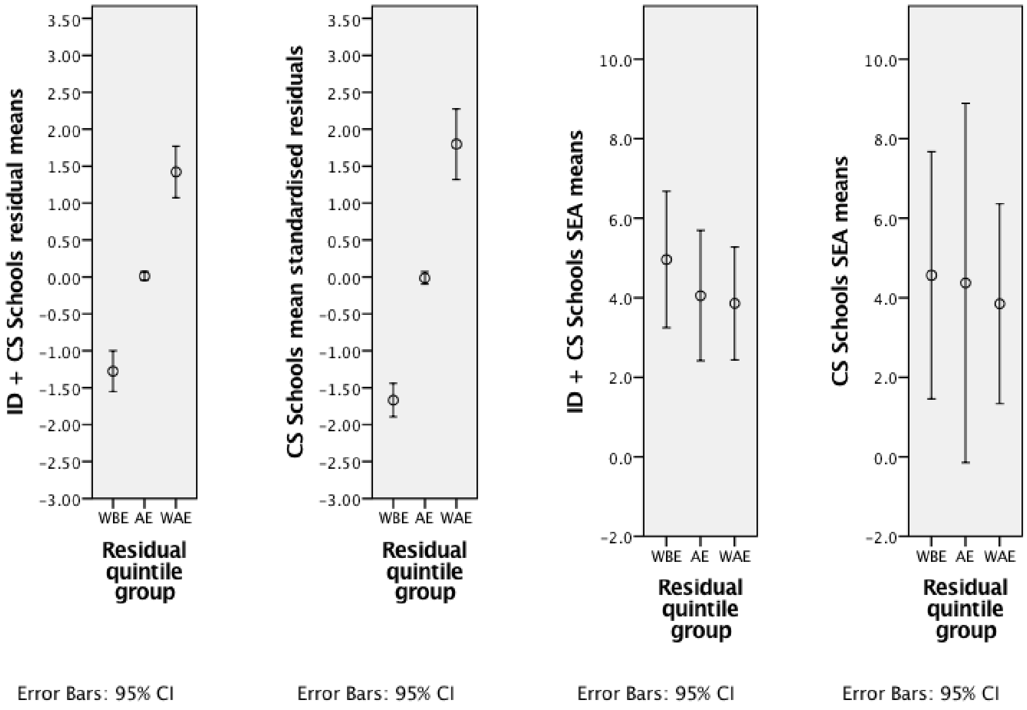


Figure 5.1 Graphical representation of descriptive statistics for identified (ID) and case study (CS) schools combined and case study (CS) schools separately

The data sets for all identified schools (n = 36) from Table 5.1 were tested using SPSS for normality (Shapiro-Wilk test) and homogeneity of variance (Levene tests). Three of the nine data sets (AE schools SEA scores and EV results and WAE schools EV results) failed the normality test ($p < .05$). All three data sets of residuals failed the homogeneity of variance test ($p < .05$), which was not unexpected given the non-random way the schools associated with each group were selected. Correlation results (n = 36) are reported in terms of the nonparametric Spearman coefficient (ρ), degrees of freedom (df) and a two tailed significance of either .01 or .05 (as shown with the reported correlation coefficient).

The correlation between the residuals and actual EV results (ρ = .18, df = 34, p = .283) was slightly positive but not statistically significant (p > .05).

The correlation between the SEA scores and actual EV results (ρ = .84, df = 34, p = .000) was very highly positive and highly statistically significant. The SEA score and residual correlation (ρ = -.08, df = 34, p = .627) was slightly negative and not statistically significant. These two findings were hoped for given that the residual was supposed to show an effect of teaching once student background and school factors had been taken out of the EV result.

An ANOVA performed on the SEA scores and residuals related to each of the three school groups further supported the correlation results. Welch test statistics for the three data sets ($W_{\text{SEAS (2, 12.525)}}$ = .281, p = .759) indicate that the mean SEA scores for each of the three groups were not statistically significantly different.

However, the means for the EV results for the three groups ($W_{\text{EV (2, 13.133)}}$ = 4.98, p = .025) did show at least one statistically significant difference between a pair of the three group means. The Tukey multiple comparisons test (EV results passed the homogeneity of variance test) shows statistically significant means differences between the EV results for the WAE and WBE school groups ($\bar{x}_{\text{wae}}$ - $\bar{x}_{\text{wbe}}$ = 4.94, p = .03). The finding from that testing was that the EV results of WAE schools had a statistically significantly higher mean than the EV results of WBE schools. This was confirmation of a statistically significant association between high EV results and high positive residuals and lower EV results and low negative residuals.

As explained in Section 3.2, the intention was to have the residual means for the WBE and WAE schools as widely separated as possible. This was to provide the best possible chance of finding differences in the teaching associated with the residuals given that the impact of classroom teaching on learning is a relatively small contribution to the accounted for variability in achievement (around 30% according to Hattie (2003b)). The extent to which the residuals represent "maximum variability" (Flyvbjerg, 2011, p. 306) can be seen in the means plots (the first and second plots on the left in Figure 5.1) and the ANOVA results.

The teacher survey analysis (Chapter Four) attributed the residual differences to the frequency with which teachers in each of the three school groups used activities associated with the five dimensions of formative practice. EV results that were well above expectation (WAE) were statistically significantly associated with more frequent use by teachers in WAE schools of activities associated with the second, third and fifth dimensions of formative practice as summarised in Section 4.5.

5.2 Three predictions and the case study schools

This section explains the data about achievement and engagement relevant to the three predictions. Research question three asks how formative practices help improve students results and achievement. The hypothesis was that exposure to formative practices produces self-regulated autonomous learners. As outlined in the opening section of this chapter, the intention was to provide credible evidence that self-regulated autonomous learners are the engineers of their improved achievement and engagement in science.

5.2.1 Prediction one: Year 8 achievement and engagement.

Participating teachers at the case study schools were asked to transcribe results from the Schools Measurement, Assessment and reporting Toolkit (SMART) into a proforma sent well before the school visit. Teachers were asked to bring the completed proforma to the interview when it would be discussed. The proforma is provided as Appendix E. Results are reported in SMART against six, SOLO-related, levels. Schools were asked to aggregate the results into three achievement bands. Levels 5 and 6 were labelled as top band results; levels 3 and 4 were middle-band results and levels 1 and 2 were bottom-band results.

Achievement data is reported in three achievement bands for five result categories: an overall EV result; a knowledge and understanding result; an extended response task result; a working scientifically result; and a communicating scientifically result. For the purpose of this exercise, results from four of the five achievement categories were asked for (the knowledge and understanding category was not provided for on the proforma).

Engagement data were also reported in SMART to teachers against achievement levels. Teachers were asked to record the engagement scores against the three achievement bands in the proforma. The survey had 21 items in it. Only six were chosen for reporting on in the proforma. The items were labelled A to F for the purpose of this analysis.

Students responded to Items A to D by choosing from a four-point scale: strongly disagree, disagree, agree, strongly agree. Individual responses to the survey items were aggregated by school, groups of schools, and the state and reported back to schools as graphs where the scale ranged from -2 to +2. The results for Items A to D are reported on a different scale in this thesis. The effect is to shift the scale so that the lowest possible score is zero. The closer the score is to zero, the stronger the disagreement with the item statement. A score close to four means a strong collected student agreement at the school with the statements. An even mix of agreement and disagreement in the school population would produce a score close to 2.

The statewide responses for Items A to D follow. Table K.2 in Appendix I has the full data set for the case study schools for the six Items.

In relation to Item A, which said: ***I want to study a science subject in Years 11 and 12***, top band students agreed (2.78 out of 4.00), middle and bottom band students disagreed (1.76 and 1.37 respectively).

Item B said: ***Science is the hardest subject I learn***. Top band students disagreed (1.56 out of 4.00), middle band students disagreed also (1.69) but bottom band students agreed – just – that it was the hardest (2.03). Disagreement in response to this item was taken as a positive result.

Item C said: ***In primary school, I enjoyed lessons that were about science***. Top band students agreed (2.76 out of 4.00), middle and lower band students also agreed (2.35 and 2.01 respectively).

Item D said: ***In secondary school, I enjoy science lessons.*** Top band students agreed (2.83 out of 4.00), middle band students also agreed (2.23) but bottom band students disagreed (1.91).

Item E asked students to ***nominate their three favourite subjects*** (15 were listed including science). Of the top band students, 13.5% nominated science in that group, as did 6.65% of middle band students and 4.58% of bottom band students.

Item F asked students to ***nominate the three subjects they thought they learned most in***. Again, 15 options, including science were provided. Of the top band students, 25.13% (one in four) nominated science in that group, as did 16.5% of middle band students (just under one in seven) and 9.71% of bottom band students (about one in ten).

The following generalisations can be made about student responses to the items across the state. The higher the students' achievement band:

- the greater was their agreement with the propositions in Items A, C & D
- the greater was their disagreement with the proposition in Item B that science was the hardest subject they studied
- the greater was the proportion of students nominating science as one of their three options for Items E & F.

5.2.2 Prediction two: Year 10 achievement

The 2011 Year 8 cohort of students provided the 2013 Year 10 results, the 2012 Year 8 cohort provided the 2014 Year 10 results, and the 2013 Year 8 cohort the 2015 Year 10 results. Schools transcribed onto the proforma the grade patterns endorsed by the Board for each year from 2009. Data from Year 10 results were used in conjunction with Year 8 EV results to provide findings in relation to prediction two. Data from 2012 to 2015 was aggregated here for the purpose of interschool comparison.

This researcher's assumption was that the impact of syllabus changes (introduced in 2003) and the introduction of the EV test in 2007 on formative practices would have been institutionalised into school practices by 2011 and continued up until 2014, after which a new syllabus became the basis for EV testing. Correlation statistics reported in Section 5.4 were applied on that assumption.

5.2.3 Prediction three: Year 12 engagement

Prediction three involves the proportions of students at a school completing the Year 11 and 12 (senior) science courses offered at the school. A student could take from one to three of the following four courses, depending on the size of the school and resources available to it: Biology, Chemistry, Earth and Environmental Science, and Physics. Many students traditionally took one or two of these courses. It was very rare for a school to provide students with three science courses in Year 12. A fifth course, Senior Science, was an option for students not wanting to undertake further study in science after school. All five courses were developed by the Board. Some schools offered in the senior years courses in science they had developed and had endorsed by the Board, but none of the case study schools reported offering additional Board-endorsed courses in the years of interest for this project.

Year 12 completions from 2011 to 2015 were provided by schools on the proforma. Only Year 12 completions for 2015 were directly comparable with the Year 8 cohort that sat the EV test in 2011 (Year 8 results were only available in SMART from 2011 to 2014). Nevertheless, data for Year 12 completions for the years 2012 to 2015, inclusive, are provided in the tables and were considered in assessing the degree of support for prediction three. Students choose their subjects for study in Years 11 and 12 half-way through Year 10. In the experience of this researcher, the great majority of students complete the subjects they choose then. As well, given that the EV test had been in place since 2007 for all Year 8 students, any impact of the EV program and teacher use (or not) of formative practices on later engagement would, arguably, have occurred before that. As for

Year 10 results, Year 12 completions from 2012 to 2015 were used in the correlation analyses reported in Section 5.4.

This section identified and described achievement and engagement data provided by teachers in case study schools. The next section will endeavor to show that the schools in which students had the greatest exposure to formative practices (WAE schools) were able to sustain better than expected achievement and higher levels of engagement with science beyond Year 8. The quantitative data were supported where appropriate with qualitative evidence from the assessment-related work narratives to support the credible attribution of self-regulation and autonomy to learners in those schools. The data provided by schools and narrative evidence will be discussed in the context of paired school comparisons reported in the next section, Section 5.3.

5.3 Compared case study schools

Paired schools with the same (or closely matched) SEA scores are argued to have students with equivalent collective learning potentials by virtue of the sociocultural capital they bring to school. The residual is taken to be the measure of the extent to which exposure to formative practices has enhanced students' scientific literacy and produced an EV result that is above (or below) expectation.

The survey results provide a measure of student satisfaction with their school science experience. In Section 5.2.1, the connection between higher achievement and level of satisfaction with their school science experience was established for the case study schools. This satisfaction is attributed to interest in and motivation to continue with learning science and is the benchmark for engagement (as defined for the purposes of this thesis in section 3.5.5) attained at the end of Year 8.

The assumption of self-regulation and learning autonomy is based on differential evidence of later achievement and engagement in comparable schools as explained earlier.

Tables with data about achievement and engagement for each of the schools in the paired comparisons below are sourced from data tables presented in Appendix I. The numbers in those five tables were either transcribed directly from teacher-completed proformas or derived from them as explained in the keys associated with each table in Appendix I.

5.3.1 Pair ONE: PCWAE1 and MCWAE1

PCWAE1 is a relatively small provincial school in the west of the state. Two small Year 7 parallel, ungraded classes are formed each year (around 15 students in each class). Students remain in those classes for the first two years of secondary school. The school had the largest positive residual of all schools in the state. Three science teachers, including the relieving deputy principal and relieving head teacher science, attended the interview, which went for over an hour. The completed proforma was brought to the interview. A selection of assessment artifacts was provided both during and after the interview. The school had engaged with VALID10 and planned to continue doing so.

MCWAE1 is a metropolitan high school to the west of the Sydney CBD. Three Year 7 graded classes are formed each Year using feeder school data. Feeder school data is mostly literacy and numeracy based. Students remain in those classes with few changes until the end of Year 8 at least. Thirty percent of its student intake are from refugee families and some of them have had no formal schooling. Many students attend an Intensive Language Centre before entering secondary schooling at the school. Four teachers, including the head teacher, attended the interview. The teachers had accessed SMART data before the interview and the proforma was completed. Assessment artifacts were provided during the interview and some were forwarded later as well. The school had engaged with VALID10 and planned to continue doing so.

Table 5.3 contains information about achievement and engagement at these two schools. It was compiled from data provided in Appendix I, Tables K.1 & K.3.

Table 5.3

Pair ONE selected statistics

School	Y8 ACH		Y8 ENG		Y10 ACH	
	SCH (%)	STA	ALL / 12	TOP / 16	SCH (%)	STA
PCWAE1 EV = 89.95 ± 0.79 SEAS = 2.7 ± 0.22 RES = 2.68 ± 0.38	T 29 B 2	156 15	10	13	A-B 33 D-E 23	87 88
MCWAE1 EV = 82.14 ± 1.91 SEAS = 2.8 ± 0.46 RES = 1.85 ± 0.48	T 7 B 27	38 20	1	1	A-B 10 D-E 69	27 265

Y8 ACH = the proportion of Year 8 students in the top (T) and bottom (B) achievement bands. SCH (%) = school proportions represented as a percentage. STA = the proportion of students at the school expressed as a ratio (school proportion as a % over the state proportion as a %) relative to the state designated as 100.

Y8 ENG = the rank order of schools based on engagement scores. ALL = all three achievement bands / 12 = the rank out of 12 non-selective schools based on the total survey scores for students at a school (the state figure is counted as a school) / TOP = top achievement band students / 16 = school rank for top band students in the 16 case study schools for which data had been provided (the state figure is counted as a school).

Y10 ACH = the proportion of Year 10 students attaining grades A and B and D and E. SCH (%) = the proportion of students at a school with grades A&B and D&E represented as a percentage. STA = the proportion of students at the school as a ratio (school proportion as a % over the state proportion as a %) relative to the state designated as 100.

YEAR 8 ACHIEVEMENT AND ENGAGEMENT

From Table 5.3 it is clear that at the end of Year 8, compared to the state, the proportion of PCWAE1 students attaining top band EV results was very much higher. MCWAE1, on the other hand, had a lower proportion of its students in the top band compared to both PCWAE1 and the state which is consistent with prediction one in terms of achievement.

When we look at engagement, as measured by responses to the six items in the student survey (Appendix I, Tables K.5A – K.5C), at the end of Year 8, when compared to MCWAE1 students, PCWAE1 students in the context of the 11 non-selective entry case study schools:

- were less enthusiastic about wanting to study science courses in the senior years (ranked 5th; MCWAE1 students ranked 1st);
- found science easier (3rd compared to 11th);
- enjoyed their primary school science less (9th compared to highest);
- enjoyed their secondary science classes less (9th compared to highest);
- fewer had nominated science in their three favourite subjects (10th compared to highest); and,
- fewer had nominated science as one of the three subjects they learnt most in (9th compared to highest).

PCWAE1's highest ranking on any of the items was 3rd for wanting to study science in the senior years (Item A). However, the score on which that ranking was based, was below the state score, as were the other five scores. This was an unexpected result given that a positive school experience in science up to the end of Year 8 was associated with high achievement (reported in Section 5.2.1). This anomaly will be discussed in the summative comments part of this section.

YEAR 10 ACHIEVEMENT

By the end of Year 10 the distribution of results at both PCWAE1 and MCWAE1 had changed when compared to the state. PCWAE1 top band numbers decreased from three to two across the state to nine for every ten across the state. In their highest band results MCWAE1 went from having two students compared to five in the state to one in four. The reduced proportions of students in the top band was far greater for PCWAE1 than MCWAE1.

In the bottom band, PCWAE1 numbers, compared to the state, increased from one in seven to nine for every 10 in the state. MCWAE1 numbers also increased from one in five to more than five to two compared to the state. This result still had PCWAE1 with better overall results in science than MCWAE1 and confirmed prediction two.

Year 12 engagement

Table 5.4 shows the proportions of Year 12 science course completions at both schools. A higher proportion, relative to the state, of PCWAE1 students complete science courses by the end of Year 12 than do students at MCWAE1. These figures confirm prediction three. Student are asked in the middle of Year 10 to nominate courses for the final two years of schooling. Given the low rating by PCWAE1 students of their school science experience at the end of Year 8, the expectation would be that very few students would nominate to do science courses in the last two years of schooling. The apparent contradiction will be discussed in the summative comments part of this section.

Table 5.4
Year 12 science course completions (2013-2015 averages)

School	PCWAE1		MCWAE1	
Subject (state proportion%)	School	State	School	State
Biology (28.5)	40	140	32	112
Chemistry (18)	22	122	12	67
Earth and Environmental Science (2.4)	N/A	N/A	N/A	N/A
Physics (16)	22	138	14	88
Senior Science (10.4)	50	481	21	202

School = proportion of students relative to English at the school (relative to 100)
State = proportions of students at the school (relative to the state set at 100) completing Year 12 courses.

Comparative Summative Comments for Pair ONE (PCWAE1 and MCWAE1)

The following discussion of findings in relation to the predictions for the schools compared here and their contribution to answering research question three draws on the school data provided above and refers to the assessment-related work narratives for the schools in Appendix H.

The expectation from the findings in Section 5.2.1 was that PCWAE1's relatively high results would have been accompanied by a positive view of their school science experience. The assessment-related work narratives (see Appendix H) for the two schools have much in common. Both reveal a group of teachers that give a high priority to helping students recognise the science in their everyday lives and the teachers go out of their way to provide a diversity of experiences for their students, both at school and beyond the school gates, including showing students places where science is the basis for the work being done there. This diversity of experiences is used as the basis for teaching activities that provide evidence of learning (in the form of written and oral reports) as well as the traditional pen-and-paper tests that teachers use to produce formal assessments for the purpose of reporting to parents.

Teachers at both schools are very aware of the limited literacy skills possessed by many of their students and they actively promote the use of appropriate scientific terms in student talk. Whole class discussion is an important strategy and students are encouraged to learn and use the vocabulary of science relevant to the topics being studied at school. Learning intentions and success criteria are prominent in the work they do with students and teachers make use of them to inform feedback to students. Groupwork is encouraged and supported. PCWAE1 appears to provide more opportunities for peer assessment (e.g. feedback to each other on a toy that students make and present to the class) than MCWAE1. Both schools make use of older students to mentor younger students. Teachers meet regularly and collaboratively prepare teaching programs and assessment issues as well as sharing marking.

Given the above, and the absence of school factors negatively impacting on classroom environments (absenteeism is low, student relationships are reported as being good), the difference in student rating of their school experience seems to be related to attitudes to science that students bring to school. Evidence to support this was provided by students in answer to Item C, which asked about their enjoyment of classes in primary school where science was the focus. PCWAE1 students ranked 9th out of 11 here, and MCWAE1 students ranked 1st. The comparable question (Item D) for secondary science classes produced a similar result (9th compared to first). In response to a question asking which three subjects students thought they learnt most in (of 15 provided, including science), PCWAE1 ranked science 9th and MCWAE1 listed it the most.

No other evidence about student or community attitudes to science was purposefully collected in this project. The anecdotal evidence from teachers at PCWAE1 was that students at that school thought the teachers were tough on students and followed up on work set. This response was provided when teachers who had read the survey results before the interview had then asked students about it.

It appears that PCWAE1's relatively high aspiration to do senior science courses expressed at the end of Year 8 (3rd in the ranking, but still below the state's score) did come about. A higher proportion of students at PCWAE1 completed Year 12 science courses than their counterparts at MCWAE1. It may be that the higher take-up of senior science courses at the provincial school was a pragmatic response to the perception of more job opportunities related to science than other subject choices. However, teachers at both schools had been providing that information to students through excursions to places where science was a required qualification for the work observed (medicine, agriculture and universities in the case of PCWAE1). Teachers at MCWAE1 mentioned high parental expectations and support for students to do well at school, including buying science textbooks to support independent work by students on science at home.

5.3.2 Pair TWO: MCAE2 and MCWBE3

MCAE2 is a metropolitan school between Hornsby and Newcastle city. It establishes three or four mixed ability classes for students using feeder primary schools school data. One selective entry class is established for high-achieving students with a particular interest in STEM. To gain entry to that stream students sit an entry test set by the school and/or are invited. Students remain in their classes until the end of Year 8; the special class ends at the end of Year 9 when those students have completed Stage 5. Only the head teacher was at the interview.

Assessment artifacts and the proforma were provided later. The school had engaged with VALID10 and planned to continue doing so.

MCWBE3 is a metropolitan school to the south-west of the Sydney CBD. The school provides for six Year 7 classes using feeder school data. A top stream of two large graded classes is established and a second stream of four ungraded classes. The classes remain largely unchanged until the end of Year 8. Only the head teacher was present at the interview and assessment artifacts and results were sent later. The staff were not prepared to engage with VALID10 at the time of interview.

The two schools have comparable SEA scores but statistically significantly different residuals as shown by the data provided in Table 5.5

Table 5.5

Pair TWO selected statistics

School	Y8 ACH		Y8 ENG		Y10 ACH	
	SCH (%)	STA	ALL / 12	TOP / 16	SCH (%)	STA
MCAE2 EV = 85.45 ± 0.48 SEAS = 3.9 ± 0.30 RES = .03 ± 0.42	T 16 B 7	86 52	8	12	A-B 28 D-E 25	74 96
MCWBE3 EV = 82.85 ± 0.29 SEAS = 4.0 ± 0.25 RES = -1.69 ± 0.13	T 12 B 12	65 89	12	14	nil nil	nil nil

Y8 ACH = the proportion of Year 8 students in the top (T) and bottom (B) achievement bands. SCH (%) = school proportions represented as a percentage. STA = the proportion of students at the school expressed as a ratio (school proportion as a % over the state proportion as a %) relative to the state designated as 100.

Y8 ENG = the rank order of schools based on engagement scores. ALL = all three achievement bands / 12 = the rank out of 12 non-selective schools based on the total survey scores for students at a school (the state figure is counted as a school) / TOP = top achievement band students / 16 = school rank for top band students in the 16 case study schools for which data had been provided (the state figure is counted as a school).

Y10 ACH = the proportion of Year 10 students attaining grades A and B and D and E. SCH (%) = the proportion of students at a school with grades A&B and D&E represented as a percentage. STA = the proportion of students at the school as a ratio (school proportion as a % over the state proportion as a %) relative to the state designated as 100.

Year 8 achievement and engagement

MCAE2 results are positively skewed with a higher proportion of students in the top band compared to the bottom band; MCWBE3 has a higher proportion of its students (relative to the state) in the bottom band. The comparison here confirms prediction one in relation to achievement.

In relation to engagement with science at the end of Year 8, compared to MCWBE3 MCAE2 students were:

- slightly less enthusiastic about taking science in the senior years (8th compared to 7th out of the 11 non-selective schools)
- slightly more likely to disagree that science was the hardest subject they studied (5th compared to 6th)
- liked their primary science classes less (8th compared to 6th)
- liked their secondary science classes more (7th compared to 10th)
- proportionately more likely to include science as one of their three favourite subjects (7th compared to 9th)
- proportionately more likely to include science in the group of three subjects they thought they learnt most in (8th compared to 11th).

These figures show that MCAE2 had a slightly more positive view of their school science experience than students at MCWBE3.

Year 10 achievement

MCWBE3 did not provide any data for Year 10, so no comparison can be made here. As for the first pair of schools, Year 10 results for MCAE2 changed from Year 8 to Year 10. The proportion (relative to the state) of top band students at MCAE2 went from 17 compared to 20 in the state down to three compared to four in the state. The proportion of students in their bottom band went down from one to two in the state to the state figures (almost one for one).

Year 12 engagement

Table 5.6 shows the proportions of students completing Year 12 senior science courses at the two schools. MCAE2 has proportionately more of its students completing Biology, Chemistry and Physics courses compared to MCWBE3. These data confirm prediction three.

Table 5.6

Year 12 science course completions (2013-2015 averages)

School	MCAE2		MCWBE3	
Subject (state proportion%)	School	State	School	State
Biology (28.5)	57	200	21	74
Chemistry (18)	19	106	7	39
Earth and Environmental Science (2.4)	N/A	N/A	N/A	N/A
Physics (16)	10	63	9	56
Senior Science (10.4)	N/A	N/A	N/A	N/A

School = proportion of students relative to English at the school (relative to 100)
State = proportions of students at the school (relative to the state set at 100) completing Year 12 courses.

Comparative Summative Comments for PAIR TWO (MCAE2 and MCWBE3)

Of interest here was the fact that MCAE2 actively promoted itself as a STEM school with a particular interest in the Biosciences. That said, it appears to be succeeding and it performed better in science than its comparable school pair. However, given the special status of science at the school, the differential on engagement with science by students at the two schools is not particularly marked. As well, students at the end of Year 8 at MCAE2 were only slightly less enthusiastic about taking senior science classes than were students at MCWBE3.

5.3.3 Pair THREE: PCWAE2 and MCWBE5

PCWAE2 is a relatively small coeducational regional school in the central-west of the state. The school establishes three Year 7 classes each year from students in their feeder primary schools. The classes are initially ungraded, but after six months students are graded using science assessment results. Classes are reviewed every six months and changes made depending on assessment results. This continues until half way through Year 10. The science head teacher was the main contributor at the interview and had moved from a metropolitan coeducational school to take up that position before 2011. There are four full-time and two part-time science teachers at the school. A full-time laboratory assistant and a part-time agriculture assistant support the work of the science department. One of the science teachers was trained as an agriculture teacher. The head teacher said she had been involved over the years in junior and senior secondary science syllabus consultation processes as well as reviewing items for inclusion in EV tests. Another science teacher who had been at the school for several years joined the interview towards the end. Artifacts of Year 7 and Year 8 assessment-related work were provided and the proforma was completed and forwarded after the interview. The school did the first of the VALID10 tests, but at the time of the interview it was not planning to continue with it.

MCWBE5 is a medium-sized metropolitan coeducational high school. Over the years of interest, it provided from four to five Year 7 classes each year depending on the intake numbers from feeder primary schools. One class is a combined Year 7-8 class that has a gifted and talented student intake of around 15 students each year. Students wanting to enter this class sits an entrance test set by the high school. A second class of high achieving independent learners (identified by their feeder schools) was also established each year. Two or three smaller ungraded classes were then created from the remainder of the intake. These classes are retained mostly unchanged until the end of Year 8. The science head teacher had occupied the position throughout the period of interest and was the only science staff member met with and interviewed at this school. His previous school was a provincial high school in the west of the state. There were six full-time and one part time science teachers at the school. Artifacts of Year 7 and Year 8 assessment-related work were provided and the data proforma was completed and forwarded after the interview. The science department had no plans at the time of interview to take up VALID10. Table 5.7 provides relevant data about achievement at the two schools.

Table 5.7

Pair THREE selected statistics

School	Y8 ACH		Y8 ENG		Y10 ACH	
	SCH (%)	STA	ALL / 12	TOP / 16	SCH (%)	STA
PCWAE2 EV = 84.79 ± 0.31 SEAS = 1.8 ± 0.45 RES = 1.69 ± 0.21	T 12 B 12	65 89	7	11	A-B 17 D-E 37	45 142
MCWBE5 EV = 82.54 ± 0.56 SEAS = 2.1 ± 0.11 RES = -1.48 ± 0.28	T 13 B 18	70 133	3	3	A-B 29 D-E 24	76 92

Y8 ACH = the proportion of Year 8 students in the top (T) and bottom (B) achievement bands. SCH (%) = school proportions represented as a percentage. STA = the proportion of students at the school expressed as a ratio (school proportion as a % over the state proportion as a %) relative to the state designated as 100.

Y8 ENG = the rank order of schools based on engagement scores. ALL = all three achievement bands / 12 = the rank out of 12 non-selective schools based on the total survey scores for students at a school (the state figure is counted as a school) / TOP = top achievement band students / 16 = school rank for top band students in the 16 case study schools for which data had been provided (the state figure is counted as a school).

Y10 ACH = the proportion of Year 10 students attaining grades A and B and D and E. SCH (%) = the proportion of students at a school with grades A&B and D&E represented as a percentage. STA = the proportion of students at the school as a ratio (school proportion as a % over the state proportion as a %) relative to the state designated as 100.

The SEA scores for the two schools were very low, indicating that there were many more socio-educationally disadvantaged students at the two schools than advantaged students. The SEA scores were comparable but their residuals were statistically significantly different.

YEAR 8 ACHIEVEMENT AND ENGAGEMENT

At the end of Year 8, it was clear that PCWAE2 was outperforming MCWBE5 when it came to EV results (see Table 5.7). While MCWBE5 had more students in the top band, it had a much greater proportion of its students in the bottom band than did PCWAE2.

In relation to engagement, of the 11 non-selective case study schools, at the end of Year 8, compared to MCWBE5 students, PCWAE2 students:

- were slightly less wanting to study science in the senior years (ranked 3rd compared to 2nd)
- found science slightly harder (8th compared to 7th)
- liked science at primary school less (7th compared to 2nd)
- liked science classes less in secondary school (6th compared to 3rd)
- had the lowest proportion of students nominating science in their group of three favourite subjects (MCWBE5 ranked 3rd)
- had a lower proportion of their students nominating science as one of the three subjects they thought they learned most in (5th compared to 3rd).

These figures are inconsistent with prediction one and against the pattern discussed in Section 5.2.1 (students at the WAE school should be more positive about their school science experience).

YEAR 10 ACHIEVEMENT

In the two years from Year 8 to Year 10, compared to the state, PCWAE2's proportion of students in the top band had declined and increased in the bottom band. MCWBE5's proportion of top band students had increased and bottom band proportions had decreased. These results were inconsistent with prediction two. However, there is a question mark over the assumption of comparability of the Year 10 results because of the pattern change in grades from 2011 (state-wide exam related) to 2012 (when grades were school determined). The proportion of A+B+C grades in MCWBE5 went from 72% (up to 2011) to 82% (2012 to 2015). In that same time span, PCWAE2's results were effectively unchanged (they went from 62% to 63%).

YEAR 12 ENGAGEMENT

Table 5.8 shows the proportions of students at the two schools who completed science courses at the end of Year 12.

The proportions of HSC science course completions over the three years compared to the state were the same in both schools for Biology. PCWAE2 had less in Physics and Chemistry than MCWBE5. PCWAE2 had more of its students complete Senior Science than MCWBE5. The finding in relation to prediction three for this pair of schools was inconclusive (see the summative comments below). However, given the low SEA scores for both schools, the proportions of students completing senior science courses were above state figures in Biology, well above for Senior Science, but close to state figures in Chemistry and Physics (proportionately more WBE students than WAE students completed Physics and Chemistry).

Table 5.8
Year 12 science course completions (2013-2015 averages)

School	PCWAE2		MCWBE5	
Subject (state proportion)	School	State	School	State
Biology (28.5)	38	133	38	133
Chemistry (18)	16	89	18	100
Earth and Environmental Science (2.4)	N/A	N/A	N/A	N/A
Physics (16)	13	81	17	106
Senior Science (10.4)	30	288	20	191

School = proportion of students relative to English at the school (relative to 100)
State = proportions of students at the school (relative to the state set at 100) completing Year 12 courses.

COMPARATIVE SUMMATIVE COMMENTS FOR PAIR THREE (PCWAE2 AND MCWBE5)

The discussion of findings for the pair of schools compared here and their contribution to answering research question three draws on the school data provided above and refers to the assessment-related work narratives for the schools in Appendix H. The narratives for the two schools reflect their very different priorities for science learning. The narratives for the two schools provide evidence of school-factor differences, particularly in relation to summative assessment.

In the WAE school, the focus was on preparing students to undertake senior science courses, and students class placement was reviewed each semester in the light of assessment performance. The science department's staff were active participants in the school-wide literacy program and provided one period of science per timetabling cycle to it. Science teachers also provided students with specific science vocabulary homework linked to the topics being studied. Science teaching was highly differentiated and sensitive to student literacy needs. Talk comes first, then teacher directed reading (by students to the class), followed by writing.

Considerable laboratory-based practical work is also undertaken by students in the name of learning the skills of working scientifically. What is talked about and written is highly managed by teachers. Whilst other tasks contribute to overall assessment results, there are two formal tests per year. Rubrics for scoring students work were prepared to reflect learning intentions and success criteria described in the Board syllabus. The rubrics were provided to students before, during and after assessment and feedback is provided on the extent to which intentions were met. Students do a major research project each year and practical tests, evidence from which contributes to students' overall assessment in science. The research task was tightly constrained by teachers and a detailed scaffold for the final report was provided.

In the WBE school, the priority was for students to enjoy their school science experiences. The focus for teachers was on providing a diversity of rich science experiences, some arising spontaneously out of student interest, within and beyond the school boundary. At the time of interest for this project, there did not seem to be a strong emphasis on using literacy strategies in science. Assessment was likely to be negotiated with students, peer assessment was used to provide feedback on one of the tasks (a model-making exercise), and there was an opportunity for self-assessment at the end of each topic. Evidence of learning was collected from a variety of tasks and there was a good deal of individual teacher judgment involved when it came to preparing reports for parents (and students). Summative assessment was a low-key affair (deliberately) and students were not shifted around on the basis of results until the end of Year 8.

The overall negative impression recorded by PCWAE2 students is inconsistent with the engagement aspect of prediction one. The learning programs at both schools encourage the use of contexts to support teaching and learning. At PCWAE2, mention was made of agriculture and biotechnology as contexts mostly used. The WAE school has a high-stakes, summative assessment approach which has been shown in the research literature (Harlen & Deakin-Crick, 2002) to impact negatively on the motivation to learn of students with poor learning histories. As the WAE school here is a provincial school, the possibility of student socio-cultural factors (similar to those operating in PCWAE1 above) impacting the engagement scores should not be overlooked.

By Year 10, the achievement pattern relative to state figures at the WAE school is below that of the WBE school, and completions of Senior Science courses two years after that were not too dissimilar at the two schools. The achievement findings at the end of Year 10 are inconsistent with prediction two (issues with comparing Year 10 results not withstanding). However, the findings in relation to prediction three are inconclusive. Overall, a higher proportion of PCWAE2 students complete science courses, but a smaller proportion complete the two most demanding courses, Chemistry and Physics.

Possible explanations for the unexpected findings in relation to the predictions will be discussed in the summary section of this chapter (Section 5.5).

5.3.4 Pair FOUR: MGFSAE2 and MGFSWBE1

Three fully selective schools were included for case study. They were MGFSWAE1, MGFSAE2 and MGFSWBE1. All three were metropolitan; the first was a coeducational school; the latter two being girls' schools. The two girls' schools were the focus for paired comparison in this section. However commentary and comparisons were made involving all three schools as considered useful to understanding similarities and differences relevant to the predictions being tested.

The head teachers at the three schools of interest in this section were at those schools at the time of interest for this project (2011–2014). All three schools each year established from four to five Year 7 classes. The classes were established using selective-school test results and feeder school information about the students. From the point of view of science, the classes were effectively ungraded.

Only the head teacher from the WAE and AE school were interviewed. The head teacher and seven science staff members were involved in the interview at the WBE school. Both schools provided a range of assessment-related artifacts for Years 7 to 10. The HT science at the WBE school brought a partially completed results proforma to the interview. The HT at the AE school had completed the proforma for the interview. The WBE and AE school have both engaged with VALID10, and the WAE school had no plans for doing so at the time of interview.

At least 94% of students in all three schools were in the top achievement band for EV results. None of the three schools had any students achieving lower than the middle achievement band. Students at fully selective entry schools are there because of their outstanding performance on pen-and-paper tests of general ability, literacy (including writing) and numeracy. The NAPLAN predictors for their EV results put them in the reverse order to that established by their residuals (see Table 5.9).

Their SEA scores (all other factors being equal) for the three schools were not comparable (see Table 5.9) should have delivered MGFSWBE1 with the best EV result; it came 3rd. MCFSWAE1, which should have been 2nd, was 1st, ahead of MCFSAE2, which came 2nd.

The international TIMSS and PISA test results do not reveal any gender bias in achievement in the first few years of secondary schooling in NSW schools (Thomson, DeBortli et al., 2017; Thomson, Wernert et al., 2017). However, there is international research evidence that adolescent girls in the most developed nations are less engaged with science than adolescent boys are (Bøe, Henriksen, Lyons, & Schreiner, 2013; Sjøberg & Schreiner, 2010). For this reason the comparisons made here will focus on the two girls schools.

Table 5.9 provides some data relevant to making comparisons and findings relevant to the predictions. The EV data for the three schools was sourced from Table K.1 in Appendix I.

Table 5.9
Pair FOUR selected statistics

School	Y8 ACH		Y8 ENG	Y10 ACH	
	SCH (%)	STA	TOP / 16	SCH (%)	STA
MCFSWAE1	T_{EV} 95	511	4	A 63	485
EV = 101.97 ± 0.71	T_{ER} 85	419			
SEAS = 8.6 ± 0.16	T_{WS} 80	412			
RES = 1.19 ± 0.29	T_{CS} 89	397			
MGFSAE2*	T_{EV} 95	511	15	A 83	639
EV = 101.00 ± 0.65	T_{ER} 85	419			
SEAS = 8.3 ± 0.16	T_{WS} 76	392			
RES = -0.09 ± 0.44	T_{CS} 89	397			
MGFSWBE1*	T_{EV} 94	505	8	A 85	654
EV = 97.99 ± 0.54	T_{ER} 70	345			
SEAS = 8.9 ± 0.14	T_{WS} 78	402			
RES = -1.42 ± 0.02	T_{CS} 93	415			

Y8 ACH = the proportion of Year 8 students. SCH (%) = school result. T_{EV} = proportion of overall EV result in the top band / T_{ER} = proportion of results in the top band extended response tasks. T_{WS} = proportion in the top band for working scientifically. T_{CS} = proportion of results in the top band for communicating scientifically.

STA = ratio of top band school achievement relative to the state score at 100 (ratio obtained by dividing school % proportion by state % proportion).

Y8 ENG = the rank order of schools based on engagement scores. TOP / 16 = the rank out of 16 (the state figure is counted as a school).

Y10 ACH = proportions of A grades at the school and relative to the state. SCH % = the proportion of Year 10 students attaining A grades. STA = the ratio of A grades at the school relative to the state set at 100 (ratio produced by dividing the school % proportion by the state % proportion).

* Girls schools.

In this comparison, the AE school with the lower SEA score (8.3 ± 0.16) had the better EV result (101.00 ± 0.65 compared to 97.99 ± 0.54) and a higher residual than the WBE school (-0.09 ± 0.44 compared to -1.42 ± 0.02). The lower SEA score for the AE school is strongly suggestive of greater value adding to its EV result than if it had a comparable SEA score. From this perspective, the achievement component of prediction one is satisfied.

The greatest achievement discrepancy between the AE and WBE school is in the extended response report category, where the proportion of girls at the WBE school was 70% compared to 85% at the AE school. This will be discussed in the summative comments part for this section.

The sources of data on relative engagement were Tables K.5A, B & C in Appendix I. The comparisons below include the relative order of schools (in parentheses). The survey results were the measure of student engagement for science at the end of Year 8. Only top band students in each of the 15 case study schools will be compared here. At the end of Year 8, girls at the:

- WBE (4th) school were more positive about taking a senior science subject (Item A) than were their AE (9th) counterparts (and both school scores were below the state figure)
- AE (13th) and WBE (14th) schools thought science harder than their counterparts across the state (Item B) and both were above the state scores in their agreement
- WBE (9th) and AE (14th) schools enjoyed their primary school science experiences in the order listed; the AE school ranking was below the state (Item C)
- WBE (12th) and AE (15th) schools enjoyed their secondary school science experiences and proportions including science in their three favourite subjects as listed here; both schools combined scores were below the state (Items D plus E score)
- WBE (6th) and AE (12th) schools listed science in the group of three subjects that students thought they learnt the most in in the order listed; the AE school's score was below the state (Item F).

It is clear that for MGFSAE2, high achievement in science is not associated with positive attitudes toward science. Possible reasons for this will be canvassed in the summative comments part for this section.

YEAR 10 ACHIEVEMENT

The school data provided by the two schools included the levels / grades awarded on the basis of the pattern of results from the external examination at Year 10. The last exam was in 2011. There was a discontinuity between the results before and after the Year 10 exam ended, thus this researcher was reluctant to draw any conclusions about achievement changes relative to Year 8 and remained silent about prediction two for these schools.

YEAR 12 ENGAGEMENT

Table 5.10 shows the proportions of students at the three schools completing Year 12 science courses. Overall the WBE school's proportions of Year 12 completions were larger than the AE school's completions. This was contrary to prediction three.

Table 5.10

Year 12 science course completions (2013-2015 averages)

School	MCFSWAE1		MGFSAE2		MGFSWBE1	
Subject (state proportion)	School	State	School	State	School	State
Biology (28.5)	34	119	20	70	22	77
Chemistry (18)	70	389	54	300	58	322
Earth and Environ. Sci. (2.4)	N/A	N/A	N/A	N/A	N/A	N/A
Physics (16)	46.4	288	23	144	28	175
Senior Science (10.4)	5.8	87	N/A	N/A	N/A	N/A

School = proportion of students relative to English at the school (relative to 100)

State = proportions of students at the school (relative to the state set at 100) completing Year 12 courses.

Comparative Summative Comments for Pair FOUR (MGFSAE2 and MGFSWBE1)

The following discussion of findings in relation to the predictions for the two girls schools compared here and their contribution to answering research question three draws on the school data provided above and the assessment-related work narratives for the schools in Appendix H.

Findings in relation to the predictions are qualified because the SEA scores are not comparable. The fact that the AE school has a lower SEA score than the WBE school provides confidence that the residual difference supports the conclusion that the AE schools EV results were better than expected due to their more frequent exposure to formative practices. A review of the assessment-related work narratives for the three schools and results in other categories points to a difference in emphasis on what was valued as sources of evidence of learning as explained in the next paragraph.

In the WBE school, students worked on cross-faculty projects and beyond the school gates. Evidence of learning was obtained from student-created models, written reports (using tightly constrained scaffolds) and group presentations supported by technology as well as traditional pen-and-paper tests. By contrast, the AE school had a strong emphasis on written evidence of learning drawn from traditional laboratory and text-based experiences mostly provided within the school boundaries. ‘Writing to learn’ was a higher priority for the AE school. Simply put, students at the WBE school did not have the same opportunities to write answers to open-ended extended response tasks as students in the AE school. This researcher suggests that differential opportunity is the main contributor to the better EV results at the AE school. Fuller accounts of the narratives for the two schools are provided in Appendix H.

Prediction three is about the proportions of students (relative to the state) completing senior science courses. The expectation from prediction three was that the AE school would have a higher proportion of its students completing Year 12 science courses than the WBE school, which was clearly not the situation here. It would appear that the AE girls’ strong dislike for science at the end of Year 8 continued and was a factor in their lower uptake of science courses in the senior years.

If we extrapolate students' ratings of their school science experience at the AE and WBE schools from Years 8 to 10, when students make choices about whether and what science to do in the senior years, it appears here that Year 8 engagement and Year 12 engagement (senior course completions) correlate better than Year 8 achievement and later engagement. Prediction three includes the understanding that self-regulation prioritises learning over enjoyment. The links between Year 8 achievement and later engagement will be explored in the next section (Section 5.4), where the findings from statistical correlations will be reported.

That said, the narratives for the two schools suggest that the more positive attitude to science at the WBE school is related to qualitatively different learning programs. At the end of Year 8, the girls at the AE school had demonstrably better writing skills, but the girls were clearly not enjoying the science they wrote about.

5.3.5 Pair FIVE: PCWAE2 and PCWAE3

The rank ordering of schools in the state that is based on the relative size and polarity of the residual from a regression of EV results over a NAPLAN-based predictor produced an unexpected finding, which was that the proportion of provincial schools ranked in the top 20% of the state went from 9% to 56% (see Section 4.1). Seven of the 12 WAE schools that identified themselves were provincial schools (see Table 5.1), which was a coincidence but reflected that state-wide finding.

In principle, comparing a provincial and a metropolitan school, or even a full selective school, should not matter as long as the SEA scores are identical and school factors (such as attendance rates) are taken into account. This was done in the earlier analyses to compare PCWAE1 and MCWAE1, and PCWAE2 and MCWBE5. The premise is that the SEA score captures all that matters when it comes to students' science learning potential.

This section compares three WAE provincial schools, two of which (PCWAE1 and PCWAE2) were looked at earlier in this chapter, but in the context of comparisons with other schools having the same SEA scores as the provincial schools. The third provincial school (PCWAE3) was selected for pairing with PCWAE2 because it had comparable SEA scores and comparable residuals ('comparable' meaning not significantly different in the statistical sense). The thumbnail sketches of PCWAE1 and PCWAE2 were provided above in the context of Pairs ONE and THREE respectively; the thumbnail sketch for PCWAE3 follows.

PCWAE3 is the largest of the three provincial schools. Each year, the school establishes four to five Year 7 classes using data from feeder schools. A single top stream class is established from the highest achievers and a bottom stream small class consists of students with the weakest literacy levels. The two or three classes in the middle have the remainder of the students allocated in no particular order. All classes are mixed ability from a science perspective. The science head teacher was the only teacher involved in the interview and had been at the school from before the period of interest. Artifacts of Year 7 and Year 8 assessment were provided at the interview and the proforma had been completed. The school had no plans to take up VALID10 at the time of interview.

Table 5.11 provides selected data sourced from data tables in Appendix I to make comparisons relevant to addressing the three predictions.

Table 5.11

Pair FIVE selected statistics

School	Y8 ACH		Y8 ENG		Y10 ACH	
	SCH (%)	STA	ALL / 12	TOP / 16	SCH (%)	STA
PCWAE2*	T 12	65	7	11	A-B 17	45
EV = 84.79 ± 0.31 SEAS = 1.8 ± 0.45 RES = 1.69 ± 0.21	B 12	89			D-E 37	142
PCWAE3*	T 12	65	11	16	A-B 24	71
EV = 83.64 ± 0.79 SEAS = 2.0 ± 0.27 RES = 1.43 ± 0.25	B 13	96			D-E 7	127

Y8 ACH = the proportion of Year 8 students in the top (T) and bottom (B) achievement bands. SCH (%) = school proportions represented as a percentage. STA = the proportion of students at the school expressed as a ratio (school proportion as a % over the state proportion as a %) relative to the state designated as 100.

Y8 ENG = the rank order of schools based on engagement scores. ALL = all three achievement bands / 12 = the rank out of 12 non-selective schools based on the total survey scores for students at a school (the state figure is counted as a school) / TOP = top achievement band students / 16 = school rank for top band students in the 16 case study schools for which data had been provided (the state figure is counted as a school).

Y10 ACH = the proportion of Year 10 students attaining grades A and B and D and E. SCH (%) = the proportion of students at a school with grades A&B and D&E represented as a percentage. STA = the proportion of students at the school as a ratio (school proportion as a % over the state proportion as a %) relative to the state designated as 100.

YEAR 8 ACHIEVEMENT AND ENGAGEMENT

The data relevant to confirming the achievement component of prediction one (Table 5.11), has PCWAE2 with a statistically significantly higher EV result than PCWAE3, though only just. On the balance of probabilities (a notionally lower SEA score and higher EV result), this supports prediction one. Note that PCWAE1's results are statistically significantly better than either of Pair FIVE's results, but it has a statistically significantly higher SEA score than either of the two schools compared here.

Table 5.12 records the proportions of students at each of the three achievement levels in three EV result reporting categories (ER, WS and CS are all identified in the legend). This level of comparison is warranted because the EV results from the two schools are very close (just as they were for MCFSAE2 and MCFSWBE1).

Table 5.12
PCWAE2 and PCWAE3 Year 8 EV results

				EV %		ER %		WS %		CS %	
School	SEAS	SRES	AB	sch	sta	sch	sta	sch	sta	sch	sta
PCWAE2	1.8	1.69	5-6	12	18.6	18	20.3	16	19.4	14	22.4
			3-4	76	67.9	66	63.4	69	63.3	71	60.3
			1-2	12	13.5	16	16.3	15	17.3	15	17.3
PCWAE3	2.0	1.43	5-6	12	18.6	15	20.3	15	19.4	14	22.4
			3-4	75	67.9	66	63.4	68	63.3	66	60.3
			1-2	13	13.5	19	16.3	17	17.3	20	17.3

Note. SEAS = socio-educational advantage score / SRES = school residual / AB = achievement band / EV % = proportions of students within each achievement band based on their total EV result (sch = school & sta = state) / ER % = proportions for extended response tasks / WS% = proportions for working scientifically / CS% = proportions for communicating scientifically

In three of the four reporting categories, there is a small but consistent positive skew in PCWAE2's results. This observation was most pronounced for the extended response task category. The relatively low top band proportions of students at both schools is consistent with their relatively low SEA scores. However, the large proportions of students in the middle band and small proportions in the bottom band is testimony to effective teaching in the two schools. Possible reasons for this result will be explored in the discussion part of this section.

Table 5.13 enables comparisons of engagement for the three provincial schools. These three are included here because the achievement–engagement pattern described in Section 5.2.1 was a general one and not linked to pairs of schools with common SEA scores. The findings in that section showed that higher EV results and positive attitudes toward school science experience were associated. Engagement findings for PCWAE1 and PCWAE2 (reported on in comparisons above) were inconsistent with that general finding. In both, engagement measures were well below the two metropolitan schools each was being compared with, despite both provincial schools having better EV results than the metropolitan school each was compared with.

Table 5.13

Case study school ranks based on student scores for the six items from the student survey

School	PCWAE1		PCWAE2		PCWAE3	
Item	ALL / 11	TOP / 15	ALL / 11	TOP / 15	ALL / 11	TOP / 15
A	5	11	3	8	6	10
B	***3***	9	8	8	***4***	4
C	9	11	7	7	11	15
D	9	13	6	10	8	15
E	10	8	11	13	8	14
F	9	14	5	8	10	13
AVERAGE RANK	7.5	11	6.7	9	7.8	11.8

ALL / 11 = all students at the 11 non-selective entry schools. TOP / 15 = top band achievers at the 15 case study schools. ***3*** & ***4*** are both better than the state figures

When considering student rankings on engagement for the non-selective schools, the three provincial schools were clearly in the lowest half of the state for all but the first two items related to engagement. Overall, PCWAE2 was more positive than either PCWAE1 or PCWAE3. This was also true for the top band achievers in all three schools.

When comparing only PCWAE2 and PCWAE3, it was found that, overall, of PCWAE2 students at the end of Year 8:

- more wanted to do a senior science course;
- fewer thought science their hardest subject;
- more enjoyed their primary science classes;
- more enjoyed their secondary science classes;
- a smaller proportion listed science in their list of three favourite subjects;

- a larger proportion listed science in the group of three subjects the students thought they learnt most in.

YEAR 10 ACHIEVEMENT

Turning now to later achievement, PCWAE3 did not provide Year 10 achievement data from 2012 to 2014 but did provide results for the three years up to and including 2011, the year of the last external science test. It is obvious that a direct comparison between PCWAE2's and PCWAE3's Year 10 results would not be a valid exercise. However, it is possible to compare the grade / level distributions compared to state figures for the appropriate years and then to infer, with appropriate caution, the extrapolation of that pattern to the years of interest (2011–2014).

Table 5.11 provides the data showing changes in the pattern of results from Year 8 to Year 10 relative to the state. Distributions of school results relative to the state, and the changed proportions, show that PCWAE2 students have not retained their achievement edge over PCWAE3. These data do not confirm prediction two. Reasons for the change in results patterns are discussed in the summative comments part of this subsection.

YEAR 12 ENGAGEMENT

Looking next at Year 12 completions for the two provincial schools (Table 5.14), the student proportions completing science courses at the end of Year 12 at PCWAE2 relative to the state and compared to PCWAE3 were: more in Biology, comparable in Chemistry, more in Physics, and more in Senior Science. PCWAE3 also offered Earth and Environmental Science, which PCWAE2 did not (10% of PCWAE3 students completed this course at the end of Year 12). Without knowing more details (such as whether Biology and Earth and Environmental Science were offered as an either/or option or both could be taken), it would appear that PCWAE2 had more of its students completing Year 12 courses than had PCWAE3, which was consistent with prediction three.

Table 5.14
Year 12 science course completions (2013-2015 averages)

School	PCWAE2		PCWAE3	
Subject (state proportion)	School	State (100)	School	State (100)
Biology (28.5)	38	133	19	67
Chemistry (18)	16	89	17	94
Earth and Environmental Science (2.4)	N/A	N/A	10	417
Physics (16)	13	81	11	69
Senior Science (10.4)	30	288	22	212

School = proportion of students relative to English at the school (relative to 100)
State = proportions of students at the school (relative to the state set at 100) completing Year 12 courses.

COMPARATIVE SUMMATIVE COMMENTS FOR PAIR FIVE (PCWAE2 AND PCWAE3)

The following discussion of findings in relation to the predictions for the schools compared here and their contribution to answering research question three draws on the school data mentioned above and the assessment-related work narratives for the schools in Appendix H.

For prediction one, PCWAE2 had the better achievement and engagement overall. For the two schools the assessment narratives discussed the priority given in both schools to working on improving students' literacy skills. The assessment narratives for both schools provided convincing evidence of differential teaching that aimed to address the full range of literacy deficits that students bring to science classes.

Both schools are in the WAE group of schools. WAE schools were more frequent users of three dimensions of formative practice than WBE schools. It is reasonable to suggest that at the end of Year 8, PCWAE2 was more successful at lifting students results than PCWAE3 because PCWAE2 teachers were more effective at promoting discourse that elicits evidence of learning, providing feedback that advances learning, and modeling good learning behaviours to peers and students.

The evidence for this conclusion is the positive bias in the results for PCWAE2 students in the extended response component of the EV results; in the detail of the assessment narrative for PCWAE2 compared to PCWAE3, and; more students at PCWAE2 had put science in the list of three subjects they thought they had learned most in.

This being the case, how is it that by the end of Year 10, the overall results at PCWAE3 are better? The anomaly to be explained is the pattern of better results by PCWAE3 at the end of Year 10 compared to PCWAE2 (relative to the state), which is contrary to prediction two. One possibility is that PCWAE3's SEA score advantage (2.0 compared to 1.8) is real. A second possibility is that student absenteeism was higher at PCWAE2 over the four years. A third possibility is the impact of a high-stakes summative assessment regime such as was revealed in the narrative for PCWAE2 compared to the low-key approach by PCWAE3 to summative assessment.

If student absenteeism is higher at PCWAE2, this might be a decisive factor in reducing their Year 10 results. Disruption to individual learning progress due to absence and disruption to group learning as a result of absenteeism was identified by the head teacher in the interview at PCWAE2. From the *MySchool* website, the proportion of indigenous students compared to non-indigenous students at PCWAE3 is higher than at PCWAE2 (1 in 4 compared to 1 in 5). This becomes relevant because data from the *MySchool* website for the two schools shows that the attendance rates for PCWAE3 students are 15% lower for indigenous and 5% lower for non-indigenous students than at PCWAE2. Thus, on any one day the proportion of all students away at the two schools is likely to be greater at PCWAE3 than at PCWAE2. So, despite lower daily attendance rates at PCWAE3, its Year 10 results are better than PCWAE2's results.

Given that absenteeism is more likely to have a greater negative effect on achievement at PCWAE3 than at PCWAE2, it is possible that there is another more potent factor at work here related to different approaches to summative assessment. Research discussed in Chapter Two identified the negative effect of high-stakes summative assessment on the motivation to learn of students with poor learning histories. Both schools have relatively high proportions of students with poor learning histories (reinforced by publicity around NAPLAN results, which are generally poor at both these schools).

PCWAE2 have strictly graded classes in science, the composition of which is changed after summative assessment every six months from half-way through Year 7 to half-way through Year 9. PCWAE3 takes a low-key approach to summative assessment and keeps the number of formal assessment tasks to a minimum. Once established, classes at PCWAE3 are retained relatively unchanged until the end of Year 8. It maybe that over the four years (from Year 7 to 10) that the negative impact on motivation to learn is greater at PCWAE2 than PCWAE3.

The approach to assessment at PCWAE3 is very similar to that at MCWBE5. (MCWBE5 was compared to PCWAE2 as pair THREE above). Absenteeism at MCWBE5 was the lowest of the three schools. Like PCWAE3, MCWBE5 established streamed classes at the beginning of Year 7 which they retained until the end of Year 8. Summative assessment was a low-key affair. All three schools had comparable SEA scores, thus making comparison fair, based on their SEA scores.

MCWBE5's residual is well below both the provincial WAE schools and EV test result was lower than either of the two provincial schools. Like PCWAE3, MCWBE5 performed better than PCWAE2 by the end of Year 10. This outcome is at least suggestive that summative assessment practices at PCWAE2 may have been a contributor to its lower achievement by the end of Year 10 than either PCWAE3 or MCWBE5.

As to attribution of summative assessment impact on differences in engagement at the end of Year 8 and Year 12 for the three schools, the evidence is less clear.

At the end of Year 8 PCWAE2 students were enjoying their secondary science classes more than PCWAE3 students, top students at both schools less so than their overall result indicates they should (see section 5.2.1). MCWBE5 students were more positive (3rd on Item D) than both the provincial schools and this was shared by their top students as well (3rd on Item D).

A smaller proportion of students at PCWAE2 included science in their list of three favourite subjects (Item E) than at PCWAE3. Top band students at both schools had even smaller proportions (out of 15 schools, PCWAE2 was 13th and PCWAE3 was 14th). MCWBE5 ranked 3rd in the state overall and its top band students were also 3rd.

When it came to the proportions identifying science in the group of subjects they thought they learnt most in, more PCWAE2 students than PCWAE3 students did so (5th overall and 10th, respectively). Top band students repeated that pattern, but were a smaller proportion again (8th compared to 13th out of 15 schools). MCWBE5 students had the 3rd largest proportion in the state and their top band was the 7th largest. Based on the above, it would be difficult to make a definitive claim about the negative impact of the assessment regime at PCWAE2 on either enjoyment (Items D & E) or sense of achievement (Item F).

An explanation for MCWBE5 students' much higher satisfaction with their school science experience compared to either PCWAE2 or PCWAE3 would appear to be less related to summative assessment than teacher use of formative practice. The assessment-related narrative for MCWBE5 points to teachers at that school giving students a greater say in what to do in the name of science education, as well as more opportunities for peer and self-assessment which seem to come to increase with the number of years spent at secondary. MCWBE5 had the highest proportions of students completing senior science courses of the three schools.

5.4 Correlation and strength of associations between school variables

Correlation provides a way of confirming (or disconfirming) the relative strengths of associations between variables. The strength of a correlation can provide more support for one or other inference when considering the qualitative evidence in the assessment narratives. In the situations being compared here we are looking at scores that are two (Year 8 to Year 10) and four years apart (Year 8 to Year 12). As mentioned earlier, students make their choices for science courses they wish to study in Years 11 and 12 (the last two years of secondary education) in the middle of Year 10.

This researcher's experience suggests that once made, students tend to follow through with those choices. Thus, the decision to use Year 12 data for completions needs to recognise that the figures reflect decisions made more than two years earlier, less than two years after the EV test, and before Year 10 results were finalised. The EV test has been in place since 2007; the results being looked at here are for the four Year 8 cohorts from 2011 to 2014. Their scores are correlated with Year 10 students who did the EV test from 2009 to 2012 and Year 12 students who did the EV test from 2007 to 2010. The first EV test for students across the state in NSW was in 2007. The advice about assessment for learning was promulgated with the 2003 syllabus. The point being made here is that changes in response to both initiatives were as strongly embedded in practice as they were ever going to be by the end of 2014, when the new syllabus became the basis for ongoing EV testing. From this perspective the correlation between sets of results that are asynchronous was not considered a major issue when it came to assessing the limitations of correlation statistics as applied here.

Another assumption here is that science results are a function of all the science teachers' efforts at a school and that staff changes or traumatic events at any one school during that time were relatively minor. Nevertheless, any statistically significant correlations need to consider specific school circumstances. School circumstances that were likely to impact results and engagement were disclosed to the researcher and were included in the assessment-related work narratives for the case study schools as appropriate.

As explained earlier, SPSS software was used by the researcher in this project to perform bivariate correlations using either parametric or nonparametric models as appropriate.

5.4.1 Correlations: fully selective entry case study schools (n = 3)

A two-tailed correlation analysis using SPSS was carried out to test the observation that engagement at the end of Year 8 is the better predictor of later engagement. Measures of the following variables for the three schools were used.

1. Year 8 results (Year 8 achievement)
2. Year 8 scores for Item A of the student survey (aspiring to do senior science courses)
3. Year 8 scores for Items D and E from the student survey (Year 8 engagement)
4. Year 10 proportions of A grades (later achievement)

Year 12 mean senior course completions in Biology, Chemistry and Physics only (later engagement).

The data sets satisfied the Shapiro-Wilk test for normality (p >.05). Results are reported in terms of the Pearson's correlation coefficient r, degrees of freedom (1), and a two-tailed significance (p) value at either the p = .01 or p = .05 level (as shown by the value quoted with the reported correlation coefficient).

For engagement at Year 8 (two items from the student survey) and achievement at Year 8, the correlations for the top band students (at least 94% of all students at the schools) on Item D (r_D (1) = -.14, p > .05) and for Item E (r_E (1) = .41, p > .05) were slightly negative and moderately positive, respectively, but neither was statistically significant. Thus, it would be difficult to defend any conclusion that liking science classes and doing well in the EV test were related at these three schools.

The correlations between Year 8 engagement (the same two items as before) and Year 10 achievement were r_D (1) = -.88, p > .05 and r_E (1) = -1.0, p < .05. The former was highly negative and not statistically significant, the latter was very highly negative and statistically significant. Students who put subjects other than science in their list of three favourite subjects at the end of Year 8 achieved very good results in science at the end of Year 10. In these three schools it seems that not liking science was no impediment to achieving well in it at the end of Year 10.

In relation to Year 8 engagement (two items as before) and Year 12 engagement (Biology, Chemistry and Physics completions), the correlations were highly positive but not statistically significant (r_D (1) = .68, p > .05) and for r_E (1) = .96, p > .05). At the end of Year 8, aspiring to do science in the senior years (Item A in the survey) and actual engagement figures for those students who had chosen science at the end of Year 10 (two years after their EV test) and finished it at the end of Year 12 (four years after that EV test) were highly positively correlated but not statistically significant (r_A = .92, p > .05). The correlation between Year 8 achievement and Year 12 completions (r = .63, p > .05) was also highly positive but not statistically significant.

Thus, it seems that for the three metropolitan fully selective schools, the combination of wanting to do senior science courses (Item A in the student survey) and liking science at the end of Year 8 (Items D and E) was likely to be a better predictor of Year 12 science course completions than Year 8 achievement.

5.5.2 Correlations: non-selective entry case study schools (n = 11)

The testing of correlations between variables was repeated for the non-selective entry case study schools (n = 11). Two more variables were added to the list for the purpose of this analysis. The variables tested were:

1. Year 8 results (an achievement measure)
2. Year 8 aspiring to do senior science courses (Item A on the student survey)
3. Year 8 student survey items D plus E (a collective measure of Year 8 engagement)
4. Year 10 achievement (the cumulative proportion of As, Bs and Cs awarded to the cohort)
5. Year 12 engagement (the average of school proportions completing Biology, Chemistry and Physics courses at the end of Year 12)
6. Residuals (a measure of teaching effect / scientific literacy scores)

7. SEA scores (the measure of socio-educational advantage).

All seven data sets to be compared passed the Shapiro-Wilk test for normality (p > .05). On that basis it was decided to use the Pearson parametric correlation (r) two-tailed test in the SPSS software. The model provides for nine degrees of freedom (based on n = 11) and a significance (p) value at either the .01 or .05 level (as reported with the correlation coefficient produced by the SPSS model).

The first tests were to assess whether Year 8 engagement or Year 8 achievement was the better predictor of later achievement (Year 10 results) and engagement (Year 12 senior science course completions).

The correlation between Year 8 engagement and Year 10 results was strongly negative and statistically significant (r (9) = -.69, p > .05). This figure suggests that not liking science at the end of Year 8 and doing well in it later on (at the end of Year 10) was the norm for the provincial and non-selective entry metropolitan case study schools.

Between Year 8 engagement and Year 12 engagement, the correlation was moderately positive but not statistically significant (r(9) = .384, p >.05). This is an expected result but in no way predictive in this context. On the other hand, the correlation between Year 8 achievement and Year 10 achievement (r(9) = .70, p <.05) was highly positive and statistically significant. The correlation between Year 8 achievement and Year 12 engagement (r(9) = .65, p < .05) was also highly positive and statistically significant.

For the non-selective case study schools compared for this exercise, Year 8 achievement is a much better predictor of later achievement (as measured by Year 10 results) and engagement (Year 12 senior science course completions) than Year 8 engagement.

In the comparisons looking at measures of Year 8 engagement in provincial schools and metropolitan case study schools relative to the state, it appeared that provincial schools had a lower level of engagement with science relative to the state and relative to the metropolitan schools they were being compared with.

The three provincial schools all had low SEA scores. The non-selective metropolitan schools had slightly higher SEA scores overall. One possibility is that a low SEA score might be an indicator of low interest in science. The correlation between SEA scores and Year 8 engagement for the non-selective case study schools was shown to be moderately negative but not statistically significantly so ($r(9) = -.42, p > .05$). Thus, any suggestion that a low SEA score and low engagement with science at Year 8 are necessarily related would not be supported by this finding.

The correlation between the residuals (a measure of scientific literacy achievement) for the 11 non-selective entry schools and engagement (liking their school science experience) was moderately negative but not statistically significantly so ($r(9) = -.30, p > .05$). The conclusion from this result is that for the case study schools, good EV results and students not liking their science experience is the more likely combination.

5.5.3 Correlations: provincial case study schools (n = 3)

To assess the strength of the associations between variables, the following variables involved in the comparisons between the three provincial schools were tested for statistically significant correlations using SPSS:

1. EV results (Year 8 benchmark measure of achievement)
2. Student survey Items D + E combined levels score (Year 8 benchmark measure of engagement with science)
3. Year 10 sum of grades A + B + C (later achievement)
4. Year 12 completions of Biology, Chemistry and Physics (average mean proportions compared to English at that school)
5. Residual
6. SEA score.

Most of the variable data sets passed the Shapiro-Wilk test for normality ($p > .05$). As a result, the SPSS procedure for a two tailed, bivariate, parametric correlation of the variables was used. Results are reported in terms of Pearson's Correlation Coefficient (r), degrees of freedom (1) and whether the correlation was statistically significant relative to the model reported value at either the $p < .01$ or $p < .05$ level of significance.

The SEA score was included to test the possibility that engagement may be positively correlated with it. The correlation for the three provincial schools on the Year 8 engagement variable (Items D + E) and SEA score produced a moderately negative but not statistically significant correlation ($r(1) = -.43$, $p > .05$). This was consistent with the correlation for all 11 non selective schools ($r(9) = -.42$, $p > .05$) and for the full complement of case study schools ($r(13) = -.38$, $p > .05$). The evidence here is that SEA score and engagement are, if anything, negatively correlated. The higher the students' learning potential, the less they liked their school science experience.

Another check is to see if the residual and engagement (Items D + E) are positively correlated. The residual is a measure of the impact of science teaching on achievement, but it might, arguably, be an indicator of student attitudinal responses to that teaching. For the three provincial schools, the correlation was highly negative but not statistically significant ($r(1) = -.76$, $p > .05$). For the 11 non-selective entry schools the figure was moderately negative and also not statistically significant ($r(9) = -.30$, $p > .05$). For all case study schools $r(13) = -.27$, $p > .05$. Again, the analysis does not support any definitive conclusion but is suggestive that the more capable students across the state are not enjoying their science lessons.

When the correlation between student satisfaction with their Year 8 school science experience and being in either a provincial (1) or metropolitan school (2) was tested for the 11 non-selective entry schools, the result was moderately negative ($r(9) = -.46$, $p >.05$) but not statistically significant. As well, comparing the average levels of satisfaction (descriptive statistic) recorded for Items D and E for all the case study schools (n = 15) shows that $\bar{x}_{metro} = 34.3$ versus $\bar{x}_{prov} = 22.1$. Thus, it is not unreasonable to conclude from the above analyses that provincial students in this sample were less positive about their experience of school science than their metropolitan counterparts.

5.5 Summary

The compared pairs of schools were PCWAE1 and MCWAE1, MCAE2 and MCWBE3, PCWAE2 and MCWBE5, MGFSAE2 and MGFSWBE1, and PCWAE2 and PCWAE3. The first three pairs of schools had comparable SEA scores but statistically significantly different residuals. The fourth pair were fully selective entry girls' schools. The girls' schools were paired on the basis of being selective entry girls' schools (but they did have statistically significantly different SEA scores and residuals). The fifth pair were coeducational provincial schools with comparable SEA scores and residuals. 'Comparable' means the scores were not statistically significantly different.

The first and fifth pair of schools were WAE schools because they had highly positive residuals, which meant that their EV results were well above expected. The residuals for the other three pairs were different enough to assign each school in the pair to a different school group based on their EV results being as expected (AE) or well below expectation (WBE). Expectation was relative to a NAPLAN-based predicted science score, as explained in Section 3.3.

The findings reported in Section 4.5 were that teachers in WAE schools were more frequent users of three of five dimensions of formative practice than were their colleagues in WBE schools. As well, overall, teachers in AE schools were more like their WAE counterparts in the frequency of their use of formative practices.

The research question to be answered in this chapter was:

> Does the use of (and if so, how do) formative practices by teachers improve students' EV results and later achievement in and engagement with science?

Section 3.6 explained that by identifying schools with matching SEA scores in a list of schools sorted from top to bottom according to the size of their residuals the possibility arises of showing that better than expected EV results (in terms of a predictor) are higher actual EV results (in absolute terms). That said, it can be seen from the tables in Section 5.3 that in terms of EV results, PCWAE1's EV result is higher than MCWAE1's, MCAE2's is higher than MCWBE3's, PCWAE2's is higher than MCWBE5's, and PCWAE2's is higher than PCWAE3's. Thus, it is possible to show that for four pairs of the case study schools where SEA scores could be matched, the schools with the biggest residuals had the better EV results. This was the claim made in prediction one. The high residuals are associated with more frequent use by teachers of three dimensions of formative practice, the use of which is linked to higher than expected scientific literacy content, thus boosting EV results.

The second part of prediction one links residual size to engagement, as measured by student scores on the six items chosen for consideration here. The presumption in making the link between achievement and engagement is that student exposure to formative practices has produced students who are not only good at science but enjoy learning it. This presumption was supported by research findings discussed in Chapter Two that had linked exposure to formative practices with the acquisition of good learning behaviours and positive dispositions toward learning. This researcher chose to use student enjoyment of their school science experience as a measure of positive commitment to learning science. Additional support for the linking of achievement and enjoyment was provided by the finding reported in Section 5.2.1 that at the end of Year 8, across the state, higher achievement and enjoyment of their school science experience were positively associated.

At the end of Year 8, students in the higher-achieving school in four of the five pairs of case study schools scored a combined Item D + E below the score of students in the school it was paired with (see Table K.5A in Appendix I). Item D was a rating of enjoyment of their secondary science classes and Item E was the proportion of students who had included science in the group of their three favourite subjects. The exception was the second pair, MCAE2 and MCWBE3, where the higher achieving school, MCAE2, was slightly above MCWBE3 in both the overall and top band comparisons. The closeness of the paired results here was somewhat surprising, given that MCAE2 promoted itself as a STEM school and established each year a class of students who had sat a selective entry test for that class on the basis of their interest in doing STEM.

These results appear to contradict the general finding in Section 5.2.1 that across the state, higher achievement and enjoyment of school science were positively associated. It seems that at the end of Year 8, high achievement had been accomplished at the expense of student enjoyment of their school science experience. This finding is also supported by the correlations reported in Section 5.4.

Of interest also was the observation (for 10 of the case study schools) that provincial students were more negative about their school science experience than metropolitan students. It also seems that the highest achieving students in the 10 schools were the ones most negative about this experience. To the extent that enjoyment of science was an indicator of self-regulation at the end of Year 8, these findings are not supportive of that conclusion, nor are the findings promising as predictors of later engagement with science (Year 12 science course completions).

Prediction two was that the school (in the pairs of schools) with the bigger residual at the end of Year 8 would go on to have the better results at the end of Year 10. The prediction was confirmed for the first pair of schools (PCWAE1 and MCWAE1). It was not possible to make the comparison for the second pair (MCAE2 and MCWBE3) because MCWBE3 did not provide Year 10 results. It was not confirmed for the third (PCWAE2 and MCWBE5), fourth (MGFSAE2 and MGFSWBE1) and fifth (PCWAE2 andPCWAE3) pairs because of uncertainty about the comparability of the Year 10 results.

In an ideal world, results from a Year 10 EV test and related student survey would have been the best option for doing this comparison. Unfortunately, such a test and related survey did not become available until after 2014. It would therefore be unsafe to say that there were more self-regulated learners in WAE schools on the evidence from one pair of schools.

Findings related to prediction three were meant to demonstrate the persistence of positive attitudes to science provided by the presence of self-regulated learners in post Year 8 years of WAE schools. The independent evidence of the presence of self-regulated learners in greater numbers in WAE schools was supposed to be confirmed by higher proportions of students completing science courses at the end of Year 12. These were courses that students had initially chosen half-way through Year 10. Having questioned the validity and reliability of the data used to verify prediction two, we are left with data about Year 8 achievement, Year 8 engagement and Year 12 engagement.

Two ways of making that interschool comparison are provided. The first is the proportion of students at each school (relative to English which is a compulsory course for students wanting to receive the school exit credential at the end of Year 12) completing one or more of the five senior science courses that were available to students. All the schools researched here offered Biology, Chemistry and Physics in two or more of the three years of interest. Most also offered Senior Science and one offered Earth and Environmental Science (PCWAE3 in 2014) in the three years of interest (2013 to 2015).

The second is to compare this school proportion to the statewide proportions. The assumption behind both methods is that schools try to accommodate students' preferences to the best of their ability, given the resources schools are able to allocate. As a starting assumption, it was accepted that the school proportions shown here accurately reflect student demand for science courses more than the constraints of available resources; this will be less true the smaller the school is.

For the pairs of case study schools matched by SEA scores and different residuals indicating their degree of exposure to formative practices, the finding is that the better EV results were, the higher the proportion of students taking up and subsequently completing science courses. It was observed that achievement at the end of Year 8 was a stronger correlate with completion than liking the subject at that time. On balance, the combination of high achievement in science and not liking the experience was the norm for the case study schools, which was contradicted by the finding reported in Section 5.2.1 from the larger sample of schools that identified themselves.

In conclusion, the evidence discussed here confirms the positive association between better EV results and the frequency of exposure to:

- discourse that elicits evidence of learning
- the provision of feedback known to progress learning
- the use and modeling (to peers and students alike) of good learning behaviours.

The attempt to demonstrate that more frequent exposure to these three dimensions of formative practice had produced more self-regulated students in WAE schools than AE or WBE schools has not been demonstrated convincingly.

The assessment-related work narratives for the schools with better than expected EV results all had strong programs aimed at building student capacity to use the language of science to explain phenomena in the natural and made worlds they inhabit. It appears to this researcher that the literacy focus was in response to a wider school priority and/or in response to science teachers' awareness of the importance of scientific literacy for success in school science and as preparation for life and work after school. In schools where results were well below expectation, the assessment narratives had little explicit evidence of a priority for building student capacity to use the language of science as a tool for managing their learning of science.

CHAPTER 6: DISCUSSION AND FUTURE DIRECTIONS

6.1 Introduction

In Chapter One it was said that the objective of this thesis is to answer the broad question: To what extent is the assessment-related work of science teachers in NSW government schools formative and why it matters? Chapter Two gave two reasons for why this study matters. The first is that teacher use of formative practices (Black & Wiliam, 2009) is linked to high achievement (Hattie, 2012) as measured by traditional pen and paper summative tests. The second reason is that teaching students to use the strategies of formative assessment that underpin formative practices has shown considerable promise as a way of helping students to learn how to learn. According to the OECD, “laying the foundations for lifelong learning” (CERI, 2008)p. 1) should be a priority for the initial phase of schooling; knowing how to learn would be important preparation for that.

A 2018 updated list of effect sizes of particular teaching strategies on test results show formative practices to be amongst the most effective (Hattie, 2018). Strategies such as classroom discussion (0.82), providing feedback (0.70), response to intervention (1.29), jigsaw method (1.20) and scaffolding (0.82) are amongst the most powerful ways for teachers to operate in the classroom. Two curriculum strategies known to have above average effect sizes include repeated reading programs (0.75) and core and specific vocabulary programs (0.62). Both of these were in evidence in WAE case study schools. The effect-size of each strategy is provided in parenthesis; higher than 0.42 is an above average effect.

Research shows that teaching students the strategies of formative assessment is associated with them acquiring the skills of learning how to learn (LHTL) and becoming autonomous learners (Black et al., 2006; James, 2006). Learning autonomy is highly valued in the context of preparing people for life in the knowledge society and related global economy as discussed in Chapter Two. Again, according to Hattie (2018), the effect size on achievement of students’ acquiring and using these strategies is very high. Examples include: transfer strategies

(0.86), deliberative practice (0.79), strategies to integrate with prior knowledge (0.93) and summarization (0.79). Boyle et al. (2001) would refer to these strategies being used by students as “good learning behaviours” (p. 200).

As outlined in Chapter One, two initiatives introduced into NSW schools in 2003 and 2007 respectively, were designed to shift teachers’ assessment focus from summative to formative. The need for that shift had been elaborated in the review of the status and quality of science education in Australia published earlier (Goodrum et al., 2001).

The initiatives took the form of strong advice to teachers in the official curriculum about bringing teaching and assessment together (assessment for learning as it was called there) and a compulsory summative test for all Year 8 students. The test also had a diagnostic purpose which was to provide a progress report on science achievement half-way through the four-year science course. The diagnostic purpose of the EV program was enhanced by using the SOLO model in the design of the assessment framework for the EV program. Test items and tasks were designed to challenge students across six levels of thinking described by the model. At the time, both the NSW Department of Education (the Department), which was responsible for the test, and the curriculum authority that had produced the curriculum, provided additional support to teachers to assist them achieve the shift in emphasis. Examples of that support are outlined below and were described in earlier chapters.

The impact on the assessment-related work of science teachers was of both personal and professional interest to this researcher for reasons explained in Chapter One. To assess the impact of the two initiatives on assessment-related work as described in earlier chapters, three research questions were posed, a research design was developed and data gathered.

The first research question asked about teacher use of the resources related to the EV program. The program components included a test, a related student survey, provision of a report to parents and comprehensive results (to their teachers, school and school system), teacher support in the form of marker training and online professional learning modules. Discussion related to reasons for their use (or not) are reported in Section 6.2.

The second research question sought to find out the extent to which science teachers are using formative practices. Factors supporting or impeding the use of formative practices will be discussed in Section 6.3.

The third research question asked whether teacher use of formative practices improved student EV results and whether that use was linked to later achievement in and engagement with science. The answers to that question involving Year 8 students at a school, their later achievement (Year 10) and later engagement (Year 12) in science at school are discussed in Section 6.4.

The research methodology used to provide the findings informing the answers to research question three is the basis for claims by this researcher of originality and contribution to the international body of work on formative assessment.

Section 6.5 provides suggestions for further work to confirm findings.

The final section (Section 6.6) of this chapter provides recommendations to relevant authorities arising from the findings reported in this thesis.

6.2 Discussion of findings addressing research question one.

The question was: What use are science teachers making of the EV program and why is it used or not used?

The assessment framework for the EV test discussed in Chapter Two provides a map of learning along two axes, one axis being what should be learned in the name of science in Years 7 and 8 in NSW schools. The other axis describes six levels of thinking about science that a student can demonstrate in their responses to test items and tasks. The SOLO model provides descriptions for the six levels against which responses are to be judged. The broader context includes the tools for collecting evidence of learning (items and tasks in the test), assigning value to that evidence (marking), reporting results and making use of results to improve learning will be reported on here as well. Subsection 6.2.1 will focus on teacher use of the EV program resources more broadly; subsection 6.2.2 will explore the extent of teacher engagement with SOLO.

6.2.1 Teachers and the EV program

The following discussion relates to the collected responses from WAE, AE and WBE teachers (n= 85) to the online teacher survey and to evidence from the assessment narratives (Appendix H) as appropriate. The first five questions (Q1-Q5) in the teacher survey collected data about eight categories of actions describing the scope of teacher engagement with EV resources (Q1 and Q2), their level of understanding of the EV program (Q3), what the main purpose of the program was (Q4) and Q5 asked whether they would participate in the extension of the program to Year 10.

Chapter Four provided the detailed analyses of their responses. In brief, the findings were:

- the overall level of WBE teacher engagement with EV resources was lower than that for AE and WAE teachers (see Figure 4.1);
- that teacher understanding of the EV program, on a five-point scale ranging from very poor to poor, acceptable and then good to very good, located WBE teachers at acceptable and AE and WAE teachers at good (see Figure 4.1 B);
- most respondents wrote that the purpose of the EV program was to provide teachers with feedback on student learning (see Table 4.5); and
- that fewer WBE schools than AE or WAE schools would be taking up the VALID10 test opportunity.

Teachers from all three groups had discussed results with each other (66%) but less so with students (22%). A possible reason for not discussing results with students was provided in case study school narratives where several teachers had mentioned the large time gap between doing the test (November) and when the results were returned (March-April the following year).

None of the schools mentioned using items or tasks that students had done poorly in as the basis for reteaching. Poor performance in working scientifically or communicating scientifically are processes that could be retaught in the context of any topics, including those being done in Year 9. Reteaching in response to feedback is a characteristic of formative practice. More broadly, the literature on feedback is consistently of the view that the shorter the time difference between action and feedback, the more likely it is to be acted upon by the learner (Black (2007), Hattie & Timperley (2007), Masters (2013), Ruiz-Primo & Li (2012) and Shute (2007)).

Almost 40% of respondents had marked extended response tasks and almost 30% said they had attended workshops about the EV program (separate from training for marking extended response questions). Responding to the teacher survey was voluntary and anonymous. Teachers exposed to those two components of the EV program were possibly more inclined to respond to the survey than those not so aware. It may also be a factor in the high proportion of the same respondents who rated their understanding of the EV program (see Q3 reference above) as acceptable and higher (87%). That said, the EV program appears to be well understood by most of the respondents, including those in regional areas, a finding supported by answers to the next question in the survey, Q4.

The collation of teacher responses to the free response question (Q4) about the most important purpose for the EV program revealed multiple purposes from some respondents. Overall, the majority (70%) saw the purpose as being about providing feedback to teachers on learning and / or teaching, which was consistent with the Department's rhetoric about its purpose (see Chapter One). A minority (21%) saw its purpose as providing feedback on comparative performance with other schools and the state. A small proportion (9%) wrote about its purpose in terms of direct student benefit, which suggested they saw its potential for student self-evaluation which is a characteristic of formative thinking (Black and Wiliam, 2009).

In relation to the EV program overall, five responses provide an insight into issues some teachers have with the program. The head teachers at MCWBE5 and MCFSWBE1 were as not happy that science had been singled out for special treatment (in the form of an external test). Three other anonymous comments from respondents to the teacher survey included:

> *No idea. It's an imposition into an already crowded curriculum that requires an inordinate amount of time and resources for something that only appears to be there to justify a well-paid job or two elsewhere.* (WBE teacher)
>
> *[The Board] ticks the box for more tests for school. Justifies funding based on a test that doesn't necessarily match to the curriculum that the students are doing at the time.* (WAE teacher)
>
> *To keep people in Head Office in a job.* (WBE teacher)

These were the only negative comments in a total of ninety-five different responses to the question about the main purpose of the EV program.

Q5 from the survey asked about intentions to take-up the VALID10 test (introduced in 2015 on a voluntary basis; data collection for this project was in 2016). VALID10 is the Y10 equivalent of the Y8 test (as explained in Chapter One). It was impossible to be definitive about the intended take-up because this was an anonymous teacher response survey and there was no way of knowing which teachers were at what schools and whether there was more than one teacher from a school responding. Based on the raw data, 72% of teachers in WAE schools said they would be taking up the test that year, 52% of AE teachers and 47% of WBE teachers. The overall result for the sample (n = 84) was 56% which suggests that around half the state's Year 10 classes were preparing to take up the test on a voluntary basis in 2016.

Four schools reported wanting to see evidence of change from Year 8 to Year 10 (MGFSAE2, MCWBE4, MCWAE2 and PCWAE1) as the reason they took up the offer of participating in VALID10. Reasons given for not taking it on included to reduce assessment pressure on students (PCWAE3 and MCWBE5); issues to do with computer access (PCWAE2 and MCWBE5); teachers were too busy at that time of the year (MCFSWAE1); not much point given that students all went on to Year 11 anyway (MCWBE5 and MCFSWAE1). An aside: in separate conversations with science teachers outside the context of interviews in case study schools, some had reported they did not want to engage with VALID10 because, unlike the Year 8 test, training and marking was onsite (at school) and unpaid.

Asking case study teachers to complete a proforma with a sample of data for students at their school was meant to provide this researcher with an opportunity to find out the breadth and depth of analysis teachers do with both EV test and student survey results as well as their own teacher devised assessments. Only four of the case study schools had engaged with the proforma before the interview. Thus discussion at the interview of their practices in relation to data analysis was limited by the low overall response at that time. The low response was taken as an indicator that using data for formative purposes was not high on teachers' assessment agenda. This impression was confirmed and recorded in assessment narratives where learning intentions and success criteria were primarily used by teachers as the basis for feedback on strengths and weaknesses in answers to summative assessment tasks. There was little evidence of students being given the opportunity to use learning intentions and success criteria to provide feedback to peers or in self-assessment activities. (overall, only 16% of teachers said they often asked students to redo work to a higher standard). Both Hattie (2012) and Mitchell et al. (2009) describe research supporting the effectiveness of reflection as an aid to improving learning.

Most of the head teachers interviewed said that the level of results analysis asked for in the proforma was something they had not considered doing before. However, the three who did come to the interview with completed proformas said it was beneficial to look at the data over time and to identify trends. The head teacher at MCWAE2 suggested that providing a data downloading capability from SMART would encourage greater access and use by science teachers of the data, particularly the student survey data. She, along with the head teachers at MCWBE3 and MCPSWBE2, said they saw value in keeping a record over time of results from Year 8 to Year 10 science.

All head teachers interviewed said they kept faculty records of HSC results over time. No analysis was done by head teachers to find the proportions of students doing senior science courses each year in case study schools before their participation in this project. Most did not have faculty records over time of Y10 grades after external testing stopped. This was not a priority because almost all students went on to Year 11 and many took up senior science courses. There seemed to be little awareness by head teachers that these records provide a basis for monitoring engagement in science (Year 12 proportions relative to English) or of progress in learning (from Year 8 to Year 10).

In relation to monitoring progress in learning from Year 8 to Year 12, doing this was not helped by the fact that Year 8 results are reported against six levels, Year 10 results are reported against five grades and Year 12 results are reported against six levels (not commensurate with the Year 8 levels). The possibility for monitoring student achievement and engagement from Year 8 to Year 10 using VALID10 results is now available to those schools taking up the VALID10 test. K-6 schools taking up the VALID6 option can report their results to the secondary schools receiving their students.

It is also possible that some science head teachers and classroom teachers do not have the statistical skills and / or spreadsheet fluency and expertise to confidently manage the transfer and transformation of the EV data. This was found to be a barrier to meaningful engagement with NAPLAN results for some secondary teachers of English and Mathematics (Pierce & Chick, 2011).

Student EV results are distributed to parents after printing out by the school. Typically, results are sent home in the same way the bi-annual school reports on all courses are distributed. When asked what feedback, if any, was provided by parents to science teachers about the EV reports, none of those interviewed could recall any parent commenting on or asking for more information. This was also true for the two schools that said they handed the reports to parents at their regular parent – teacher night held early in Year 8. When asked why there was no apparent interest from parents in the results, several commented that the time interval between doing the test (November the previous year) and receipt of the report (March-April the next year) may have been a factor, though none suggested how that might have influenced the apparent lack of parent interest. The research literature on the reduced effect of feedback provided well after the assessment was mentioned earlier in this section.

The inclusion of a student survey with the EV test was a unique addition to large scale whole of cohort testing in NSW schools. Only students at schools chosen in national samples to participate in TIMSS testing (in Year 4 and/or Year 8) had completed surveys and tests before EV testing began. Teachers from all three school groups responding to the survey had individually looked at the student survey results (67%. Yet only 20% had discussed the results with colleagues or students. Case study schools said in the interviews that the main reason for not having those discussions was because teachers had not been given support or encouragement to do so. On the other hand, almost all the case study schools said they met regularly as a staff and that assessment was a frequent item on the agenda for those meetings. Had the student survey been of interest or seen as relevant, given the regular meetings, it could have been on the agenda. It would appear that science achievement was of more interest than student engagement with science.

The personal and professional discomfort of teachers to student dissatisfaction is understandable. In recognition of that, EV results are deliberately not publicized in the same way NAPLAN results are. All interviewed said there was no pressure from the school executive over EV results, one way or the other. Whilst this is in keeping with the low-stakes intentions of diagnostic assessment, the main reason the feedback is provided is to promote change leading to better overall levels of student achievement and engagement. There is a strong element of trust being placed in the professionalism of teachers to respond to the feedback. Based on the high level of intention (more than 50% saying they would take up the voluntary Year 10 test), the relative absence of negative feedback about the program (see individual teacher comments above) and the fact that 48% of teachers had used the results to inform changes to teaching and learning programs, that approach by the education system and school managers seems to be sound.

PCWAE1 teachers were aware students at their school did not like science or their experience of it but could not offer any reason apart from reporting a comment from students that teachers at their school were strict about students completing their work. As reported in Chapter Five, students at both PCWAE2 and PCWAE3 had low rankings of their school science experiences as well (see Table K.5A, B and C in Appendix I). That negativity was also reported for their primary school science experience and all three were below the state figures for the proportions including science in their list of three favourite subjects (PCWAE2 was the best of the three there). The apparent paradox of better than expected achievement and dislike for their school science experience will be discussed further in section 6.6.

Of the three fully selective schools, students from MGFSAE2 recorded the least positive views of their school science experience. MGFSAE2's ranking on Items D and E combined was 16th (out of 16). The three selective entry schools' top achievement band students recorded the three highest levels of agreement with Item B which said that science was the hardest subject I learn (1st, 3rd and 2nd respectively for the WAE, AE and WBE schools in that order). A review of the assessment related artifacts provided for the three schools showed that the expectations for knowledge and understanding were well above syllabus expectations which may be a factor contributing to them not enjoying their school science experiences.

In contrast to the above, as shown in Tables K.5D and K.5C in Appendix I, MCWBE5, MCWAE1 and MCWBE4 were at the top of case study school rankings (and above the state) for student enjoyment of their secondary school science experience (student survey Items D and E). The three schools also had the largest proportions of students nominating science as the subject they learnt most in (Item F). EV results for all three of the schools were relatively low (82.54, 82.14 and 73.63 respectively). However, students at MCWBE5 thought science was not as difficult (5th out of 11 non-selective schools and counting the state as one school) as students at MCWAE1 (1st) or MCWBE4 (2nd) did. For students at these two schools, perceived difficulty did not seem to impact enjoyment of their school science experience. Enjoyment of science and / or engagement, as was pointed out in Section 5.4 for all students (three levels of achievement together) at the different schools, was not obviously related to either SEA scores or residual rankings.

Analysis of the assessment narratives for MCWBE5 and MCWAE1, and the schools they were compared with (PCWAE2, PCWAE1 respectively), did not provide consistent, substantive evidence that students at any of the four schools at the end of Year 8 had acquired skills associated with self-regulation. MCWBE4 was not compared to any school and it had a SEA score of 0.7 and a residual of -1.58. Their assessment narrative was more focused on how science contexts were being used to improve student's literacy and numeracy skills and identify formation. A hypothesized link between high achievement and engagement at the end of Year 8 and self-regulation could not be supported. In essence, whilst self-regulation (Boekaerts & Corno, 2005) and learning how to learn (James, 2006), are seen as important, the methods used in this project and related findings did not show a hoped for consistent, pattern that could reasonably be attributed to student self-regulation.

Overall, based on teacher comments in the interviews, students like doing the online EV test which teachers said students find inherently interesting. In only two schools was it suggested that (some) students did not take the test seriously (MGFSWBE1 and MCWBE4). None of the schools reported spending time preparing students for the test apart from the basic requirements to ensure login success and for students to familiarize themselves with how to respond to the items and tasks. The common message given to students was that the Year 8 EV results would not be used in school assessments, but that students should do their best because the test results would help teachers to improve their teaching.

None of the case study teachers interviewed mentioned they had used the teaching strategies advice provided in SMART to address misconceptions identified in feedback to the school. The overall survey response to that question was fewer than half saying 'yes' (one in three WBE and WAE teachers said 'yes'; AE teacher response was two in three saying 'yes'). The provision of this resource in the feedback package was overlooked by most teachers it seems. This researcher's explanation for that is the overall lack of incentive for teachers to engage with the mass of data available in the SMART package. That appears to be the case for WAE schools' low response compared to AE schools where results were perhaps not as good or teachers in those schools were keen to do better for their students. WBE schools had a lower level of engagement for all aspects of the program.

The national tests in Australia for literacy and numeracy (NAPLAN tests) were used in this project to develop predictors of EV success. These tests are examples of summative tests also being used for diagnostic purposes (as well as other purposes discussed in Chapter Two). The anecdotal feedback from the science head teachers in case study schools was that NAPLAN feedback attracts more interest, attention and time from parents, students and their schools' senior executives than does the feedback on EV results. The reasons most gave for the attention to NAPLAN was the publication of the school's results on a well-publicized website for all the world to see (the MySchool website), media interest in comparing schools and the requirement to report NAPLAN results in annual school reports.

In summary, teachers are using or adapting EV test items and tasks from past tests to enhance their science department formal assessment programs (69%). Teachers in schools where results are well above (WAE) or at expectation (AE) are using the resources more and in a wider variety of ways than their colleagues in schools where results are well below expectation (WBE). However, overall, fewer than half (48%) of the teachers that responded to the survey said they were using the feedback from EV results to amend their teaching and learning programs. Teachers in schools where results were as expected (AE) reported the highest 'yes' response rate (75%) to the item about using the feedback from EV results to amend their programs.

6.2.2 Teachers and SOLO

Engagement with the SOLO model was addressed in the online survey in three questions, questions six to eight. SOLO was a key element in the assessment framework for the EV program because it provided the basis for feedback about the level of thinking evident in student responses to items and tasks in the test.

As reported in Chapter Four, the overall finding was that differences between the responses of WAE, AE and WBE teachers on any of the aspects of SOLO engagement investigated here were not statistically significant. Also the overall yes responses to items began at 54% and declined from there to a low of 5% on the second last item in Q6 which was about reporting to home using SOLO.

On Q7, which asked teachers to rate their understanding of SOLO, on a rating scale going from very poor to poor, then acceptable and good to very good, the modal response was "acceptable" (29% chose that option).

When teachers were asked in question eight (Q8) where they learnt most about SOLO, the most commonly mentioned situation was training for marking the EV test or marking EV tests (35%). The next was in EV workshops (9%) followed by never heard of SOLO (7%). It was not possible to distinguish whether the responses were about the Year 8 marking for extended response tasks which is done externally to the school by experienced, trained, science teachers or Year 10 marking. Training for the latter is done at school or home by working through online modules.

Of the sixteen schools visited, only two were actively using SOLO to inform their assessment feedback to students (MGFSAE2 and MCWAE2) at school. MCWBE4 indicated that the school was considering using SOLO as an enhancement to its assessment policies and practices. Neither school used it to report to parents or carers. MCWAE2 recognised its potential to provide feedback to help students with their expressive language skills in science and were using it to mark extended response questions teachers at the school had constructed or appropriated from other sources. The HT at MCWBE3 was actively working on building staff understanding about SOLO in order to use it as the basis for feedback to students in science.

A cogent reason for not using SOLO was given by the two teachers involved in the interview at PCWAE2. They said that students found it confusing to reconcile SOLO and NSW Board of Studies (the Board) provided feedback (reported in levels and grades respectively and based on different criteria as explained in Chapter Two). Given that the school's priority (see their school narrative in Appendix H) to have students do senior secondary science courses, the teachers felt their efforts would be better spent having students understand the Board's *Common Grade Scale* approach to assessment. Head teachers at MCWAE2 and MGFSAE2 who were actively using SOLO to improve student learning in science appeared to have a reasonable understanding of SOLO levels. The head teacher at MCWAE2 was working with the original SOLO taxonomy rather than the version being used in the EV program.

From the above, the SOLO component of the EV program was not very well understood by science teachers responding to the survey and was largely ignored as a basis for providing feedback to students about their level of thinking in science.

The EV program and the SOLO model are exclusively Department initiatives and the above feedback will be of interest to the Department. However, the use of a formal, externally (to the school) developed and imposed summative test to provide feedback to teachers on progress in learning is of general interest to all systems where such testing is done with diagnostic intent. As was discussed in Chapter Two, the last round of PISA testing (2015) in science included a cognitive demand dimension in its assessment framework. SOLO was considered for that role but the test developers chose an alternative, simpler model that recognised three levels of cognitive demand (OECD, 2017). Recognising cognitive demand in the assessment framework of an international test, such as PISA, represents a qualitative improvement in the sophistication of measurement-based, assessment models of which the EV and PISA tests have been described as exemplary (Fensham, 2013).

6.3 Discussion of findings addressing research question two

The question asks: what formative practices are evident in the assessment-related work of science teachers and why are they used or not used?

The focus here will be to look first at case study schools' assessment related-work narratives (provided at Appendix H) for examples of science department practice that reflect formative intentions (6.3.1) before looking at evidence of formative practice in the classroom (6.3.2). In the section on classroom practices, discussion will be linked to the five dimensions of formative practice which comprised the theoretical framework for assessing the extent to which practices were formative.

6.3.1 Science department assessment practices

As was described in Chapter Five, student allocation to classes in the junior secondary years of high school for the purposes of instruction was done in case study schools almost always on the basis of achievement in literacy and numeracy as assessed by teachers at the end of Year 6. The Department's staffing formula provides teachers on the basis that no junior secondary class in the core subjects (which includes science) "need exceed 30 students" (NSW D of E, 2017). In practice however, some classes in a given Year were allocated (with staff agreement) more than 30 students in order to create smaller classes for 'lower ability' students (generally meaning students with poor learning histories). So called 'bottom' classes were generally assigned close to 20 students or fewer if possible. From the perspective of science teachers, the classes assigned to them were "ungraded" in terms of prior science learning. The science head teachers involved in interviews said their expectation was that teachers would work from that assumption. The range of responses by teachers to the diversity of students in their classes is discussed in Section 6.3.2.

It is important to understand that teachers in NSW government schools were required by their employer (the Department) to make use of a specified curriculum and employer-provided policy documents (NSW D of E, 2013) to guide preparation of their teaching and related assessment work. This requirement applied well before the period of interest for this project and continues today. The response by science teachers in case study schools to the above was to use the syllabus and related implementation support and policy advice to guide their construction of a planned program of work for their students mapped to the forty weeks of the school year. Important structural features of the program of work were the curriculum standards described in terms of outcomes and related content to define the scope and level of expected learning related to each outcome. The curriculum expectation was that the learning would be spread equally between knowledge and understanding of science and related contexts and the acquisition of skills related to working and communicating scientifically (BOS, 2003).

In case study schools, the science department's program described for teachers the science knowledge and understandings, skills and attitudes they were expected to "teach" to students in their classes in the four years from Year 7 up to the end of Year 10 and how it would be assessed along the way for the purposes of collecting evidence of learning to be used in preparing progress reports about student learning for parents. There is a requirement to report to parents at least twice a year. The curriculum (called a syllabus in NSW because of its specificity about what was expected to be taught) in place at the time of interest for this project included advice that teaching and learning need to be closely linked, an intention captured in the phrase "assessment for learning" (BOS, 2003, p. 70). To help teachers do that, the Department and Board provide a range of support materials and professional learning activities that teachers can access and work through to devise learning and assessment tasks that better reflect the full range of curriculum intentions and that are fair, valid and reliable reflections of those intentions.

In response to that support, the range of tasks and activities described in the narratives as being used by teachers to collect evidence of learning, apart from pen and paper tests, included student research projects (a mandated activity in the curriculum), field work reports, excursion reports, written responses to laboratory tasks, internet and other text-based research tasks, oral presentations and creative activities such as model making and diary writing, to name some. These activities were either done entirely in class time or in both class and home time. Student responses to the activities provided evidence of learning that was used by teachers for both formative and summative assessment purposes. Whilst a wide range of tasks was being used, what students know and understand as reported to parents (typically expressed as a mark or grade) was dominated by the weight of evidence from tasks returning marks based on teachers' judgments of the quality of expressive language used by students in the construction of responses to the requirements of those tasks.

Accompanying many activities were rubrics setting out the learning expectations and success criteria that would be looked for in assessing the worth of the evidence of learning demonstrated in student responses. The learning intentions, as written down, were typically derived from curriculum outcomes and related content that described the scope and form of expected responses (descriptions including comparison and contrasts, graphic representations with appropriate labels, explanations, justifications and aspects of performances to name some). The judgment to be made of the quality of the response was almost always referenced to the five grades in the Board's *Common Grade Scale* advice (BOS, 2013). SOLO levels and related language only appeared in artifacts provided by two case study schools (MGFSAE2 and MCWAE2).

Of interest was the place AE teachers occupied in the analysis of the responses against the five dimensions of formative practice. The AE teachers are the group of teachers whose students' EV results were as expected based on the predictor. AE teacher responses provided reference levels for this exercise. From the analysis reported in chapter five the frequency means for the AE group of teachers were always between the WAE and WBE teachers. However, for the third dimension (feedback that advances learning) both the WAE and AE means were statistically significantly different to (above) the WBE mean which meant that these two groups of teachers were more frequent users of a variety of feedback sources than were their WBE colleagues.

How teachers engage students with curriculum content and conduct assessments of the extent of learning is a decision for classroom teachers (BOS, 2003). The extent to which the forms of evidence from the surveys and assessment narratives can be said to be formative is discussed below in Section 6.3.2.

6.3.2 Formative classroom practices

Together, the questions and related activities described in the items from the science teacher survey in Questions 9 to 15 address what were called in earlier chapters five dimensions of formative practice. The dimensions, bringing together science instruction and assessment strategies, are:

1. Clarifying and sharing learning intentions and success criteria (LISC);
2. Engineering effective classroom discourse and using learning tasks that elicit evidence of learning (CDEL);
3. Providing feedback that moves learners forward (FTAL);
4. Activating students as instructional resources for one another (including peer assessment) and their teachers (ASIR);
5. Activating students (and teachers) as owners of their own learning (including self-assessment) (ASTL).

These five dimensions provide the framework for assessing the extent to which practices discussed here can be described as formative. As discussed in Chapter Four, the intention was to find out the extent to which teachers were themselves actively using, as well as promoting student agency with, formative practice dimensions. The examples and contexts discussed here relate to Years 7 and 8.

Teacher use of learning intentions and success criteria (LISC)

As mentioned in the findings from the teacher survey (Section 4.2.3.1), students had very little input into the choice of task, learning intentions or success criteria which for the most part appeared to be given determined by the teacher. There were no statistically significant differences between the three school groups when it came to the frequencies with which learning intentions and success criteria were used. Also, the relative frequencies of opportunities for students to take ownership were fewer than teacher led situations.

In the context of teaching and learning tasks, the purpose of tasks was typically explained to students in terms of curriculum intentions. When helping students prepare for assessment, teachers used rubrics to describe features of answers that would attract 'full marks'. Almost all schools provided written rubrics to students to help them understand the criteria that would be used to assign scores. Students typically attempted formal assessment tasks individually and without assistance from others. Their responses were typically scored by teachers either working alone (most often) or shared with other teachers. As will be clear from the discussion following relating to the next two dimensions, the use of LISC to focus discourse and feedback was more frequent in the sample of teachers in WAE and AE schools when compared to the sample of teachers in the WBE schools.

Classroom discourse eliciting evidence of learning (CDEL)

In this dimension of formative practice, statistically significant differences were found between the three school groups relating to teacher-directed classroom discussion. WAE teachers were more frequent users than WBE teachers of wait-time before responding, of discussion about items from tests and assignments and student responses to those items. WAE teachers more frequently asked students to explain their thinking as well as explaining their (teacher) thinking to students. Teachers in WAE schools, in particular, had a strong commitment to developing students' literacy skills and helping students to acquire the scientific vocabulary needed to describe and explain the science in the world around them.

MCFSWAE1 provided a mostly school-based range of science activities focused on laboratory work linked to text-book practical activities and related skills development. The head teacher reported an emphasis on writing explanations as a focus for Year 7 and 8 science. MGFSWBE1 on the other hand, had an emphasis on science process skills but students also worked on projects (involving Visual Arts and Personal Development, Health and Physical Education) as well. Students were also provided with experiences beyond the school gate (an excursion to the Zoo). The girls were given opportunities to discuss science in groups and in whole class discussion and to make models and deliver presentations about what they had learned. The girls at MGFSWBE1 were provided with a more diversified set of science-rich contexts than coed students at MCFSWAE1 and a wider range of experiences in which to explore the meanings of science.

Both PCWAE1 and its paired school, MCWAE1 (pair one) provided students with a range of science rich contexts both in the school science laboratory and beyond the school gate. The teachers worked hard at both schools to fit syllabus intended science learning with contexts relevant to the experience of students. Both schools had the smallest Year 7and 8 classes of all the case study schools. The experiences provided were well used by teachers to develop students' oral and writing skills as well as helping them to acquire the vocabulary needed to describe and explain the science in the experiences provided.

MCAE2 and its paired school MCWBE3 (pair two) provided a range of school-based science laboratory and text-based activities for their students. The focus at both was on developing skills related to scientific investigations in those contexts. MCAE2 engaged its students in a wide range of science projects and it sends the best projects to the NSW Science Teachers Association Young Scientist Awards. Each year it establishes a class of Year 7 students with an interest in science and who have done well in a science-based test set by the school and completed in Year 6.

PCWAE2 and MCWBE5 (pair three) both provide a range of school-based science laboratory and text-based activities for their students. PCWAE2 makes use of a range of agricultural contexts outside the classroom and beyond the school gate to widen the opportunities for its students to engage with science. MCWAE2 provides a range of science rich activities outside the classroom and beyond the school gates to its students as well. PCWAE2 has particular emphasis on developing the literacy skills of its student with a particular emphasis on whole class discussion and reading aloud.

Providing feedback that advances learning (FTAL)

This was another dimension where there were statistically significant differences between the three groups of teachers. In this dimension both WAE and AE teachers were more frequent users than their WBE colleagues of a wide range of opportunities for and sources of feedback ranging from digital polling, to ticks, marks, grades and comments, both encouraging and diagnostic (including the provision of model answers, in terms of success criteria, misconceptions, SOLO levels, elements of the Quality Teaching model, syllabus expectations and Bloom categories); WAE and AE teachers were also more likely to ask their students for feedback on their teaching and to change direction in lessons in response to student feedback.

It was established in Chapter Four that EV results for students in WAE schools were better than comparable AE or WBE school results. The better results indicated that WAE students were more scientifically literate than students in comparable AE or WBE schools. Thus it was no surprise that the dominant theme to emerge from the WAE case study school narratives was the focus WAE teachers had on providing feedback with the explicit purpose of developing expressive literacy skills and student acquisition of science vocabulary related to the science topics being studied at that time.

Activities included requiring students to learn the vocabulary related to the concepts being taught in the current topic (all six WAE schools), by getting them to write extended answers on worksheets with scaffolds and space to write descriptions, comparisons, explanations and justifications (MCFSWAE1 and MGFSAE2). At the end of Year 8, MCFSWAE1 and MGFSAE2 had almost identical result profiles in the four result categories monitored for this project. MGFSWBE1 had a much-reduced top band performance in the extended response category of results than the other two schools (see Table 5.10). The difference there was attributed by this researcher to fewer opportunities being provided to the girls at the WBE school to perform in this way and consequent less feedback to support that way of representing what they knew. Their potential for performing strongly in this way was suggested by the fact that their performance in the communicating scientifically report category was stronger than the other two schools (see Table 5.10).

PCWAE1 students, compared to MCWAE1 students had the better EV result overall and result profile as well, but then almost all their students were from English as a first language background. That was not the case for MCWAE1 students (85% of their students came from language other than English backgrounds and around a third of them were recent refugees with little or no primary school education). That said, the use of feedback to improve learning outcomes at PCWAE1 was outstanding in that it produced a result profile for its students at the end of Year 8 that was better than many schools with higher SEA scores such as MCWAE2, MCWBE3 and MCWBE5 (see Table 5.1 and Table K.1 in Appendix I).

Whole class oral discussion of science contexts and related concepts were explicitly mentioned by all six WAE schools as well. PCWAE1 explicitly referred to pretesting when starting new topics. PCWAE2 provided the most evidence of a differentiated approach to dealing with the diversity of students' literacy and numeracy levels at the time of interest for this project. The feedback provided by PCWAE2 teachers in the context of classroom work was very effective in supporting science learning (as demonstrated clearly in the better result profile for the extended response report category when compared to MCWBE5 which was its paired school (see Table K.1 in Appendix I). By comparison the narrative for MCWBE5 showed an emphasis on process over the acquisition of conceptual knowledge (and related vocabulary). The profiles for working scientifically were very similar. Overall, PCWAE2 had a positive skew in their result pattern; MCWBE5 had a negative skew).

Of interest, as discussed in Chapter Five were the very different levels of student satisfaction with their school science experience. That was recorded by these two schools in their responses to the six items from the student survey. On the combined scores for items D and E (enjoyment of science lessons and science as one of their three favourite subjects, PCWAE2 ranked 14th (out of 16 schools);

MCWBE5 ranked 4th. See Tables K.5A-D in Appendix I for their comparative scores on all six items. Not enjoying science did not deter students at PCWAE2 from achieving highly and nor did it appear to deter their take up of senior science courses, relative to English at their school, the state and MCWBE5 (see Table 5.9). This apparent paradox will be further discussed in the next section, Section 6.4.

A similar outcome for the two fully selective girls schools was in evidence as well. MGFSAE2 outperformed MGFSWBE1 despite the latter having a higher SEA score. On their combined scores for Item D and E, MGFSWBE1 ranked ahead of MGFSAE2 (13th compared to 16th out of 16). MGFSAE2's program was strongly text-based and linked to conventional science laboratory-based skills. The girls at both schools had the 3rd and 2nd highest levels of agreement with the statement that science was the most difficult subject they learnt (MCFSWAE1 was 1st). Their artifacts, when compared to those of the other case study schools, showed knowledge and understanding demands way above the other schools (and for that matter syllabus expectations as well). In the end the take-up of science subjects overall by the WBE school was greater than in the AE school (relative to the state) by a wide margin (see Table 5.11).

Summative assessment at PCWAE2 was much more consequential for students than in other case study schools. Students were moved to a different class at six monthly intervals if performance and achievement was either very good or poor. The reason given for that was to better prepare students for success in senior science courses as a means to the end of obtaining good science-related jobs after school.

Activating students as instructional resources for one another (including peer assessment) and their teachers (ASIR)

There were no statistically significant differences between teachers in the three school groups when it came to activities linked to this dimension of formative practice (see Table 4.15). Teachers in the three school groups had comparable usage frequencies for activities such as collaboratively preparing assessment tasks, marking criteria or rubrics and shared marking (approximately 95% said sometimes or often). When it came to providing students in Years 7 and 8 with opportunities for peer assessment, they were limited and subjective (not well grounded in the language of learning intentions and success criteria). Examples mentioned in the artifacts were not limited to WAE schools (PCWAE1 and MCWBE5 included examples). In terms of frequencies (combining sometimes and often responses to items) for the provision of feedback to peers using success criteria, working in groups on think-pair-share-report activities, writing learning intentions and success criteria, constructing assessment items and tasks, the proportions ranged from 86% being provided with opportunities to use success criteria or assessment rubrics and guidelines to 24% being given the chance to construct assessment items and tasks. A number of schools mentioned that they gave more opportunities for students to provide feedback to each other in Years 9 and 10 (MGFSAE2, MGFSWBE1, MCWBE5 and PCWAE2).

Activating students (and teachers) as owners of their own learning (including self-assessment) (ASTL)

Analysis of teacher survey results for this dimension revealed statistically significant differences in the teacher-initiated aspects of this dimension of formative practice (see Table 4.35). WAE teachers, compared to their WBE colleagues were more frequent evaluators of lessons, keepers of notes on learning issues individual students have, accessors of information about assessment, more frequently engaged with colleagues in activities related to improving personal and shared knowledge about syllabus learning intentions and what progression in science learning 'looks like'. The means between the three groups were not statistically significantly different in terms of the opportunities provided to students to redo work to a higher standard (71% of teachers said they did this sometimes or often), getting students to self-select items for portfolios (30% said sometimes or often) and keeping a journal of reflective writing on science (23% said sometimes or often).

As the above shows, opportunities for students to self-assess were not limited to WAE schools and those opportunities were infrequent. Two examples were recorded in the narratives for case study schools. PCWAE1 provided an opportunity in the context of a toy project and MCWBE5 gave students the opportunity to self-assess against five criteria on a number of tasks.

The anecdotal evidence from interview and artifacts was that where peer and self-assessment were discussed during the interviews, they were opportunities given more to students in Years 9 and 10 than Years 7 and 8 at the time of interest for this project. The same was true for extended groupwork and use of strategies such as think-pair-share-report or jigsaw methods.

The finding that WAE teachers, compared to their WBE colleagues, were more frequent users of a wide-range of activities involving the use and modeling (to peers and students alike) of good learning behaviours was indicative of them 'practicing what they were teaching'. This was most mentioned when it came to staff meetings where assessment-related work was being discussed, when assessment items and tasks were being collaboratively developed or selected, when marking rubrics were being developed and collectively used with each other and students to assess student responses to tasks (see section 4.3.2.5).

Overview of and reasons for using or not using formative assessment

Science teachers in NSW government schools, after a decade of externally provided Year 8 science testing and related feedback on achievement informed by the SOLO model, have not taken up SOLO in a substantial way. The most probable reason for it not being more widely adopted being the requirement to report achievement in terms of grades linked to syllabus standards not, themselves, defined with any reference to the SOLO model. This was explicitly mentioned by PCWAE2 as one reason for not continuing with the VALID10 test after the year of its introduction in 2015.

EV science test results were best in schools where science teachers were more frequent users than their colleagues in other schools of activities related to three dimensions of formative practice. The dimensions were:

- discourse eliciting evidence of learning (second dimension);
- the provision of feedback known to progress learning (third dimension); and
- the use and modeling (to peers and students alike) of good learning behaviours (fifth dimension).

The first dimension about learning intentions and success criteria was being well used by teachers in all schools to guide both instruction and assessment. The language of intentions and criteria were almost invariably derived from the language of outcomes and related content that defined curriculum standards in the official curriculum for NSW schools. Teacher use dominated this dimension. The overall result, relative to the four point scale of never, seldom, sometimes and often, was between sometimes and often as shown in Figure 4.12. Opportunities for students to develop skills in their use was rated between seldom and sometimes, but closer to sometimes.

The fourth dimension of formative practice relates to activating students as instructional resources for each other and their teachers. Student performances provide teachers with feedback they can use to adjust and improve instruction. Providing opportunities for peer assessment is another way of doing that. Teachers overall were evenly distributed in their responses to items related to this dimension by answering from seldom to sometimes (Figure 4.15). There was anecdotal evidence of more frequent opportunities for students to engage in both formal and informal (structured groupwork) peer assessment in Years 9 and 10. However teacher's working together to develop assessment programs, items and shared marking was rated between sometimes and often, but closer to often).

No explicit reasons emerged from the interviews as to why students weren't being given more opportunities to develop the skills of formative assessment in science for themselves. However, a possible explanation may be found in the official science curriculum where the language used to describe skill outcomes for the first two years of secondary science (outcomes 13 to 22 of 22 outcomes) is explicit about the need for teacher guidance. The 'guidance' provided by teachers of students in those years took the form of worksheets that effectively led students from beginning to end of a task (as evidenced in the artifacts supplied). The student research task was in almost all case study schools a heavily scaffolded project telling students what they could and could not research, how to do it and what needed to be included in a written report at the end.

6.4 Discussion of findings addressing research question three

Research question three asks: Does the use of (and if so, how do) formative practices improve students' EV results and later achievement in and engagement with science?

The short answer for schools that self-identified is 'yes' when it comes to the use of EV results. Table 5.1 lists all the schools that identified themselves. Schools are listed in order of the size and polarity of the residual (second last column from the right) from regressing their EV results over an EV result predictor derived from NAPLAN scores as explained in Chapter Three.

The residual is the measure of an effect size of teaching on the EV result as was also explained in Chapter Three. The schools were grouped according to the size and polarity of the residual. WAE schools had residuals that placed them in the top 20% of schools and WBE schools had residuals that placed them in the bottom 20% of schools. In Chapter Four the findings from a survey of teacher assessment related practices were that teachers at WAE schools compared to WBE schools were more frequent users of activities associated with three of the five dimensions of formative practice. The EV results of schools associated with large positive residuals were also schools where science teachers were more frequent users of activities associated with three of the five dimensions of formative practice than their colleagues at other schools with smaller residuals.

Other research discussed in Chapter Three explained that three major factors contribute to the accounted for variability of test results, namely, student socio-cultural background and previous learning history (50%), the actions of their teachers (30%) and school environment factors (20%). As discussed in Chapter

Three, the SEA score for a school is an independent measure of the learning potential students bring to school and this was the basis for creating “comparable pairs” of schools. Because it was impossible to account objectively for the 20% of school environment factors in two different schools, comparing different schools with the same SEA score but widely different residuals provided the best opportunity for confirming that differences in the use of formative practices provide the most likely reason for EV result differences.

On that basis, it was established in Chapter Four that for comparable school pairs, the school with better EV results was associated with more frequent use by teachers of activities associated with three of the five dimensions of formative practice. Those activities were:

- promoting classroom discourse that elicits evidence of learning;
- providing feedback known to progress learning; and
- the use and modeling (to peers and students alike) of good learning behaviours.

The assessment-related work narratives for WAE schools all included strong references to using science contexts for the specific purpose of helping students to acquire scientific vocabulary and the skill to use it appropriately and fluently (orally and in writing). MGFSAE2 also had a high priority for ‘writing’ science. The assessment narratives of the other case study schools gave more prominence to other priorities, such as investigation skills (MGFSWBE1, MCAE2, MCWBE3, MCWBE5) or identity building (MCWBE4). By putting together the analysis of the surveys and a priority for using the language of science, the following picture of formative practice in WAE schools emerged.

Teachers in WAE schools managed classroom discourse that produced evidence of learning (the second dimension of formative practice) that informed teacher feedback (the third dimension) on how well students were doing in using scientific language. Teachers spent a lot of their class-time modelling to students good learning behaviours (the fifth dimension) for acquiring the skills and text types related to scientific literacy, including using prescribed learning intentions and success criteria related to scientific literacy, to self-evaluate. The answer to how formative practices improve EV results rests on the credibility of the claim that the formative use of literacy strategies in science contexts is the most powerful influence on science learning operating in the case study schools. References made to Hattie's (2018) work on effect sizes of different interventions on learning in the opening section of this chapter provides independent confirmation of the power of such approaches.

A hoped-for lasting effect of student exposure to formative practice was their acquisition of the skills and attributes of self-regulated, autonomous learners. This was an expectation based on work reported in the literature review (Chapter Two) linking explicit teaching of the skills of formative assessment to student self-regulation (Black et al., 2006 and James et al., 2007). Ongoing exposure after Year 8 to higher frequency teacher use of formative practice and, perhaps, acquisition of student self-regulation, in that light, should continue to produce better results for those students at the WAE school (at both Year 10 and in senior science courses). Also, the expectation was that higher proportions of students would be completing senior science courses than in their paired school. That legacy may be the explanation for better later achievement and higher engagement.

However, in relation to the later achievement and engagement part of research question three, analysis reported in Chapter Five of data from assessment-related work narratives associated with the case study schools was not a sound basis for making any claims about an ongoing effect. The correlation between the measure of Year 8 engagement and Year 12 science course completions inconclusive. However, the correlations between Year 8 achievement and Year 12 science course completions at a school was persuasive for the case study schools. Given the unreliability of comparing Year 8 and Year 10 results and absence of a persuasive supporting correlations between Year 8 engagement and Year 12 engagement, the acquisition of self-regulation by more students in high residual schools compared to other, lower residual schools as proposed here could not be justified.

From the above discussion of evidence, it is reasonable to claim that teacher use of formative practices helped students to achieve better results in science at the end of Year 8. An additional conclusion that students exposed to those practices had acquired the skills of self-regulation as a consequence of exposure to those practices could not be supported by the available evidence.

6.5 Suggestions for further research

Given the importance of producing students who are self-regulated, autonomous learners by the time they leave school, further studies using the research design at the core of this project is warranted by the findings reported in this thesis. The use of reliable, comparable data on achievement and engagement after Year 8 to investigate the worth of teaching formative assessment strategies to students may be worthwhile. Additional research to that end is discussed below.

Provincial students apparent low regard for science

The findings reported in Chapter Five add weight to concerns already expressed by other researchers who have reported similar findings from their research for provincial students in the early years of secondary schooling. Lyons and Quinn (2010, 2012, 2014) confirm that Australian provincial school students' negative attitudes to science relative to their metropolitan counterparts persist up to Year 10. The researchers could only speculate as to the reasons for that negativity but did see it as a barrier to be overcome (curriculum mismatch with student experience, a shortage of specialist teachers and lack of perceived relevance were some of the possibilities they listed). Tytler and Symington (2015) writing in *Teaching Science* list other researchers who reported similar findings.

As mentioned in the opening paragraph of this chapter graduating students who know how to learn is important. That being so, then it is important to find out why provincial students don't enjoy an experience that many are clearly doing well at (after taking into account their lower literacy and numeracy levels compared to their metropolitan counterparts) is also important. If provincial students both understand why they are doing better than expected (by acquiring the fluency with and control over the language of science at the very least) and feel they are doing better than expected that might provide the motivation for even more students to take up science in the senior years. Palmer (2015) identified student enjoyment of science as a reason for taking it up in the senior years of schooling and hopefully beyond into preparation for a STEM career.

A first suggestion for future research in this area may be to try and understand why provincial students have a less positive view of their school science experience than their metropolitan counterparts. An initial project might undertake a full analysis of the student surveys for the case study schools in this project. The Department has that data from 2005 up to the present time. At the very least, it may provide a more nuanced understanding of students' views about their experience of science at school and additional clues as to why they don't like science. School factors external to the science classroom may be a contributor, but the consistency of the low regard by students in provincial settings (all three of the provincial WAE schools in this project) may well have more to do with parent socio-cultural dispositions that accord a lower value to science in those communities than elsewhere. A related question to explore would be why provincial students take up senior science courses when they clearly do not like the subject.

Student backgrounds, according to Hattie (2003b), are responsible for up to half the accounted for variation in test results. The suggestion that these values may be implicated comes from the finding that students at the three schools also recorded low enjoyment of their primary science class experiences (Item C in the student survey). Top band achievers at the three provincial schools ranked their experience at primary school years science experiences at 12th (PCWAE1), 7th (PCWAE2) and 16th (PCWAE3) out of the 16 schools compared here (the state result was counted as a school in Table K.5B in Appendix I).

The importance of self-regulation

Given the growing importance of producing self-regulated, autonomous learners as a valued outcome of schooling it may be useful to confirm whether putting more effort into explicitly teaching students the strategies of formative assessment is the most effective way of doing that. It may be that other teaching approaches can do the job more effectively. Two approaches have already shown promise in that regard. The first is inquiry, the second is problem solving. Their importance as aspects of science education is reflected in the working scientifically and communicating scientifically EV reporting categories respectively.

The assumption is that pen and paper tests are able to provide sufficient valid, reliable and authentic evidence of the attributes of self-regulation and learning autonomy. The methodology described in Chapter Three could just as easily be applied to exploring whether, for example, inquiry or problem-solving approaches would be a more effective means to that end. Teacher surveys designed to characterise teaching that reflects best practice in teaching inquiry and problem solving may be substituted for the formative practices survey used in this project. An appropriate set of interview questions could be developed, artifacts collected and related narratives generated to examine for corroboration of findings.

Another approach to investigate is that of representational pedagogies which were speculatively posited as a "signature pedagogy" for science by Tytler, Prain, Huber & Waldrip (2013). Research papers already published could provide the activity descriptors (such as the one on forces espoused by Huber, Tytler and Haslam (2010)) with which to generate survey items that could be piloted with schools known to be early adopters of these pedagogies.

Representational pedagogies are essentially formative because they shift the emphasis from what the teacher is doing to what the student is doing. That approach to teaching engages students in creating representations of what they are learning and challenges students to test the limits of their explanatory power. The representations produced may be in a variety of forms such as diagrams and 3-D models, written texts, presentations using ICT and including audio and video content or any combination that is deemed appropriate for purpose and audience. Curriculum intent, pedagogy and assessment are evaluated for alignment by all participants in the back and forward negotiation of meaning. Representational pedagogies may well be a more effective way to produce students who are self-regulated and autonomous learners than the current approach in the UK to explicitly teach students the strategies of formative assessment (James, 2006).

Confirmation of findings from the initial research project

As mentioned above, the introduction of the VALID10 test provides standardised achievement results for students in all participating schools at the end of Year 10. The school sets of science results can be used to evaluate the prediction that in pairs of comparable schools, the school with the higher residual will continue to produce better results at the end of Year 10 and again in science subjects at the end of Year 12. Data sets for this project should be available from the Department for cohorts of students doing VALID8 (beginning) in 2015 (based on the new national curriculum), VALID10 in 2017 and Y12 results from 2019.

Smaller studies may choose to test the validity and reliability of aspects of the methodology used in this project. Given the closeness of the coefficients of determination for the four predictors used, it might be simpler to use the Year 7 NAPLAN reading results on their own as the basis for the predicator used in the regression analysis without serious loss in the integrity of the findings.

The teacher survey instrument would benefit from including a wider array of strategies that may be being used by teachers to enhance student agency as autonomous learners. Hattie (2018) has a list of 33 strategies under the heading of Strategies emphasizing student meta-cognitive / self-regulated learning. Also, existing PEEL resources (Mitchell et al., 2009) could be accessed for appropriate "good learning behaviours" (p. 172). Procedures could be tested for recognition by teachers and selected for inclusion. A strategy not on Hattie's list is the Predict-Observe-Explain sequence (White & Gunstone, 1992). The expanded item set so produced could be added to the teacher survey for a repeat of the original study along with the enhancements mentioned above.

An additional enhancement would be to include interviews with and artifact collection from Year 10 teachers responding to a science teacher survey. The survey should be the same for both Year 8 and Year 10 teachers.

This project used the average of four consecutive years of standardized residuals as the basis for choosing maximum variation cases (Flyvbjerg, 2011). In a future study, researchers could look for schools where the residuals were increasing over the years; were declining over the years and look for changes in the assessment-related work narratives for those schools in a before and after study.

Evidence gathering, apart from the methods described above, could be expanded to include classroom observations (recorded by people, audio and or video technology) of teacher enactments of target strategies and student responses to them. These observations could be used to corroborate teacher responses to surveys and used to confirm the fidelity of strategy interpretation.

Given the above suggestion that community valuing of science may be a factor inhibiting student engagement with science, it may be useful to have samples of parents respond to appropriate items from the current student survey in the first instance. Their responses to the same (or tested equivalent items) may provide insights into the source of student attitudes, particularly if students and their parents independently complete the survey and their responses matched and compared with their child's responses. In the event that this does not provide the needed insight, a wider range of questions about science may be helpful. To that end, Barry Fraser's (Fraser, 1978) *Test of Science Related Attitudes* (TOSRA) survey might be a good starting point.

6.6 Recommendations

This section provides recommendations to the Department, the NSW Educational Standards Authority, the Australian Curriculum Assessment and Reporting Authority and a wider audience of educational researchers with an interest in the theory of formative assessment, its integration with instruction (formative practices) and its potential for guiding students to learn how to learn. The recommendations are supported by the findings reported in this thesis.

The interest of a wider audience of educational researchers is predicated on their prior interest in testing the power of formative assessment to improve student achievement in and engagement (especially beyond school) with science. In that light, it is hoped that some researchers might be prepared to undertake further work along the lines suggested in the previous section to add more weight to the body of research supporting the power of formative practice to improve achievement in and engagement with science.

One of the claims for importance of this research is the methodology developed by this researcher to isolate the contribution of teaching from other contributions to a test result. Here it was used to separate the contribution of general literacy and numeracy skills from the scientific literacy component in a science test result. The scientific literacy component is what students have learned in the context of their science lessons. Other researchers might be interested to use it and confirm its utility in other learning areas apart from science, such as geography or history.

The latest round of PISA testing completed in 2015 (see Chapter Two) emphasized that providing feedback on the level of thinking demonstrated in student responses is useful because it differentiates between recall of an attribute of one science concept and being able to “relate and evaluate many items of knowledge” (OECD, 2017), p. 40). Demonstrating the latter in a science context is arguably a higher value response in the context of assessing competence in scientific literacy. To do so requires a student to use more cognitive resources than the recall of a single attribute. To be scientifically literate, as the PISA framework specifies, requires students to operate at a level where they can relate and evaluate more than one item of knowledge. Including cognitive demand as a dimension in the assessment framework enhances the validity of the test construct and results from it (Messick, 1995; Mislevy, 2008).

After considering a number of schema’ to operationalize the construct of cognitive demand, including the Biggs and Collis (1982) SOLO taxonomy, the developers of the PISA test adapted Webb's four level Depth of Knowledge model (Webb, 1997) to that end.

The OECD-PISA decision to include cognitive demand as an aspect of competency in scientific literacy vindicated the decision by the Department to include a cognitive demand dimension from the outset (2005) in its assessment framework for the EV program. The Department used the SOLO model developed by Pegg, Panizzon and others at the University of New England (Panizzon, 2003). The SOLO model was an evolution of the SOLO Taxonomy first published by Biggs and Collis (1982; 1991) and was described in Chapter Two. Given international support for and acceptance of cognitive demand as an explicit enhancement to the PISA assessment frameworks, it would be a pity to discontinue the one large scale assessment project in Australia where it is a feature.

At this time neither the NSW curriculum authority (NESA, 2017) nor the ACARA published achievement standards for the current national *Australian Curriculum: Science* (ACARA, 2018) include either explicit or implicit recognition of cognitive demand as a dimension in their definitions of competency and related assessment support materials or advice. In the interests of improving the validity of assessment and ensuring on going alignment between curriculum intent, related instruction and assessment validity as discussed in the NRC (2001) report, ACARA might want to consider how future iterations of the science curriculum (at the very least) respond to PISA leadership and include references to different levels of complexity in thinking in its *Science Sequence of Achievement* descriptions (ACARA, 2018). Rather than drop SOLO for Webb's model as used by PISA, externally designed, large-scale tests could look at using SOLO as their model because of its historical prior use in science education and elsewhere in Australasia.

The two reports from the last round of Year 6 NAP testing in Science Literacy have dropped the Appendices carried by successive reports up to and including 2012 where the connection between SOLO and that test was explained (ACARA, 2017). SOLO was used as a classifier for items produced by ACER to populate the data base of science assessment items in the context of the national *Science Education*

Assessment Resource project (ACER, 2004a). The current version of that resource is being managed now by Education Services Australia (ESA, n.d.). SOLO is also used to inform feedback to primary teachers in New Zealand who have used items from the, so titled, e-asTTle data base of items to assess student achievement and progress in reading, mathematics and writing (Hattie & Brown, 2004).

The presumption of the Department's interest is based on the fact of their tangible support for this project by providing this researcher with access to EV test data and related statistical analysis. The recommendations (below) for change in practice are an expected outcome from using a transformative mixed methods research design in this project (Creswell & Plano Clark, 2011).

One part of the rationale for the Department continuing with the VALID program is that it includes the dimension of cognitive demand in its assessment framework and has done so from its inception in 2005. Including cognitive demand in the assessment framework of tests improves the validity of the test as discussed earlier. The inclusion of cognitive demand in the OECD-PISA test assessment framework is vindication of the Department's earlier decision to use SOLO as the basis for measuring cognitive demand in its EV test. A second is the endorsement of the EV test provided by Fensham (2013) who says that the [EV] test development process is comparable to the PISA and TIMSS development processes. In the same book, a chapter by Miller (2013) argues that assessment models are an important complement to curriculum documents because they help teachers to operationalize curriculum standards and show how best to assess curriculum intentions.

Two components of the current EV test design are singled out for further comment. The first are the three extended response items included in each test. The extended response tasks model open-ended questions that enable students to respond, using written text, at the highest levels of thinking they are capable of. The capacity to write scientific explanations is a highly valued outcome of science education which some, at least, of the case study school participants explicitly acknowledged. Inclusion of the extended response items in the test signals to science teachers and students the importance of this skill. The evidence reported in this thesis showed that the results of students exposed to explicit teaching of literacy strategies in formative ways leads to better than expected results.

The second is the use of science rich stimulus material drawn from the wider reading and Internet experience of students as contexts for science questions is an important signal to students and teachers of the relevance of science for dealing rationally with the world. It also provides opportunities for item construction that provides higher levels of cognitive challenge to students in a form that is an authentic test of aspects of scientific literacy (see above in the discussion of the PISA rationale for including cognitive demand in its assessment framework). With some modification, items and tasks could easily be amended to provide for a wider range of responses than written texts alone. This will be important once representational pedagogies and other more progressive approaches become more widely used. The capacity to upload video and sound as well as photos and diagrams should be considered in addition to the construction of written texts now that the test is delivered online. In the context of a test, the set of items a testee is provided with can be changed to better meet their demonstrated ability (as assessed by the software managing the item set being delivered to the testee as they do the test).

The capacity to upload a wider range of responses to items and tasks would be made easier by transforming the once-a-year test to an online repository of items, related stimulus materials and extended response tasks from which teachers could choose. They could retain and store items online until they enabled access for their students as they work through the topic or at the end or both. The capacity for immediate feedback on their learning, this being one of the most powerful means for supporting learning, would then be provided. There are already a number of items (and related stimulus material) and extended (open-ended) response tasks going back to 2005 held by ESA (SEAR, 2004) that could be used to populate such a repository.

Online availability of assessment items and tasks has a number of potential advantages which include the capacity to:

- provide immediate feedback to teachers about student experience of science (using items from the current student survey);
- provide a brief description of item and task links to curriculum intentions;
- information about the level of cognitive demand of the item or task and possible real-world situations where engaging with the particular item and its stimulus material or task has benefits for the individual, society or the environment;
- provide explanations of alternative conceptions indicated by student selection of particular distractors (in multiple choice items) in feedback to students (and teachers);
- suggestions for activities to correct misconceptions (already provided in SMART for the current version of the EV tests);
- provide a range of answers that would be scored at different levels according to the SOLO model (for extended tasks only); and
- the history of item and task use and student answers could be retained online and made accessible to both teachers and the education system for monitoring purposes.

The last point would enable stronger measures of item reliability and difficulty (psychometric data) to be confirmed over time as well as enabling monitoring of change in the quality of learning over time by both teachers and the system. Also, transparency about the uses of that data would need to clearly provided and agreed to by all participants.

Student self-assessment is seen as an important skill for students to acquire in the context of becoming autonomous learners (Black et al., 2006). With that in mind, ways for direct access by students to a future repository of assessment items and tasks should be developed and trialed. In this scenario, students would be able to select and complete items and obtain immediate feedback on their responses. Student access could be managed in a way that protects the integrity of the items, related stimulus material and extended response tasks but enables the data on responses to be generated and retained. Student responses should also be retained for teacher access as well to enable them to evaluate how learning is progressing as it happens. This would provide the opportunity for teacher interventions based on what they see happening online as students engage with the material there.

At the present time, EV data is provided to NSW schools and not published in the same way as NAPLAN data (on a school-specific website for all the world to access). The findings reported in Chapter Five were that science teachers understood the purpose of the EV test, were willing to engage with it and feedback from it and appreciated the absence of pressures experienced by their colleagues more directly associated with the publication of NAPLAN results. It is strongly believed by this researcher that shifting the items and tasks into an online repository accessible as discussed above would increase usage because the feedback would be immediate and thus most useful to teachers and students (Black, 2007; Hattie & Timperley, 2007; Shute, 2007). Delay in receiving feedback was identified by teachers involved in this project as a disincentive to greater engagement with the EV program.

In the event that public accountability is seen as important, consideration could be given to sample testing along the lines of the current NAP program for Year 6 science or simply continue using the current program of TIMSS and PISA testing which Australian students have been doing for the past two decades already. Using only the international tests would avoid duplication and free up resources for other purposes.

6.7 Conclusion

Section 2.2 of this thesis outlined the gap between ideal and actual practice found by Goodrum et al. (2001) in their review of science teaching and learning in Australia at the end of the 20th century. The three researchers in their report drew attention then to the strong emphasis, particularly in secondary schools, on summative assessment and the negative impacts (not just in Australia), it was having on science teaching, on achievement and on engagement with school science (see Table 2.1). The writers recommended greater alignment between syllabus intentions (outcomes that focus on scientific literacy), instruction and assessment. Assessment, they said, should be used more to support instruction as it was happening (formative assessment).

This thesis reports on the impact of two initiatives designed to help teachers shift their assessment focus from summative to formative. The initiatives were in response to the 2001 report by Goodrum et al. Data for the thesis was collected in 2016 and covered school years 2010 to 2015. The first initiative was in the form of curriculum advice for teachers about assessment for learning (an alternative name for formative assessment). It was promulgated in the new science syllabus which was introduced from 2003 (BOS, 2003). The second initiative was a large scale, diagnostic science test and student survey at the midpoint of a mandatory, four-year secondary science course. The test was piloted in 2005, trialed in 2006 and implemented across the state of NSW for all Year 8 students from 2007. The test gathered evidence of student learning relative to syllabus standards (described as outcomes). The survey gathered evidence of student understanding about science in the world and about their experience of science in the school setting.

Parents (and their students) received a progress report about their learning in terms of both syllabus expectations and level of understanding demonstrated in relation to those expectations. The levels were referenced to the six levels of understanding described in the SOLO model. Teachers received a comprehensive analysis of individual performance on every task and item in the test as well as students' collective views about science and their experience of it at school. Teachers were expected to use the results of the test and the survey to diagnose strengths, weakness and gaps in student learning (and level of engagement with learning science) and to respond accordingly.

Impact of both the curriculum advice on assessment for learning and the EV program on teachers' assessment-related work was explored against the five dimensions of formative practice described in Chapter Two. The evidence from analysis of the teacher survey responses revealed that fifteen years after the

Goodrum et al. (2001) report, instruction and assessment were more aligned to curriculum expectations (described in terms of outcomes) than was the situation in 2000 (first dimension of formative practice). This was a consistent feature of teaching across all three groups of schools, regardless of whether EV results were well above (WAE), at (AE) or well below expectation (WBE).

However, in schools where results were WAE and AE, the teachers there were more frequent users of discourse eliciting evidence of learning (second dimension of formative practice) and providers of feedback that advanced learning (third dimension of formative practice) than were their colleagues in schools where EV results were WBE.

When it came to providing students in the first two years of secondary school with opportunities to take the lead as instructors for each other, none of the three school groups stood out for doing so (fourth dimension of formative practice).

In schools where results were WAE, teachers there were more frequent demonstrators of good learning behaviours both with students (and with each other) than were their colleagues in either AE or WBE schools (fifth dimension of formative practice).

In WAE schools teachers also focused on developing students' capacity to use the language of scientific literacy appropriately. This was most evident in the WAE / AE – WBE comparisons of schools with comparable socio-educational advantage (SEA) scores. In those comparisons, WAE / AE schools had larger proportions of their students in the top band for the extended response category of results (PCWAE2 and MCWBE5, MCAE2 and MCWBE3 and MGFSAE2 and MGFSWBE1).

Because it was not possible to ensure the comparability of Year 10 results across schools nor to be sure that the proportions of students doing senior science courses was a direct reflection of student demand (rather than school resource limitations), three predictions developed as indicators of student self-regulation could not be reliably verified. The only other independent measure of self-regulation available to this project (students reporting in the survey a positive school science experience at the end of Year 8) was not consistently found in high achieving case study schools. In fact, high achievement was consistently linked to low ratings by students of their school science experience. This was very evident in the three provincial case study schools.

The combination of teacher survey results and assessment narratives supports the conclusion that in case study schools at least, teachers retain strong control over the activities associated with formative practices, at least up to the end of Year 8. Whilst this is associated with better than expected scientific literacy outcomes, students in provincial schools in particular do not appear to be enjoying the experience. This was in contrast to two coeducational metropolitan case study schools, also with relatively low SEA scores (MCWAE1 and MCWBE4) but with high ratings of their school science experiences.

At the end of Year 10 both MCWBE5 and PCWAE3 have better result profiles than PCWAE2. All three schools have comparable SEA scores. This is a reversal of the Year 8 position. Despite the uncertainty around the comparability of the actual results, the distribution of the results across the grades is telling. It seems that the rigorous application of a summative assessment policy at PCWAE2 may be a contributor to the decline in achievement from Year 8 to Year 10.

Taking all the above into account it is the view of this researcher that progress is being made toward helping students acquire the tools needed to manage their own learning, as the focus on mastering the language of science has shown. However, the broadening of that to encompass the full meaning of being scientifically literate (OECD, 2017) will require that students are explicitly taught the skills of formative assessment and given opportunities to use them at school. This will only happen when the community accepts the validity and reliability of evidence of learning obtained by means other than pen and paper tests (or their on-line equivalents).

APPENDICES

Appendix A: Competencies, Basic Skills, Generic Skills and Key Competencies

Table 2.1

Competences, basic skills, generic skills and key competencies

SECTION ONE: Quality Education Review Committee (QERC,1985) general **competences** and basic skills

1. Acquiring information;
2. Conveying information;
3. Applying logical processes;
4. Performing practical tasks as individuals;
5. Performing practical tasks as members of a group (Recommendation1, p.201).

Basic skills in (the curriculum) including:

- communication skills;
- Mathematics;
- Science;
- Technology;
- the world of work; and
- Australian studies (Recommendation 10, p. 203)

SECTION TWO: Australian Education Council Review Committee (AECRC, 1992) Finn review **generic skills**

1. Language and communication;
2. Mathematics;
3. Scientific and technological understanding;
4. Cultural understanding;
5. Problem solving; and
6. Personal and interpersonal understanding.

SECTION THREE: Australian Education Council Review Committee (AECRC, 1992) Mayer **key competencies** (NSW version)

1. Collecting, analysing and organising information;
2. Communicating ideas and information;
3. Planning and organising activities;
4. Working with others and in teams;
5. Using mathematical ideas and techniques;
6. Solving problems;
7. Using technology; and
8. Cultural understanding*

SECTION FOUR: Science 7-10 syllabus (BOS, 2003)

Key Competencies are embedded within the objectives and content of the Skills. The content develops students' ability to:

1. plan, organise and perform first-hand investigations to test a hypothesis or question that can be researched;
2. collect, analyse and organise information from first-hand investigations and secondary sources, organising data using a variety of methods including diagrams, tables and spreadsheets, and checking reliability of gathered data and information by making comparisons with observations or information from other sources;
3. communicate ideas and information using a range of text types including explanation, procedure and report formats to present data and information from first-hand investigations;
4. identify the nature of issues and problems, framing possible problem-solving strategies and developing creative solutions in a logical, coherent way;
5. use technology including CD-ROMs and the internet to access information
6. work individually and in teams where appropriate, safely, responsibly and effectively with realistic timelines and goals; and
7. use appropriate mathematical processes including appropriate units, graphs, spreadsheets and mathematical procedures and relationships.

*This was a NSW addition to the list

Source: report documents as listed in the Table (see reference list).

Appendix B: Goals for Schooling (1989 – 2008)

Evolution of Australia's Common and Agreed National Goals for Schooling in the Twenty First Century.

Hobart Declaration on Schooling (1989). The ten goals…

1. To provide an excellent education for all young people, being one which develops their talents and capacities to full potential, and is relevant to the social, cultural and economic needs of the nation.
2. To enable all students to achieve high standards of learning and to develop self-confidence, optimism, high self-esteem, respect for others and achievement of personal excellence.
3. To promote equality of education opportunities, and to provide for groups with special learning requirements.
4. To respond to the current and emerging economic and social needs of the nation, and to provide those skills which will allow students maximum flexibility and adaptability in their future employment and other aspects of life.
5. To provide a foundation for further education and training, in terms of knowledge and skills, respect for learning and positive attitudes for life-long education.
6. To develop in students:

   a) the skills of English literacy, including skills in listening, speaking, reading and writing;

   b) skills of numeracy, and other mathematical skills;

   c) skills of analysis and problem solving;

   d) skills of information processing and computing;

   e) an understanding of the role of science and technology in society, together with scientific and technological skills;

   f) a knowledge and appreciation of Australia's historical and geographic context;

   g) a knowledge of languages other than English;

   h) an appreciation and understanding of, and confidence to participate in, the creative arts;

   i) an understanding of, and concern for, balanced development and the global environment; and

   j) a capacity to exercise judgement in matters of morality, ethics and social justice.
7. To develop knowledge, skills, attitudes and values which will enable students to participate as active and informed citizens in our democratic Australian society within an international context.
8. To provide students with an understanding and respect for our cultural heritage including the particular cultural background of Aboriginal and ethnic groups.
9. To provide for the physical development and personal health and fitness of students, and for the creative use of leisure time.
10. To provide appropriate career education and knowledge of the world of work, including an understanding of the nature and place of work in our society.

Adelaide Declaration (released in 1998)

The achievement of Australia's common and agreed national goals for schooling establishes the pathway for lifelong learning, from the foundations established in the early years through to senior secondary education including vocational education and linking to employment and continuing education and training.

Schooling should develop fully the talents and capacities of every student. In particular, when students leave school they should:

- have skills in analysis and problem solving and the ability to become confident and technologically competent members of 21st century society.
- have qualities of self-confidence, optimism, high self-esteem, and a commitment to personal excellence as a basis for their potential life roles as family, community and workforce members.
- be active and informed citizens with the ability to exercise judgement and responsibility in matters of morality, ethics and social justice; and the capacity to make sense of their world, to think about how things got to be the way they are, to make rational and informed decisions about their own lives and to collaborate with others.
- have a foundation for, and positive attitudes towards, vocational education and training, further education, employment and life-long learning.

In terms of curriculum, students should have:

- attained high standards of knowledge, skills and understanding through a comprehensive and balanced curriculum encompassing the agreed eight key learning areas: the arts; English; health and physical education; languages other than English; mathematics; science; studies of society and environment; technology

  and the interrelationships between them.
- attained the skills of numeracy and English literacy; in particular, every child leaving primary school should be numerate, able to read, write, spell and communicate at an appropriate level.
- been encouraged to be enterprising and to acquire those skills which will allow them maximum flexibility and adaptability in the future.

In addition, schooling should be socially just, and should ensure that:

- outcomes for educationally disadvantaged students improve and match more closely those of other students.
- Aboriginal and Torres Strait Islander students have equitable access, participation and outcomes.
- all students have understanding of and respect for Aboriginal cultures and Torres Strait Islander cultures to achieve reconciliation between indigenous and non-indigenous Australians.
- all students have the knowledge, cultural understandings and skills which respect individuals' freedom to celebrate languages and cultures within a socially cohesive framework of shared values.

MCEETYA (2008) Melbourne Declaration (December 2008)

The Educational Goals for Young Australians

Goal 1: Australian schooling promotes equity and excellence

Goal 2: All young Australians become:

– Successful learners

– Confident and creative individuals

– Active and informed citizens

Source: Hobart and Adelaide Declarations – MCEETYA, 1998 / Melbourne Declaration – MCEETYA, 2008. See reference list for full citations.

Appendix C: A teaching sequence exemplifying different views of learning

A view of learning that clarifies curriculum intention, guides instruction and shapes assessment.

The following illustrates how a view of learning and cognition can clarify curriculum intention, guide instruction and shape assessment. The 2003 NSW Science Years 7-10 syllabus (BOS, 2003) requires students to use a particle model to explain change of state (outcome 4.7.1, 2 & 3: particle model, change of state on p. 32). A teacher who is familiar and comfortable with a cognitive, constructivist view of cognition and learning might take students through a teaching sequence that ends with an assessment task.

This sequence might involve students setting up situations involving water where they can observe evaporation, boiling, melting and freezing. In that scenario, the teacher moves around the room providing advice and support as students work through the activities. This is followed by a teacher-led explanation of the particle model and a discussion of how it can be used to explain each of the examples the students had worked with. Students are then given a pen-and-paper task that has short response items, an extended task of a simulated experiment involving ice melting, and a set of questions asking students to create labelled diagrams using particles to represent the change from ice to water to steam. When marked, the teacher would lead a discussion of the results with the class. This description is a truncated version of the 5Es approach (AAS, 2017), although the outline given here is not from the 5Es materials.

A teacher who is familiar and comfortable with a situative view of learning might see an opportunity to address two other syllabus (BOS, 2003) requirements at the same time as teaching the particle model. Syllabus outcomes related to the nature and practice of science (outcome 4/5.2a to evaluate the role of creativity … in describing phenomena on p. 28) and working in teams (outcome 4/5.22.2 to practice aspects of team work described in content items a to h on p. 44) are outcomes that lend themselves to groupwork. The teaching sequence might involve a cooperative learning strategy, such as the jigsaw technique (Mitchell et al., 2009, pp 75-76) to engage with all three outcomes. The assessment might involve a presentation by each member of the group (individual or group role-playing water particles and moving in ways that simulate evaporation, melting, boiling and freezing followed by a class Q & A led by performers).

Team members could also be asked to complete a checklist identifying aspects of teamwork for themselves (self-assessment) and other members of the team (peer-assessment) in terms of their own contribution to the preparation and delivery of the content in the presentation. The test could also be done by individuals. When completed, the teacher would provide feedback to the group drawing on the evidence of learning from her/his observations, a reading of the checklists, and the test results.

Assessment involves both pen-and-paper responses and observations of performance as evidence of learning. Feedback can be given in terms of both the particle model of matter and the processes involved in preparing and delivering the performance.

Appendix D: Five examples involving aspects of the SOLO model

Example one-heating ice

An example from the 2005 EV pilot test shows how the two-learning cycle SOLO model was applied to code an extended response task. The students were presented with a diagram of a beaker containing ice, a thermometer and a stirrer. The beaker and contents were sitting on a gauze mat and retort stand with a Bunsen burner under it running a low, two-zone flame (this equipment is ubiquitous still in NSW government school science labs). Students were presented with a table of results (Table 2.6) showing temperature change over time (from 0 to 9 minutes)

Table 2.6

Table of results from heating ice

Time (in minutes)	0	1	2	3	4	5	6	7	8	9
Temperature (in ℃)	1	1	1	4	15	29	45	61	75	89

Source: ESSA 2005 test booklet, NSW Department of Education and Training

Students were asked:

(a) Using the information from the result table (Table 2.6), describe what was happening in the first nine minutes of the experiment.

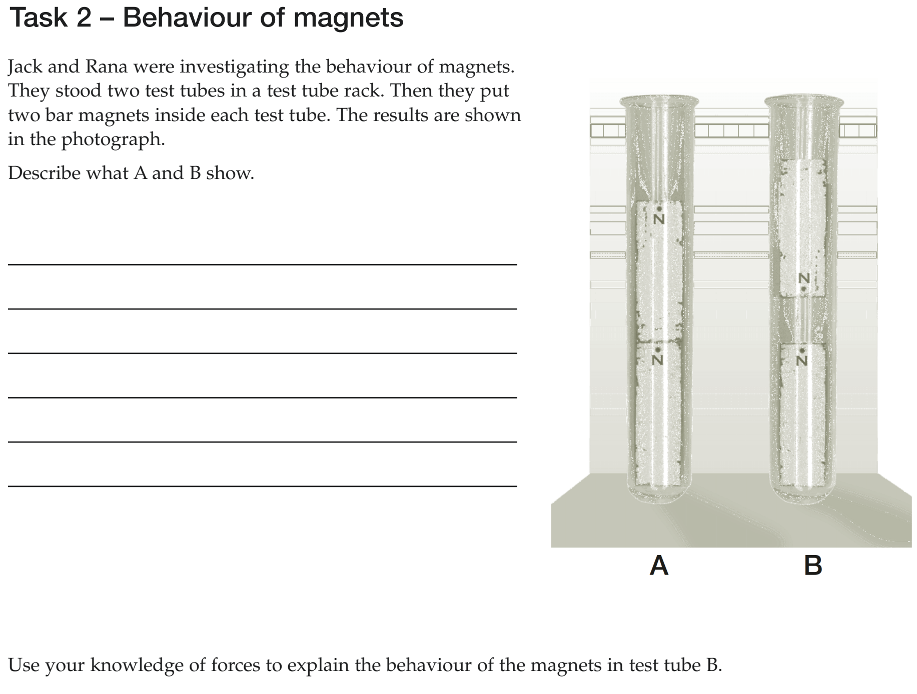
(b) Using your knowledge of the particle theory, explain why this happens.

For the first cycle any ONE of the following were accepted as a U_1 response: change of state / the ice melts / describes a part or all of the trend changes in temperature over time with or without specific reference to time intervals. An M_1 response involved TWO or more U_1 responses being provided. R_1 responses linked the trend change in temperature to an inferred melting of all the ice. Note that in cycle one, the responses made no reference to science concepts. Responses were in terms of everyday language related to the relevant observations.

In the second cycle any ONE of the following were accepted as U_2 responses: heat increases the movement or vibration of particles / heat is absorbed by ice particles as the ice melts / heat energy breaks down forces of attraction between particles of ice [and] the particles spread apart. An M_2 response involved at least TWO U_2 elements. R_2 responses needed to provide a cause and effect explanation for change of state and subsequent rise in temperature (using moving particles and heat or energy appropriately in the explanation).

The moving particle theory / model of matter is a syllabus expectation for explaining physical changes such as change of state (BOS, 2003). Clear references to syllabus concepts and related explanations using those concepts were needed to score in the second cycle. Note that there was no expectation of a reference to latent heat, but an operational definition in terms of heat absorption without temperature change until all the ice had melted was an accepted inference from the data provided.

Example two-Behaviour of magnets task



Task 2 – Behaviour of magnets

Jack and Rana were investigating the behaviour of magnets. They stood two test tubes in a test tube rack. Then they put two bar magnets inside each test tube. The results are shown in the photograph.

Describe what A and B show.

____

____

____

____

____

____

Use your knowledge of forces to explain the behaviour of the magnets in test tube B.

<<< there were seven lines below the task in which to provide a response>>>

Behaviour of magnets:

Concept: magnetic and gravitational fields cause forces on objects at a distance

Background: in the Science and Technology K–6 Syllabus: ‘forces and their effects’ and ‘magnetism and some of its characteristics’ are prescribed content; ‘pushes and pulls can make things move and stop’ is a Stage 1 outcome and ‘magnets attract some materials but not others’ is a Stage 2 outcome

SOLO Cycle 1 generalises about the behaviour of magnets using information from the stimulus photograph

SOLO Cycle 2 describes the interaction of magnetic poles and explains the interaction of magnetic and gravitational forces that produces the phenomenon in test tube B in the stimulus photograph

Code	Description
8	non-attempt; the page for responding to the task is left blank
0	a response was made but it does not meet any of the marking criteria
1	the response contains a single piece of commonsense information relevant to the major concept
2	the response contains two or more pieces of commonsense information relevant to the major concept
3	the response contains a commonsense explanation about the major concept that relates two or more pieces of commonsense information
4	the response contains a single piece of 'scientific' information relevant to the major concept that clearly reflects syllabus expectations or accepted science
5	the response contains two or more pieces of 'scientific' information relevant to the major concept that clearly reflect syllabus expectations or accepted science
6	the response contains a clearly stated 'scientific' explanation about the major concept that relates two or more pieces of information, which clearly reflect syllabus expectations or accepted science

Figure 2.9 Actual EV extended response task. Source: 2008 EV test
Note: codes 1 to 6 correspond to U_1-M_1-R_1 & U_2-M_2-R_2 in Figure 2.6

Teachers who engaged with training for marking the EV extended response tasks and subsequent marking reported that the training and subsequent marking was the most important source of their learning about the SOLO model (see Chapter 4).

The text under the headings Code and Description provide the code (a number) and the general criteria for assigning the related code consistent with the qualitative differences expected for cycle 1 and cycle 2 responses.

The expectation is that students will become scientifically literate as described in the official science curriculum (syllabus in NSW). Part of this includes being able to provide scientific explanations for a range of phenomena they observe and experience in the natural and made worlds by the end of Year 10. This expectation includes being able to identify and name the concepts that link sets of seemingly disparate arrays of phenomena and use those concepts to explain observations related to the phenomena.

> Here the quality of thinking is captured by looking at SOLO levels within modes, below the mode that science demands as that which provides a satisfactory explanation. (Biggs & Collis, 1991, pp. 73-74)

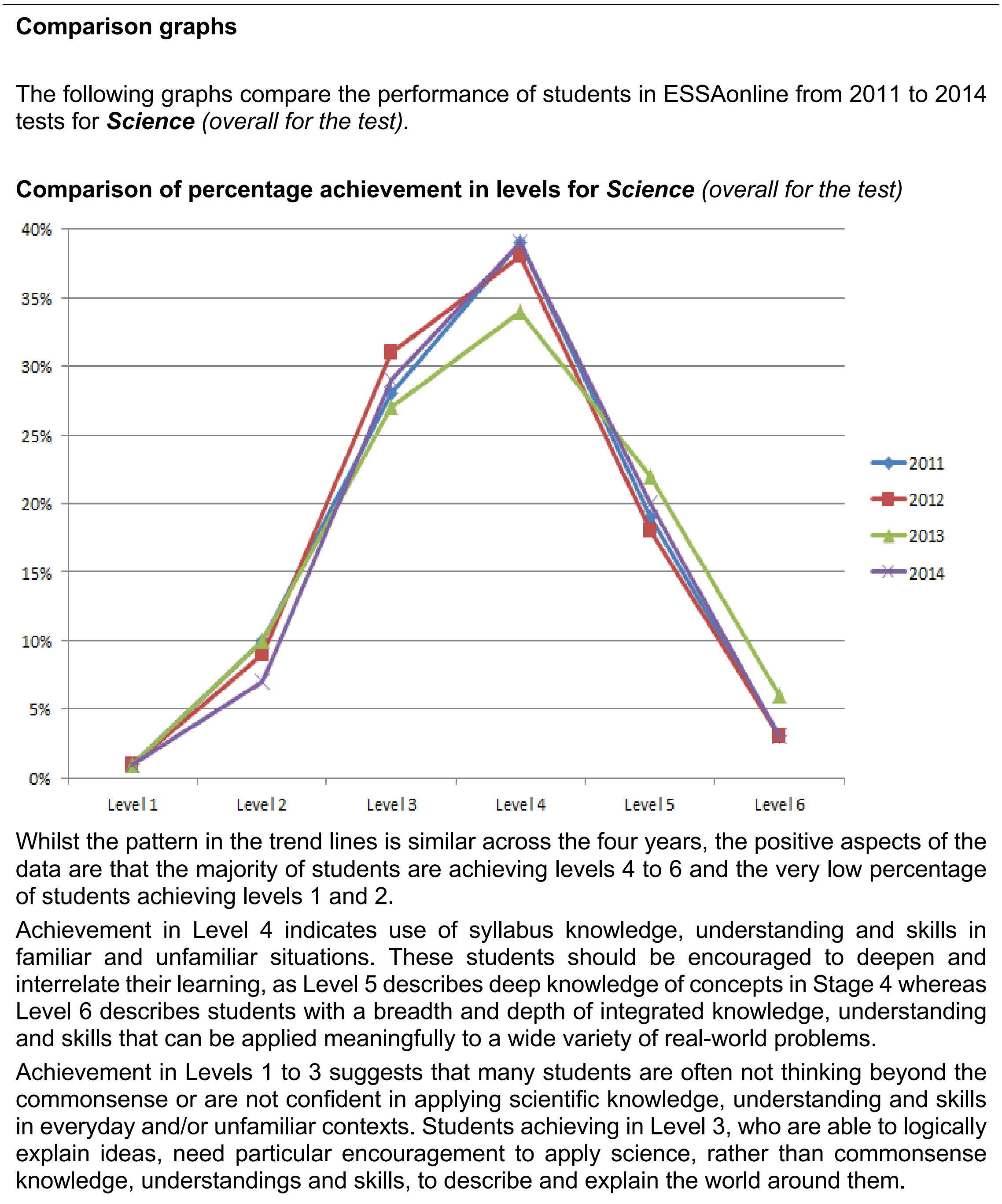
Once assessed, suitable remedial action can be taken to help the student shift their level of understanding to one closer to the syllabus version of a scientific explanation. The evaluative purposes of the test, that is, providing feedback to the Department, are explained next.

Many students are good talkers about the world they inhabit when they start school at age 5 and the level of that 'talk' for some would be at the Ikonic, R_2 level or lower (see Figure 2.6). By the end of Year 8 (age 13 -14 yrs), the graphs for the period of interest (Figure 2.10) show that around 35% of the students are operating at the second cycle, unistructural level (U_2) in science. The second cycle is where we would expect Years 7 and 8 secondary students to be working; overall, approximately 60% are. Less than 5% of students are operating at the top end of the second cycle (relational level R_2). This is the level we would hope many, if not most, Year12 (16-17 years) would be operating. This low R_2 result is in line with expectations given the age of Year 8 students. Nevertheless, it should be a goal for teachers to aim at as suggested in the commentary under the graphs in Figure 2.10.

The 2011 to 2014 results represented in the four coloured graphs in Figure 2.10 are standards referenced to an assessment framework put in place in 2011. Given this, it would have been unsurprising to see the graphs showing a progressive skew to the right with each successive year. The skew would show that teachers were working successfully to improve student levels of thinking, as evidenced by successively more of them appearing at the multistructural (M_2) level at the very least. There is some visual evidence of a shift to the right from 2011 to 2013, but it is not evident from the 2014 data. Without knowing the SE for the data points, it is impossible to comment on whether that is a real effect or not. My research findings (reported in Chapter 5) provide support for the view that the shift is not 'real'.

The use of SOLO levels to inform the reporting of achievement at the end of Year 8 is the enhancement to EV feedback referred to in Chapter 1. Parents are provided with representations of thinking (criterion referenced assessment) related to each of the six levels in the SOLO model (see Figure 2.6).

Figure 2.10 shows the proportion of Year 8 students at each SOLO level for the four years of interest in this project. The last year for the ESSA test was 2014. It became the VALID test from 2015.



Comparison graphs

The following graphs compare the performance of students in ESSAonline from 2011 to 2014 tests for ***Science*** *(overall for the test).*

Comparison of percentage achievement in levels for *Science* *(overall for the test)*

Whilst the pattern in the trend lines is similar across the four years, the positive aspects of the data are that the majority of students are achieving levels 4 to 6 and the very low percentage of students achieving levels 1 and 2.

Achievement in Level 4 indicates use of syllabus knowledge, understanding and skills in familiar and unfamiliar situations. These students should be encouraged to deepen and interrelate their learning, as Level 5 describes deep knowledge of concepts in Stage 4 whereas Level 6 describes students with a breadth and depth of integrated knowledge, understanding and skills that can be applied meaningfully to a wide variety of real-world problems.

Achievement in Levels 1 to 3 suggests that many students are often not thinking beyond the commonsense or are not confident in applying scientific knowledge, understanding and skills in everyday and/or unfamiliar contexts. Students achieving in Level 3, who are able to logically explain ideas, need particular encouragement to apply science, rather than commonsense knowledge, understandings and skills, to describe and explain the world around them.

Figure 2.10 Statewide performance data and related commentary for the ESSA test 2011-2014 (Source: DEC, Essential Secondary Science Assessment 2014 state report.)

Example three-mapping syllabus outcomes to the SOLO model

Table 2.7

Selected outcomes and related SOLO levels in the 2011 EV assessment framework

	LEVEL 1	LEVEL 2	LEVEL 3	LEVEL 4	LEVEL 5	LEVEL 6
Outcomes 4.1 to 4.5 **(2 of 7 rows)**	Identify a scientific discovery	Compare scientific discovery to other types of discovery	Link a scientific discovery to its effect on humans	Describe a development in science that has led to new developments in technology	Compare the methods of the scientist to the design model of the engineer and architect	Explain the role of scientific thinking on society
	Identify a possible career path in science	Identify a science context in a career	Link a career in science to knowledge and skills required	Identify science as a human activity	Discuss why society should support scientific research	
Outcomes 4.6 to 4.9 **(3 of 16 rows)**	Identify materials attracted by a magnet	Compare the observable effects when magnets are placed end to end	Link the observable effects when two magnets are placed end to end with their position	Describe a magnetic field as producing a force that attracts particular metals	Describe the poles of a magnet as the area/ends where the magnet's field is most intense	Explain the behaviour of magnetic poles using the term field
		Identify that objects / substances take up space and/or have mass/weight	Explain that materials are held together differently in solids, liquids and gases			Explain density in terms of a simple particle model
	Identify an observable feature in melting, freezing, condensation, evaporation or boiling	Describe observable features in melting, freezing, condensation, evaporation and boiling	Explain that, when substances melt, freeze, condense, evaporate and boil, they are still made of the same stuff	Identify that particles are continuously moving and interacting	Compare movement and interaction of particles in different states	Explain change of state in terms of rearrangements of particles
				Identify that as particles are heated they gain energy	Identify that as particles are heated they gain energy and move	Relate changes of state to the motion of particles as energy is removed or added

					further apart	
No content for Outcomes 10 - 12 is included						
Outcomes 4.13 to 4.15* (1 of 8 rows)	Make a simple observation	Compare observations made by different people	Explain strategies to increase accuracy of observation	Correctly sequence steps in a scientific procedure	Accurately and systematically record observations and data	Discuss the relationship between accuracy and reliability
Outcomes 4.16, 4.17 a-d & 4.18 (1 of 8 rows)**	Use a simple key or symbol to represent a concrete object or representation	Distinguish between different symbols	Complete diagrams and symbolic representations	Correctly sequence steps in a process described in a text	Distinguish between two related sets of data / information	Represent relationships using keys, symbols and flow chart
Outcomes 4.17e-g, 4.19-4.21* (1 of 7 rows)**	Identify a common unit of measurement	Identify the ratio of one unit to another	Complete a correct conversion of one unit to another	Create a simple scale	Compare the scale on two axes	Create an appropriate scale

Source: NSW Department of Education and Training DET, 2011. Shaded rows are referenced in the body text. * Planning and Conducting Investigations area / ** Communication area / and *** Critical thinking area

Table 2.7 and following text explains the map of syllabus outcomes and SOLO model levels 1 to 6.

While the test was delivered as a pen and paper exercise (from 2005 to 2010), the assessment framework discussed in this subsection was being developed and validated.

Table 2.7 shows an extract of the framework. It shows how the syllabus outcomes (written and published for the 2003 science syllabus) were subsequently related to the six levels of the concrete symbolic mode of thinking in the SOLO model.

The melting ice task above and a second task involving magnets described in subsection 2.6.4 were parts of tests done before 2010. Both these tasks subsequently mapped to the EV framework produced for the 2011–2014 tests (see the shaded sections across the extract from the EV framework in Table 2.7). A new framework was used to inform test development for the VALID tests that began from 2015. This was based on the new Australian Curriculum Science (NSW version). The quality control processes used to develop items and tasks for EV tests are discussed in subsection 2.6.4.

Since 2011, the EV tests have been delivered online to school computers linked to school networks and the internet. The affordances provided by online delivery will not be addressed in this thesis.

The EV framework for the 2011 to 2014 EV test was organised as a grid (Table 2.7). The columns identify the the six performance levels (LEVELS 1 to 6) related to the "two learning cycles within a mode" SOLO model. The five rows accommodate 21 of the 22 syllabus outcomes defining Stage 4 (Years 7 & 8) (DET, 2011).

The performance levels (Table 2.7) correspond to the three levels of thinking in each of the two learning cycles in the concrete symbolic mode of thinking (see Figure 2.6). LEVEL 1 = first cycle U_1, LEVEL 2 = first cycle $M_{1,}$ LEVEL 3 = first cycle R_1, LEVEL 4 = second cycle U_2, LEVEL 5 = second cycle $M_{2,}$ and LEVEL 6 = second cycle R_2.

The descriptors in each of the grid cells were identified as appropriate for Stage 4 learners by experienced science teachers and SOLO experts (as explained in the subsection 2.6.4). The wording used was based on a combination of their professional judgment and the outputs from sophisticated psychometric analysis using results from trialling and piloting and the first few years of full cohort testing.

The outcomes are numbered in the left-hand column. The first digit refers to the Science Syllabus Stage, which in this case is Stage 4 (for students to achieve by the end of Year 8). The second number identifies the outcome (from 1 to 21) The letters correspond to content related actions (indicators of outcome attainment) linked to essential syllabus content (that students will learn about) that defines the scope of the outcome.

The 21 outcomes able to be assessed by this mode of testing (pen and paper) are grouped on the full EV assessment framework to reflect groups of outcomes defined here as syllabus areas:

- Prescribed focus area (Outcomes 4.1-4.5)
- Knowledge and understanding area (Outcomes 4.6-4.12)
- Planning and conducting investigations area (Outcomes 4.13, 4.14 & 4.15)
- Communication area (Outcomes 4.16, 4.17a-d and 4.18)
- Critical thinking area (Outcomes 4.17e-g, 4.19, 4.20 & 4.21) (DEC, 2015, p. 18).

The first two bullet-point areas are both knowledge and understanding outcomes; the remaining three are related to science skills and processes. Not every cell of every row has a content description because the syllabus is silent about relevant content at that SOLO level. This is to be expected because the syllabus was published in 2003; the six levels were identified within the existing syllabus content and 'levelled' using SOLO expectations in 2010.

About half of the outcomes are about knowledge and understanding (1 to 12); the other half are science process / skill outcomes (13 to 22). The intended message is that junior secondary science is as much about science knowledge and understanding as it is about 'doing science'. Thus, the full grid provides an easy way to map the items for a particular test against syllabus expectations and all SOLO levels of thinking.

The *Critical Thinking* grouping of outcomes in the EV framework (see the bottom row of Table 2.7) is an attempt by the test developers to signal to teachers that having students engage critically with science and science-related issues is an important expectation. In the full version of the EV framework, there are seven rows of content descriptors for this area. The one chosen here is about measurement; others relate to a progression in thinking to do with mathematical relationships between variables, data analysis, evidence-based conclusions, critical analysis of scientific explanations, predictions and inferences (based on scientific evidence), and recognising aspects of a problem that may be resolved using science.

Measurement is an important component of the 'epistemic' basis of science. The meaning of 'epistemic' and examples of items and tasks related to it are explained in the PISA2015 frameworks document (OECD, 2017, pp. 29-38). The six SOLO levels related to measurement in the *Critical Thinking* in Table 2.7 begin with specific contexts for measurements and them move to relationships between aspects of the same and different measurements (SOLO cycle one). From LEVEL 4 the expectation progresses to developing a more generalised understanding of measurement scales (SOLO cycle two). The progression has the power to guide teaching and assessment aligned to syllabus intentions.

The cell descriptors along each row provide guidance to a teacher about what content is to be learned and the complexity of thinking students are expected to manage as they work through the two years of Stage 4 in science. An example describing a possible progression in learning about units of measurement (in a particular context) and its extension to measurement scales generally will now be described.

Outcome 4.7 about the particle model of matter (4.7.1) and melting ice as a particular example of change of state (4.7.3) are the contexts for this learning sequence. The end result of using different thermometers (glass-alcohol and digital) with different scales (Kelvin, Celsius and Fahrenheit) is to validate an operational definition for latent heat of melting for one substance, water. The observation that the temperature of an ice-water mixture stays the same until all the ice melts is explained by the idea that added heat is being used to 'overcome' whatever it is holding the water particles together to make ice rather than increasing the temperature of the ice-water mixture).

The initial emphasis here is to ensure accurate and reliable measuring of the temperature as ice melts. The class discussion before working in groups (each using a different thermometer with a different scale) would be to discuss the three temperature scales (Kelvin, Celsius and Fahrenheit). Why the three scales? Which one to use / or all of them? Why? What is the ratio of one scale division relative to the other across the three scales? How do we convert from one to the other? Why might this be a useful conversion to be able to do? A worksheet could be developed for use by members of the group working together to discuss and answer the questions posed there.

Once the task is completed, each group member takes the results and independently answers another set of questions (such as the ones in the 2005 test and others relating to scale conversions and drawing a graph of one scale verses another to interpolate within and extrapolate beyond) provided on another sheet. A final question to answer might be: How would you recalibrate a thermometer where the scale had been rubbed out?

The challenge to 'calibrate a thermometer with no scale' addresses the BOS *Common Grade Descriptor* criteria for an A: "Can apply these skills to new situations" (BOS, 2013). The tasks above also address three Mayer Key Competencies relating to solving a problem, using mathematical ideas and techniques as well as using technology (in this situation, analogue thermometers with three different scales and digital technology in the form of digital temperature probes linked to data loggers or computers). The Mayer Key Competencies were integrated into the syllabus at the time it was written (AECRC, 1992).

Example four-reporting achievement at the end of Year 8

The results from the EV test are organised into a summative report of achievement at the end of Year 8. The report for students, parents and teachers provides the results for five areas or categories of outcomes. The scores from items in the EV framework mapped to the *Critical Thinking* area are distributed to the working scientifically and communicating scientifically categories, depending on whether the items had an investigating or communicating context. The student report provides individual feedback on every task and item in the test.

Individual responses are also aggregated to provide a score and position on a scale from 1 to 6 related to the six SOLO levels as shown in Figure 2.7. Five scales are provided showing an overall score for science, a score for the three extended response tasks and three separate scales for knowledge and understanding, communicating scientifically and working scientifically. Providing feedback on five categories of science achievement is more useful and respectful of achievement by an individual than a single indicator of overall achievement, such as a grade or mark. It is also diagnostic in the sense that an assessment of strengths and weaknesses in particular areas of science can be easily seen.

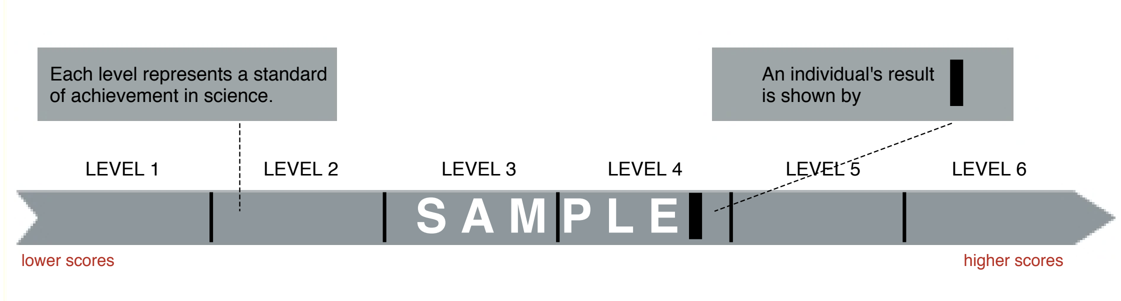

Each level represents a standard of achievement in science.
An individual's result is shown by ▌
LEVEL 1
LEVEL 2
LEVEL 3
LEVEL 4
LEVEL 5
LEVEL 6
SAMPLE
lower scores
higher scores

Figure 2.7 A sample reporting scale (Source: DET, 2007)

The student's score (represented as a thick line on the scale) and its placement on the scale is determined by a combination of factors going back to the development of items and tasks for inclusion in the test and the dependability (Harlen, 2004) of the processes used then and subsequently to produce the scores and print its representation on the proficiency scale. The quality control processes to do that will be discussed below.

Table 2.8 provides an extract from the EV student report for reference. TIMSS, PISA and NAP-SL also have comparable sets of descriptors for each of the proficiency levels related to their assessment frameworks.

Table 2.8

Extract from student report showing selected levels for three reporting categories

⇑	Knowledge & understanding	Communicating scientifically	Working scientifically
Level 6	• Explains physical phenomena using a model, theory or law • Explains the interaction of complex systems (for example, relates the role of the circulatory system to the needs of cells)	• Explains the theme and function of a complex text • Critically analyses the credibility of scientific information	• Relates the dependent and independent variables for a given problem • Describes the wider significance of conclusions (for example, accounts for the differing amounts of water loss by plant cuttings by identifying plant processes)
Level 5	• Describes examples where scientific understanding has changed • Describes interactions of systems or within systems	• Extracts related information from diagrams, tables, graphs or other texts • Compares two sets of information (for example, compares a table and graph and inserts information into the graph)	• Identifies ways to improve the reliability and accuracy of controlled investigations • Applies mathematical models to data (for example, interpolates information from a line graph)
Level 4	• Identifies scientific evidence (for example, identifies evidence that leads to change in a scientific theory) • Describes a complex process of our world or space (for example, identifies requirements for photosynthesis) • Identifies an interaction of systems or within a system (for example, identifies evidence that indicates that a chemical reaction has occurred)	• Identifies one piece of relevant scientific information • Describes an effective solution to a problem with a science context	• Identifies a prediction, inference, conclusion, aim and hypothesis • Selects one piece of appropriate scientific equipment for a task (for example, identifies a benefit of using a data logger to collect information in an investigation) • Draws a conclusion based on scientific evidence
Leve l 3 ⇑	• Explains a link between technology and science • Relates simple processes of our world or space (for example, identifies insects as consumers) • Relates a model to an aspect of our world or space (for example, identifies kinetic energy acting in an activity)	• References information within a diagram, table, graph or other text (for example, summarises ideas across a text) • Uses cause and effect to explain an observation (for example, identifies the effect of a change during a process)	• Relates equipment and appropriate use for a simple task (for example, identifies the correct use of a thermometer) • Draws a simple conclusion

Source: DET, 2007, p. 3. The lower arrow represents the transition from the R_1 level of the first cycle in the concrete symbolic mode of thinking; the higher arrow represents the transition to the U1 level of the next (formal) mode of thinking.

The formative intent of the EV program is signalled in the report to parents and students:

> Students, parents and teachers can use the [EV] levels [Table 2.8] to plan learning programs and activities so that students keep moving forward in their science knowledge and skills. (DET, 2007, p. 3)

The levels referred to are the six levels linked to the SOLO model discussed above. Progress ("moving forward" in the EV report) in science learning is defined by the language used in each of the level descriptions for a particular reporting category.

The feedback from the test and student survey is provided to schools and school systems participating in the test some six months after the tests are done, and well into a new school year when students have commenced the next stage of learning (syllabus stage 5 early in Year 9). Because the feedback is not immediate, the results are helpful to teachers when evaluating their programs and making changes for the next cohort of students as discussed in Chapter 5.

Because the primary purpose of the EV test is to provide feedback to students, parents, teachers, schools and school systems about progress in student learning, the aim is to have as many students finish the test as possible. The test administration process is managed at the school level by teachers, and schools have one week in which to complete the exercise. To ensure that students are able to complete the test, time allocations for the sections of the test are listed as approximate only. Eight minutes is advised for the preliminary, practice items; 20 minutes for the three extended response items; an hour for the short response item sets; and, about five minutes for the student survey.

In keeping with the purpose of providing feedback to individuals, students do the test individually as they would any other test. There is no competitive advantage to be had by 'cheating' because the results provide individual feedback about their learning relative to the syllabus and SOLO levels, not how well they are doing relative to other students.

Example five-EV test items, stimulus material and student survey

Schools are encouraged to keep and reuse the test items and tasks in their own school-based assessments because they are exemplary assessment items. Teachers have access to all the tests from previous years and related stimulus material as well as the assessment rubrics used to mark the three extended response items in those years.

The three extended response items in the EV test were placed immediately after the preliminary practice items when the test was in print form. Experience with external tests at that time, where extended response questions were at the end of the test, showed that many students simply ignored those questions. Placement at the beginning of the test obtained an almost 100% response. Of the three extended response tasks, one involved an investigation scenario; the other two primarily addressed syllabus knowledge and understanding expectations.

Extended response tasks are open ended so that students can respond at the highest level of understanding they are capable of demonstrating. (see examples one and two above). The relevant syllabus references related to these two tasks are highlighted in the section of the EV framework provided as Table 2.7

Short response items are written to identify not only a student's knowledge and understanding but also their ability to comprehend at or above the lowest targeted SOLO level of thinking (as reflected in the wording of the item). Items are linked to a piece of stimulus material rich in science content from the syllabus for that stage of learning. The text provided is chosen from the range of experiences an adolescent learner is likely to have had or to know about. It might be an extract from a newspaper or magazine or an advertisement or a recount of a TV news item, for example. From three to eight items targeting a range of SOLO levels might be related to any one piece of stimulus material.

Items and tasks 'look and feel' to students like items and tasks in other external tests they do each year for NAPLAN. A test would have around 75 to 85 short response items. Not all the knowledge and understandings needed to satisfy item demands are provided in the stimulus material. Students are expected to use knowledge and understanding from the syllabus to respond appropriately. Students are expected to respond by choosing one from three to five alternatives (to identify the best answer) or to write a few words or the result of a calculation on the answer sheet provided. Distractors are chosen, where possible, to identify misconceptions students may have (see Figure 2.8, Item 14).

ESSA 2014 TEST Stimulus material and related items 9 to 16

Read the following article then complete items 9 to 16.

Why use a pool cover?

A pool cover is a great investment. Over a whole year, a pool can lose up to 5 mm of water each day. By using a pool cover, the water loss is reduced by about 95%.

Pool covers also extend the swimming season by increasing the pool's water temperature by up to 8ºC.

A well-fitted pool cover keeps dirt, leaves and insects out of the pool. This also helps the cleaning equipment to keep the water suitable for swimming.

9 Choose yes or no for each reason to answer the following question:

According to the article, what are the reasons that a pool cover is a great investment?

	Yes	No
prevents water loss	○	○
saves energy	○	○
keeps the pool cleaner	○	○
extends use of the pool	○	○

10 Using a net to remove leaves and insects from a pool is an example of

○ chromatography
○ filtration
○ sedimentation

11 Swimming pools would lose most water during

○ cool and cloudy days
○ cool and windy days
○ warm and cloudy nights
○ warm and windy days

12 On a hot day, the water on the surface of a pool would most likely undergo

○ a physical change
○ a chemical change
○ no change

13 Gaseous water is less dense than liquid water because particles in gaseous water are

○ closer together
○ further apart
○ smaller in size
○ larger in size

14 On hot days, water particles in the pool collide into each other more often because the water particles

○ have more energy
○ have less energy
○ get larger as the pool warms up
○ are made as the pool warms up

15 What is one environmental impact of covering a pool?

○ Australia would have fewer droughts.
○ There would be more water in dams.
○ People could swim for more months in the year.
○ Swimming pools would stay clean and leaf-free.

16 Pure water has the chemical formula H_2O.

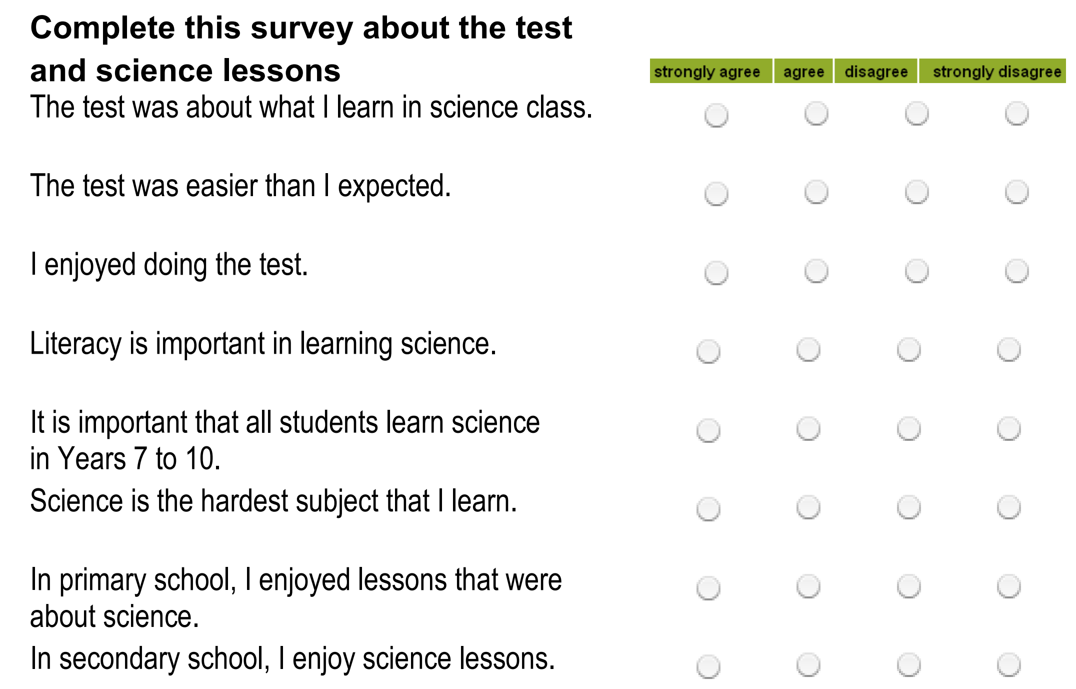
What type of chemical substance is pure water?

○ compound
○ element
○ mixture

Figure 2.8 One of the stimulus-item sets from the 2014 EV test. Source: NSW Department of Education, ESSA 2014 Test item.

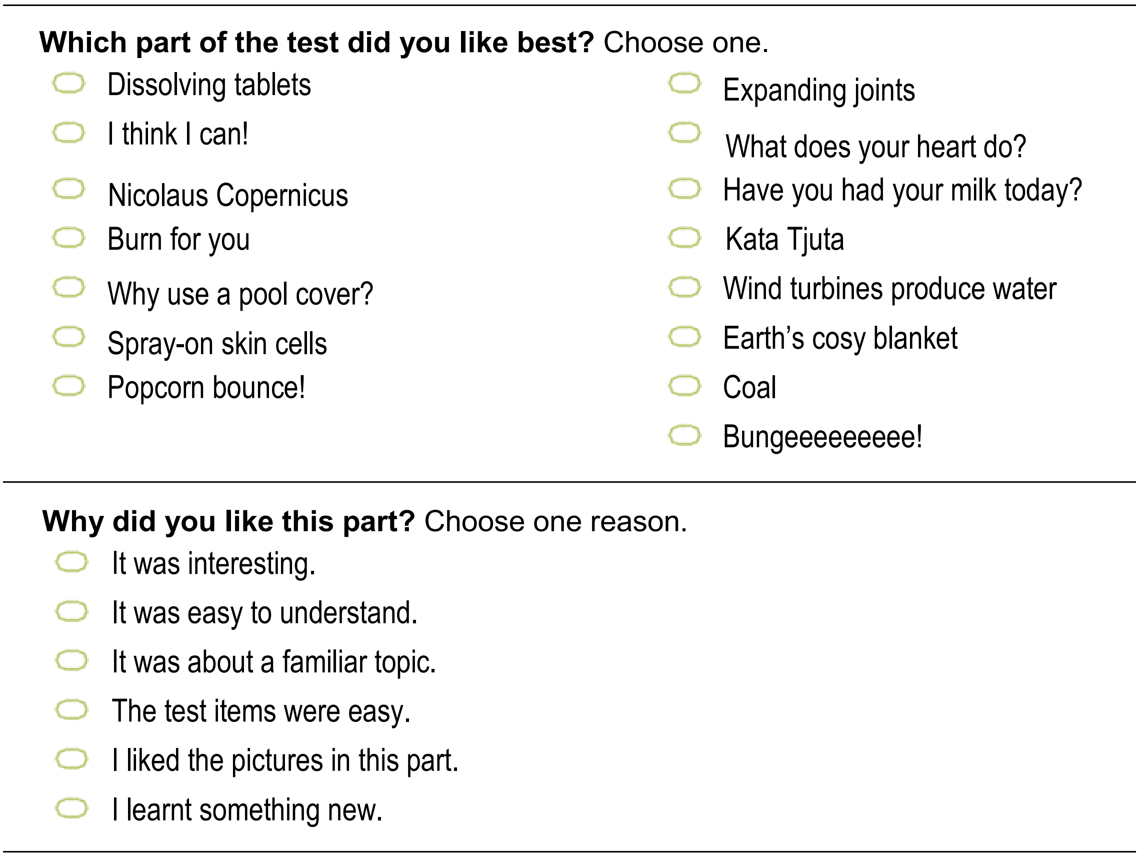
Figure 2.9 (in example 2 above) provides a task from the 2008 EV test and the descriptors for applying a code to student responses in the online marking process. A feature of the coding process is that markers are asked to code for the highest level of response evidenced in an answer. The text in the section under the task in Figure 2.9 outlines expectations for responses. For cycle 1 the language used is sourced from the expected learning related to magnets and forces in the K-6 *Science and Technology* syllabus. Cycle 2 response language is sourced from the Science 7-10 syllabus in use by schools at the time (BOS, 2003).

Some of the student survey questions specifically address issues to do with the test, as exemplified by the extract from the student survey (see Figure 2.11).



Complete this survey about the test and science lessons

	strongly agree	agree	disagree	strongly disagree
The test was about what I learn in science class.	○	○	○	○
The test was easier than I expected.	○	○	○	○
I enjoyed doing the test.	○	○	○	○
Literacy is important in learning science.	○	○	○	○
It is important that all students learn science in Years 7 to 10.	○	○	○	○
Science is the hardest subject that I learn.	○	○	○	○
In primary school, I enjoyed lessons that were about science.	○	○	○	○
In secondary school, I enjoy science lessons.	○	○	○	○



Which part of the test did you like best? Choose one.

○ Dissolving tablets
○ I think I can!
○ Nicolaus Copernicus
○ Burn for you
○ Why use a pool cover?
○ Spray-on skin cells
○ Popcorn bounce!
○ Expanding joints
○ What does your heart do?
○ Have you had your milk today?
○ Kata Tjuta
○ Wind turbines produce water
○ Earth's cosy blanket
○ Coal
○ Bungeeeeeeeee!

Why did you like this part? Choose one reason.

○ It was interesting.
○ It was easy to understand.
○ It was about a familiar topic.
○ The test items were easy.
○ I liked the pictures in this part.
○ I learnt something new.

Figure 2.11 Questions about the EV test. Source: NSW Department of Education and Communities, ESSA Test, 2014.

Student responses are used as feedback to refine and improve the test and the test experience for students going forward. Schools also receive the feedback from students at their school and their responses can be compared to the rest of the state.

Appendix E: Proforma for case study schools to complete

SCHOOL:

DATE:

For this page, fill in the boxes by estimating the scale reading in SMART for each of the six components. If pressed for time only do the odd years coming back from 2015…5/6, 3/4 & 1/2 refer to student achievement levels as represented in SMART for the survey.

YEAR 8 ESSA-VALID STUDENT SURVEY DATA…Please obtain this from SMART

		2011	2011	2012	2012	2013	2013	2014	2014	2015	2015
	LEVEL	School	State	School	State	School	State	School	State	School	State
A. I want to study a science subject in Years 11 &12	5-6										
	3-4										
	1-2										

		2011	2011	2012	2012	2013	2013	2014	2014	2015	2015
	LEVEL	School	State	School	State	School	State	School	State	School	State
B. Science is the hardest subject that I learn	5-6										
	3-4										
	1-2										

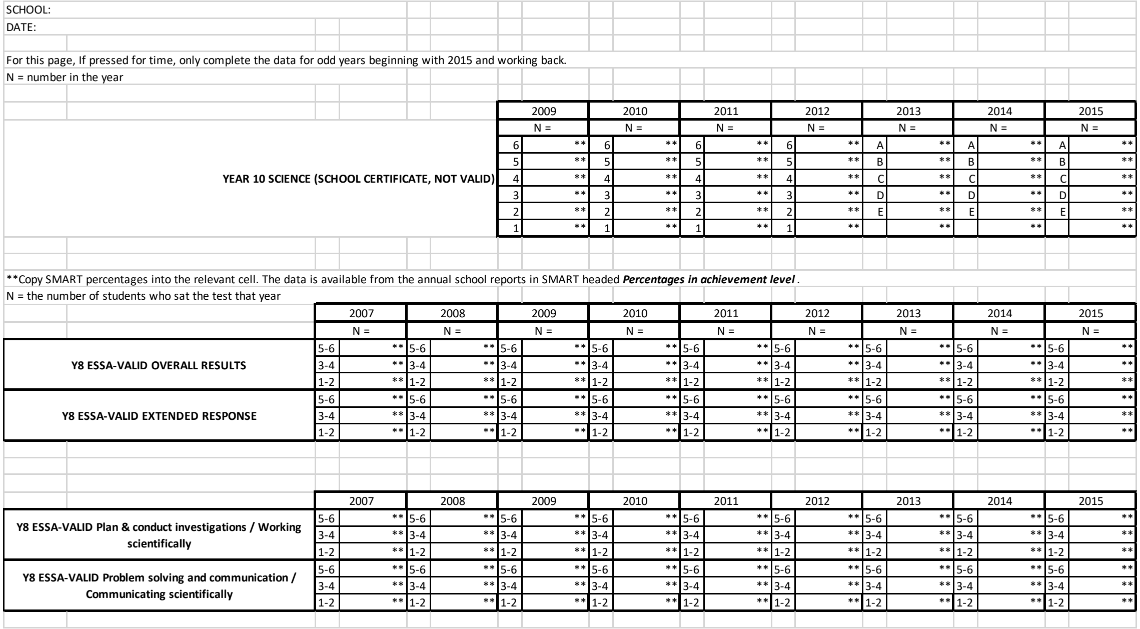
		2011	2011	2012	2012	2013	2013	2014	2014	2015	2015
	LEVEL	School	State	School	State	School	State	School	State	School	State
C. In primary school, I enjoyed lessons that were about science	**5-6**										
	3-4										
	1-2										

		2011	2011	2012	2012	2013	2013	2014	2014	2015	2015
	LEVEL	School	State	School	State	School	State	School	State	School	State
D. In secondary school, I enjoy science lessons	**5-6**										
	3-4										
	1-2										

		2011	2011	2012	2012	2013	2013	2014	2014	2015	2015
	LEVEL	School	State	School	State	School	State	School	State	School	State
E. My three favourite school subjects are (record the % for science)	**5-6**										
	3-4										
	1-2										

		2011	2011	2012	2012	2013	2013	2014	2014	2015	2015
	LEVEL	School	State	School	State	School	State	School	State	School	State
F. The three subjecs I think I learn most in (record the % for science)	**5-6**										
	3-4										
	1-2										

In terms of your priorites for science in Years 7-9, write the letter representing the six statements (A to E) in order of importance (most important first):



SCHOOL:

DATE:

For this page, If pressed for time, only complete the data for odd years beginning with 2015 and working back.

N = number in the year

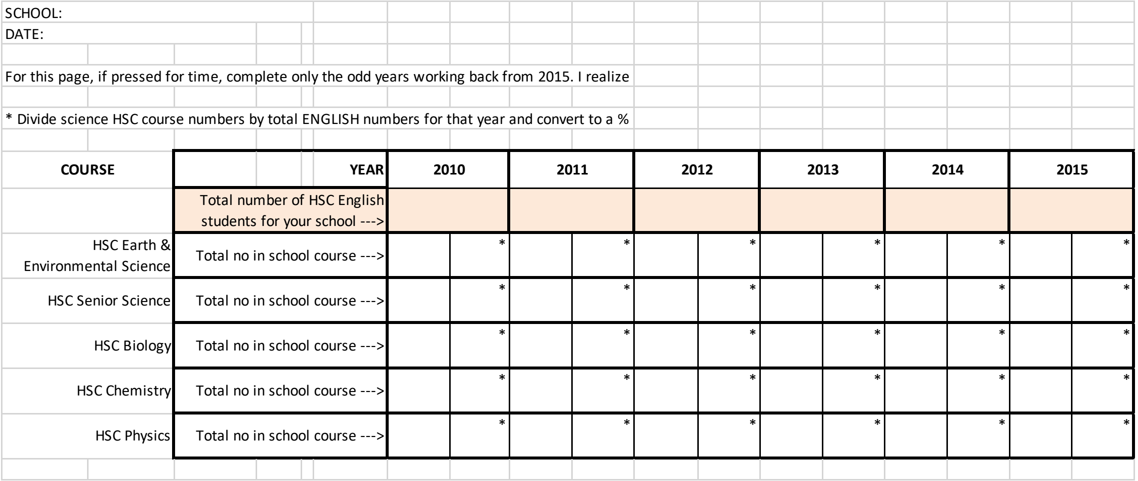
	2009		2010		2011		2012		2013		2014		2015	
	N =		N =		N =		N =		N =		N =		N =	
YEAR 10 SCIENCE (SCHOOL CERTIFICATE, NOT VALID)	6	**	6	**	6	**	6	**	A	**	A	**	A	**
	5	**	5	**	5	**	5	**	B	**	B	**	B	**
	4	**	4	**	4	**	4	**	C	**	C	**	C	**
	3	**	3	**	3	**	3	**	D	**	D	**	D	**
	2	**	2	**	2	**	2	**	E	**	E	**	E	**
	1	**	1	**	1	**	1	**		**		**		**

Copy SMART percentages into the relevant cell. The data is available from the annual school reports in SMART headed ***Percentages in achievement level* .

N = the number of students who sat the test that year

	2007		2008		2009		2010		2011		2012		2013		2014		2015	
	N =		N =		N =		N =		N =		N =		N =		N =		N =	
Y8 ESSA-VALID OVERALL RESULTS	5-6	**	5-6	**	5-6	**	5-6	**	5-6	**	5-6	**	5-6	**	5-6	**	5-6	**
	3-4	**	3-4	**	3-4	**	3-4	**	3-4	**	3-4	**	3-4	**	3-4	**	3-4	**
	1-2	**	1-2	**	1-2	**	1-2	**	1-2	**	1-2	**	1-2	**	1-2	**	1-2	**
Y8 ESSA-VALID EXTENDED RESPONSE	5-6	**	5-6	**	5-6	**	5-6	**	5-6	**	5-6	**	5-6	**	5-6	**	5-6	**
	3-4	**	3-4	**	3-4	**	3-4	**	3-4	**	3-4	**	3-4	**	3-4	**	3-4	**
	1-2	**	1-2	**	1-2	**	1-2	**	1-2	**	1-2	**	1-2	**	1-2	**	1-2	**

	2007		2008		2009		2010		2011		2012		2013		2014		2015	
Y8 ESSA-VALID Plan & conduct investigations / Working scientifically	5-6	**	5-6	**	5-6	**	5-6	**	5-6	**	5-6	**	5-6	**	5-6	**	5-6	**
	3-4	**	3-4	**	3-4	**	3-4	**	3-4	**	3-4	**	3-4	**	3-4	**	3-4	**
	1-2	**	1-2	**	1-2	**	1-2	**	1-2	**	1-2	**	1-2	**	1-2	**	1-2	**
Y8 ESSA-VALID Problem solving and communication / Communicating scientifically	5-6	**	5-6	**	5-6	**	5-6	**	5-6	**	5-6	**	5-6	**	5-6	**	5-6	**
	3-4	**	3-4	**	3-4	**	3-4	**	3-4	**	3-4	**	3-4	**	3-4	**	3-4	**
	1-2	**	1-2	**	1-2	**	1-2	**	1-2	**	1-2	**	1-2	**	1-2	**	1-2	**



SCHOOL:

DATE:

For this page, if pressed for time, complete only the odd years working back from 2015. I realize

* Divide science HSC course numbers by total ENGLISH numbers for that year and convert to a %

COURSE	**YEAR**	**2010**		**2011**		**2012**		**2013**		**2014**		**2015**	
	Total number of HSC English students for your school --->												
HSC Earth & Environmental Science	Total no in school course --->		*		*		*		*		*		*
HSC Senior Science	Total no in school course --->		*		*		*		*		*		*
HSC Biology	Total no in school course --->		*		*		*		*		*		*
HSC Chemistry	Total no in school course --->		*		*		*		*		*		*
HSC Physics	Total no in school course --->		*		*		*		*		*		*

Appendix F: Science teacher survey questions

INSTRUCTIONS AND CONSENT

The purpose of this survey is to find out about your use of the ESSA/VALID program in the context of all your assessment- related work in science.

There are 26 questions and the whole survey should take you about 25 minutes to complete. You can change your mind at any time and stop completing the survey without consequences. If you choose to identify yourself, I will keep any data you provide confidential.

###

I have read and understand the material about the research provided in ATTACHMENTS ONE AND TWO forwarded to me by my principal.

I wish to proceed with answering the questions

SECTION ONE: ABOUT ESSA / VALID

The ESSA/VALID test has been a part of the Year 8 science experience since 2007. Feedback from the ESSA/VALID test and the related student survey accompanying it is provided to schools in Term 1 of the year following the test. The following items ask about your use of the ESSA/VALID test and related feedback. [Radio buttons for YES / NO]

1. In relation to ESSA/VALID results for my school, I have in the previous twelve months:

1a. looked at the results of the student survey

b. looked at the item analysis for my class / school

c. looked at the analysis of answers to the three extended response tasks

d. looked at the student profile information

e. discussed the item or task analysis with colleagues

f. discussed the item or task analysis with students

g. discussed the results of the student survey with colleagues

h. discussed the student profile information with colleagues

i. discussed the results of the student survey with students

2. I have in the previous two years:

2a. accessed ESSA/VALID-related materials in TaLE

b. accessed ESSA/VALID-related materials in SMART

c. used in my classes teaching strategies that I found in the ESSA/VALID-related Curriculum Links materials

d. accessed the ESSA/VALID Marking Manual/s for the extended-response tasks

e. used ESSA/VALID short response items in topic tests

f. used ESSA/VALID extended-response tasks in topic tests

g. used ESSA/VALID items & / or extended-response tasks in my teaching

h. used ESSA/VALID items & / or extended response items as models for writing new items and tasks in topic tests

i. contributed amendments to faculty programs as a direct response to ESSA/VALID results

j. written items for the ESSA/VALID test

k. been on a panel to evaluate ESSA/VALID items

l. been a marker for the ESSA/VALID extended response tasks

m. have attended workshops about ESSA/VALID (NOT including training for ESSA/VALID marking)

3. Overall, I would rate my understanding of the ESSA/VALID program as

very poor / poor / acceptable / good / very good

4. I think the most important purpose for the ESSA/VALID test is…

[a box with 100 words limit]

5. My school will this year participate in the VALID science test for Year 10 students

Yes / No / Unsure

SECTION TWO: ABOUT SOLO

This next set of items is about the Structure of the Observed Learning Outcome (SOLO) model. [Radio buttons for YES / NO]

6. In relation to SOLO, I have in the previous two years:

6a. accessed material about SOLO in the Marking Manual for the extended-response tasks in the ESSA/VALID test

b. accessed material about SOLO in places other than the Marking Manual for the extended-response tasks in the ESSA/VALID test

c. explained the SOLO model to another teacher

d. used the SOLO model to explain the ESSA/VALID student profile to a student

e. used SOLO to develop assessment criteria for my assignments / tests / tasks

f. used SOLO to provide feedback to students about their learning

g. Led a discussion with science staff about ESSA/VALID results and what they mean in terms of the SOLO model

h. worked in a science faculty where using SOLO is an explicit part of science faculty policy & / or practice

i. used SOLO concepts &/or model in regular student reports of science achievement sent home to parents / careres

j. used the SOLO model to explain the ESSA student profile to a parent or carer

7. Overall, I would rate my understanding of SOLO as

very poor / poor / acceptable / good / very good

8. I consider that I learnt most about SOLO when…

[a box with 100 words limit]

SECTION THREE: ABOUT "ASSESSMENT FOR LEARNING"

The following questions are about "assessment for learning" and related practices.

I'm wanting to find out HOW OFTEN (in a relative sense) you do the things described below in your day to day teaching and related work in your Years 7, 8 and 9 classes.

You may not know about or be unsure about some things listed here in which case choose the Not known / Unsure about button. [Respondents had the options of marking:

Not known /unsure about Never Seldom Sometimes Often

9. When working in the classroom with students, I

9a. tell them what they should know, understand, be able to do by the end of the lesson

b. give students the opportunity to set their own learning intentions for an activity or series of activities

c. explain to students the indicators or success criteria I will be looking for in their work

d. allow students some input in deciding what success criteria are to be applied

e. make the significance of what they are to do explicit to students

f. ask students why they think they are being asked to do the proposed activities

g. encourage peer feedback based on success criteria

h. use results from instant digital polling technology to inform next steps in teaching that lesson

10. When managing classroom discussions, I

10 a. ask closed questions

b. ask open questions

c. use wait-time before responding

d. ask students to explain their thinking

e. use the “think-pair-share-report” strategy

f. use test or assignment items and tasks as stimulus for discussion

g. use samples of student work or responses to assessment items as stimulus material for discussion

h. explain my responses / thinking

11. I provide feedback on student's written work in the form of

11 a. ticks

b. marks

c. grades (such as A to E)

d. comments about what they have done well (eg good work, excellent, well done...)

e. advice about how to improve

IF FOR 11e YOU CHOSE NOT KNOWN / UNSURE ABOUT OR NEVER, SKIP THE NEXT QUESTION (Q. 12) AND GO TO Q.13

12. As the BASIS for my advice about how to improve, I refer to

12 a. exemplary or model answers

b. success criteria

c. misconceptions evident in answers

d. SOLO levels of thinking

e. Quality Teaching dimensions and related elements

f. Bloom’s taxonomy / hierarchy of thinking skills

g. syllabus standards (syllabus outcomes & related content)

13. I provide opportunities for students to self-assess

13 a. by getting them to write success criteria for activities & investigations

b. by getting them to construct assessment items and tasks

c. using success criteria / assessment rubrics or guidelines

d. by redoing work to a higher standard

e. by selecting items for a portfolio of work they judge as being consistent with nominated success criteria

f. by getting them to keep a journal of their reflections in their own words on what they have learned in science lessons

14. In my day-to-day preparation for and work in class, I

14 a. ask students to give me feedback on my teaching

b. respond to students’ feedback on my teaching (this may not always be an immediate response)

c. evaluate lessons and record ideas for change next time

d. keep notes on learning issues noticed for individual students

e. change the planned ‘next step’ in a lesson in response to student feedback at the time

f. access and use information about “assessment for learning” in TaLE

g. access and use information about “assessment for learning” in the BOSTES website

h. access and use information about “assessment for learning” from other places / sources (apart from TaLE & BOSTES)

15. I collaborate with my science teacher colleagues to

15 a. write items and tasks for tests & / or assignments

b. produce marking criteria / assessment rubrics

c. assess assignments / tasks / tests from each others classes

d. to develop a shared understanding of learning intentions and success criteria implicit in syllabus outcomes for junior secondary science

e. develop a shared understanding of what progression in science learning looks like

---

SECTION FOUR: ABOUT YOU AND YOUR TEACHING EXPERIENCE/CONTEXT

16. I am a

female male other

17. I have been teaching

0-5 yrs 6-10 yrs 11-15 yrs 15+ yrs

18. I am a science teacher by training / qualification

Yes No

If NO to Q. 18 my qualifications are…

____

19. I am a head teacher, science

Yes No

20. My HIGHEST science teaching qualification is

Bachelor degree + Dip ed (or equivalent Postgraduate qualification)

BTeach (4 yr degree)

MTeach (5 yr degree)

Doctorate or PhD

Other science teaching qualification

____

21. I completed my highest qualification in (what year) _________

22. My training / qualifications to teach science were untertaken

completely overseas / partly in Australia and partly overseas / completely in Australia

23. I teach / have taught Years 7-9 classes

this year / last year / the year before last / more than three years ago

24. At my current school there are this many Year 8 science classes

one / two / three / four / five / six / seven / eight / eight +

25. At my current school there are this many full-time science teachers

one / two / three / four / five / six / seven / eight / eight +

26. At my current school there are this many part-time science teachers

one / two / three / four / five / six / seven / eight / eight +

If you are happy to be contacted about this survey &/OR are interested in contributing to a case-study about ESSA/VALID and assessment practices, please provide the following information and identify yourself as requested below.

27. I am interested in finding out more about this survey (note you will need to provide your name and preferred contact details below)

Yes / No

28. I am intersted in finding out more about the case study and what it would involve (note you will need to provide your name and preferred contact details below)

Yes / No

My given name is:

My surname / family name is:

My preferred contact mode is (please provide details):

My current school is:

I was appointed to my current school in (year):

My previous school was:

Thank you for giving your time to complete this survey.

Your input will help me to better understand ‘assessment for learning’ practices used by science teachers in NSW

You might want to keep a copy of your responses so that you can compare them with the collated responses from all teachers who participated in the survey which will be provided to you in due course.

Jim Scott April 2016

*** PLEASE TAKE A MOMENT TO GO BACK OVER THE SURVEY AND CHECK THAT YOU HAVE COMPLETED ALL QUESTIONS BEFORE FINISHING ***

Save and continue later OR Finish >

Appendix G: Interview questions for case study school participants (final)

1. What prompted you to join the case study?
2. What contribution does the EV program make to the assessment-related work done by you or your science teachers?
3. How do you prepare your students for the EV test?
4. Consider a topic you have just finished teaching or are now well into teaching. By what means do you collect evidence of student learning as you work through the topic?
5. To what uses do you put evidence of student learning?
6. What sorts of things do you do in the name of student to student (peer) assessment?
7. What about student self-assessment?
8. What are the main sources of information you access to inform your assessment-related work?
9. What are the school / principal priorities and how do they impact your work as a science teacher in Years 7-9?
10. What are your / science faculty priorities for science teaching in Years 7-9?
11. Thinking back over the past five years, what are the main resources used regularly by you and your teachers to support science teaching and learning in Years 7-9?
12. Of all the things you are doing in the name of science teaching, which is having the most impact on student learning in science? / How do you know?
13. Thinking about the survey you completed (participants were handed a page of five selected questions from the survey to review), how did you decide what seldom, sometimes and often meant?
14. If asked by a parent or new science teacher what “progression in learning science” means, how would you answer?
15. What is the nature and extent of discussion about assessment at science faculty meetings?

If the HT had brought the completed school data proforma to the meeting, the following question was asked:

16. When filling out the form, what response/s from students surprised you the most? Why did it/they surprise you?

Once responses concluded, I indicated that the interview was coming to an end and asked:

17. Was there anything you want to revisit or add before the recorder is turned off?

The interview was concluded by me saying; “Thank you for your time and patience…I hope you found the experience friendly and useful…this concludes the interview and I’m turning off the recorders now”. Once the recorders were off, I explained that when the study was completed I would be providing feedback on the survey results (to all secondary schools invited to participate) and case study summaries to participants.

Appendix H. Assessment related narratives for case study schools used to make pair wise comparisons.

The criteria for comparison are sharing the same SEA score and having different residuals; the more widely different the residuals are, the easier to see differences in assessment-related work narratives.

Pair ONE: Assessment narratives compared for PCWAE1 and MCWAE1

A. Engagement with EV feedback, resources and SOLO

PCWAE1

The provincial teachers participated in the case study to find out why their EV results were better than their NAPLAN results (as they had seen by comparing the proportions of students in each of the EV and NAPLAN bands to state proportions). They were early adopters of the EV program having piloted it in 2005 and trialed it in 2006 before it became mandatory across the state from 2007. They also engaged with VALID10 when it was first offered in 2015 and indicated they would continue with it into the future. There was evidence that they used items from the EV tests in their own assessment tasks, but syllabus criteria rather than SOLO was the basis for marking student responses in the assessment related artifacts they provided.

They admitted that their knowledge and understanding of SOLO was "very low" but they said they looked at their EV results each year. The student survey results were not usually looked at and no evidence was provided that the Year 8 EV feedback was used diagnostically during the years they were reviewing their school program in preparation for the new syllabus being implemented from 2014 (in Y7 & Y9). No comments were made about their experience with SOLO when marking the VALID10 tests.

Whilst there was a long standing and strong focus on scientific literacy and getting students to make appropriate use of scientific terms in reports and explanations, it was not apparently connected by them to the second cycle SOLO levels. Two of the three teachers identified the diagnostic purpose of the EV test in their responses to the teacher survey question asking about the most important purpose of the EV program.

MCWAE1

The school has a considerable refugee intake each year (around 30%). Many students have little, if any, formal education or English language skills before arriving in Australia. Their first experience for these students is in an Intensive Language Centre before transitioning to secondary education when they reach an appropriate level of language proficiency. The HT and three of her staff attended the interview which went over the hour. The school has embraced the EV program, including VALID10 from its inception and see it as a useful resource among many for helping their students to learn science. The teachers look at the results when they come out and report the have used achievement feedback to make changes to their programs. Teachers do not have access to the student survey feedback. In relation to the EV test, they do spend some time helping students to prepare by giving them access to sample questions from past papers. Teachers report that students enjoy doing the test. They wanted to join the case study in order to receive feedback on their assessment practices.

B. Grouping for instruction

PCWAE1

Each year the provincial school established two relatively small (fewer than twenty students is not unusual) Y7 mixed ability classes based on student data provided by the feeder K-6 schools. The two classes go on to Y8 largely unchanged. The school chooses to establish two small classes in each of Years 7 & 8, but then form two combined Y9 – 10 classes which are very large (in excess of thirty students in each).

MCWAE1

Each year three Year 7 classes (with fewer than 20 students) are established based on the level of literacy skills. Classes are ungraded from the perspective of prior science experience or learning. Assistance is provided to classes by learning support teachers to help the high proportion of students with little formal education and very limited English language proficiency. Classes are retained relatively unchanged until the end of Year 10.

C. Use of learning intentions and success criteria

PCWAE1

Analysis of interview responses and artifacts provided by the provincial school established that the teaching program was explicitly based on syllabus intentions as expressed through outcomes and related content.

In program outlines, under the heading Indicators of student achievement, a list of science vocabulary students were expected to acquire was provided as was a list of what students were expected to know and understand, and a separate list of what students needed to be able to do (skills) by the end of the topic. Learning the spelling and meanings of words in the vocabulary list for each topic was the main source of formal homework. Indicators based on the contents of these lists were evident in the criteria included in rubrics /scaffolds for tasks related to the topics being taught.

Teaching activities and related assessment tasks described in the programs were aligned to syllabus intentions. In relation to the lived experiences of students, the teachers commented that about one third of students lived on rural properties, did not recognise the science in the day to day plant and animal husbandry work and the equipment used to do that work. Addressing this disconnection between science and the students' life experience was a priority for the teachers.

Assessment tasks were assigned with rubrics that clearly described expectations based on syllabus outcomes. Teachers said they used the rubrics to both introduce tasks and to provide feedback to students once tasks were assessed.

Two formal pen and paper tests for each of Y7 and Y8, based on syllabus working scientifically content. The tests included free response extension questions that asked students to explain using scientific models (particle model) and to identify and correct misconceptions in examples that were given.

The teachers said they prioritized practical activities and report writing. Students did three research projects across the four years seven to ten (the syllabus specifies a minimum of two). Examples of both laboratory and fieldwork were provided in the artifacts. Expected learning from those experiences was typically scaffolded in a worksheet or modeled using a textbook example. Open-ended questions were evident in those worksheets. The scaffolds were informed by expectations described in the skills section of the syllabus.

Reporting to parents is in grades aligned to curriculum standards (A to E), which is national policy (the same applies to MCWAE1).

Teachers reported that they did not often use ICT in Y7/8 science classes.

MCWAE1

The science department program has four ten-week topics mapped to syllabus outcomes for the four content areas (In Year 7 the topics are Forces, Chemical World, Earth and Space and Living World). Syllabus expectations are also mapped to the eighteen elements of the Quality Teaching Framework and references to the cross curriculum aspects of learning are explicitly identified as well. Syllabus outcomes targeted include Values and Attitudes, Working Scientifically and Knowledge and Understanding. Learning activities are described in terms of lesson outcomes that appear to require from one to a number of lessons to achieve. A diversity of resources are identified to work with including conventional text books (e.g. Core Science, Science Stage 4) and worksheets describing activities to be performed and writing to be done.

Assessment is by conventional topic tests and end of semester tests. Students are provided with a range of options for responding including multiple choice, short and longer response items involving students writing descriptions and or explanations. Some items have interesting stimulus material associated with the item. There is a word puzzle at the end of each test for students to engage with if they finish early. An example of a practical test and a research project scaffold was provided for Years 7 and 8. No rubrics linked to syllabus outcomes or different levels of answers was provided as models for students to work with. Students did not appear to have much, if any say in choosing or devising learning intentions or success criteria.

D. Classroom discourse and evidence of learning

PCWAE1

Teachers reported that group work is common and instruction is provided to students about how to work cooperatively in the classroom, laboratory and during fieldwork (using the local river and a 'wetlands' area).

Teachers reported that they used a predict-observe-explain strategy to focus discussion of practical work and as a preliminary step to writing up a practical report. The teachers reported that school-based learning support staff were regularly invited visitors to their science classes to help students struggling with literacy skills.

Teachers described some of the early work done in topics as opportunities for verbal pre-testing and students were helped to construct mind maps as a way of summarizing their learning.

School policy placed great emphasis on literacy learning as a key to helping all students succeed. The science faculty supported this emphasis in its homework policy (acquisition of scientific vocabulary) and in classroom work where students were supported and encouraged to verbalise their experiences using the appropriate vocabulary early and often.

The teaching programs for Years 7 and 8 were organized into ten week topics (one per school term). The programs also listed resources such as relevant videos, text book sections and excursions which were an annual event for students in Years 7-10. One excursion for Y8 students involved a visit to La Trobe university to raise student awareness of post school options and another which was an extended five day trip to the NSW south coast. Students were required to write reports of these activities.

MCWAE1

Science learning activities provided to students at this school were diverse and included conventional classroom based activities using textbooks and worksheets, laboratory activities involving equipment and report writing, excursions beyond the school gates and visits to the school by people that work in STEM careers (such as CSIRO and Questacon. Some access to ICT is provided in the library for research purposes. Students do ICAS tests and participate in the Big Science competition. Many attend the after-school homework centre and do science homework there, including science vocabulary and spelling related work. There is a heavy emphasis in lessons on talk using scientific vocabulary (whole class discussion is common). There is explicit instruction relating to groupwork and roles to be performed. Students are not confident talkers, especially in Years 7 and 8. Students have a strong preference for rote learning (that seemed to teachers to be related to expectations based on experience brought from other cultures).

E. Feedback

PCWAE1

Feedback on assessment and other tasks often took the form of discussion with students about the rubric criteria and how they were used to allocate marks that mapped onto a five-point scale ranging from unsatisfactory to outstanding.

When asked about progression of learning (one of the questions in the survey) they did not readily relate syllabus outcomes and content with the idea of learning progression.

The school was making less use now of the Educational Assessment Australia (EAA), International and Competition Assessment (ICAS) science tests (20 students sat them in 1999; last year only two did so). Science teachers did not use the results for diagnostic purposes. They did use items from them in their class and assessment tasks. Certificates about participation and achievement were handed out at a regular school assembly.

In the interview, the relieving DP drew attention to the good results in Year 10 (see the proportion of As for the school relative to the state in Table K.3 in Appendix J) and commented that results there did not translate all that well to the HSC, which he found puzzling and for which he could find no explanation.

He also commented on the student's apparent low enjoyment of science lessons compared to the state as something he could not explain (see Table K.2 data). The other two teachers said they had asked students why they didn't like science and were told that it was because "you" (science teachers) followed up to ensure work was completed. This explicit interest in asking students why they did not like science was a response to reading the survey feedback in preparation for the interview.

MCWAE1

Talk in the interview indicated a strong emphasis on oral feedback during lessons largely related to building language skills in the appropriate use of vocabulary related to the concepts and skills and processes of science being taught at the time. Pre-testing was not mentioned. Feedback on tests and work was provided by teachers in terms of marks and discussion of answers. SOLO was not mentioned nor were syllabus outcomes or expectations. Research project reports were heavily scaffolded and teachers reported to the researcher that time was given to explaining what the different components are. Class and home time was given to the projects.

F. Activating students as instructional resources for others

PCWAE1

Peer feedback was sought when oral presentations or models were produced. No other details about that feedback were provided.

MCWAE1

The only formal opportunity for that appeared to be during groupwork in the context of practical work in the laboratory. Teachers commented that the classroom set up did not support using think-pair-share strategy.

G. Activating students (and teachers) as learners

PCWAE1

This mostly took the form of teacher-led discussion about student work in the light of teacher provided rubrics. The rubrics described a range of responses showing the features of responses that achieved high marks. The rubrics were inevitably related to syllabus expectations. Intensive literacy work with students in Y7 & Y8 was followed up in Years 9 & 10 with expectations that students would use those skills to work independently whilst their teachers were providing time to various groups in the class, given that students were not only mixed ability but across two grades. Self-assessment opportunities were provided as early as in Y7 and teachers provided feedback on it (see Figure 6.1).

Example 1: PCWAE 1

Name: ____________________________

Project: ____

Purpose of toy: ____

Are you happy with your final project? Why/Why not? ____________________

What are some things you did really well? ______________________________

What are some things you could have done better? _______________________

Thanks, Good job

Example 2 PCWAE1

Marks Outstanding --6 // High -- 4-5 // Sound --2-3 // Basic/limited --1

Justification (opinion + reasons) Clear record of changes made to toy and a reason for each change. Overall justification of final design and product (toy) Some record of changes made and reasons for these given. Brief justification of final design and product (toy) Minimal record of changes made with little or no reasons given. An attempt made to justify final design and/or product (toy) An attempt made to justify either their design or final product or any changes made

Figure 6.1 Opportunities for self-assessment in Year 7 Making a Toy task

The teachers said they met regularly both informally and formally to work on aspects of science teaching and related assessment tasks, which were often collectively marked. They were clearly enthusiastic about their work. Faculty programs provided were comprehensive and whilst they did not have space for written evaluation, it was clear from the discussion that the new programs at the school for the new syllabus had been collaboratively developed.

MCWAE1

The teachers met weekly assessment was often discussed they said. The development of the teaching programs was a shared activity. Teachers participated in the interview and were supportive and respectful of each other in that discussion.

H. Comparative summative comments

When results were compared at the end of Years 8 and 10 the provincial school had the best results. Also they had a higher proportion of senior science course completions (as a proportion of the students at their school). The comparisons made here supported the three predictions, even though the schools were both WAE schools. .

The assessment narratives from both schools, when compared, revealed that in the early years of secondary science teaching both schools made use of a variety of contexts for teaching science which in turn meant that students had opportunities to provide evidence of learning. Teachers at the provincial school made greater use of rubrics related to scaffolded tasks and they provided feedback during and after completion. The feedback was in terms of syllabus expectations and marks awarded as recommended in the Board's Common Grade Descriptor outline. There were more opportunities at the provincial school for peer- and self-assessment. Summative assessment at the metropolitan school was more strongly linked to traditional testing than at the provincial school.

Whilst the three predictions lend weight to the conclusion that teaching at the provincial school was more closely aligned to the formative practices profile of WAE schools as identified in chapter five, the level of engagement with science when compared to both the metropolitan school and the state was not in line with expectations for self-regulated learners (the expected outcome from teaching characterized as formative as discussed in chapter two). Overall, students at the provincial school were less positive about their school science experience than their metropolitan counterparts.

It was impossible to identify from the assessment narratives why students at the provincial school had such poor perceptions of their school experience of science at the end of Y8. The teaching program at the metropolitan school, compared to the provincial school, was more like that described in the left hand column of Table 2.1, yet students at MCWAE1 were the most positive about their school science experience of all the case study schools (see Table K.5 in Appendix J). Teachers at the metropolitan school said that even though parents did not come to the school often, they were aware of strong support for teachers and learning by parents who often bought text books for students to keep and use at home.

Pair TWO: Assessment narratives compared for MCAE2 and MCWBE3

A. Engagement with EV feedback, resources and SOLO

MCAE2

The principal was keen for the school to be involved in the case study and expressed interest in any feedback to come out of the process. The head teacher science was also the relieving deputy principal (R/DP) at the time of the interview and the only person at the interview. He had been at the school in the head teacher position in the period of interest for the project. The science department had not engaged with SOLO but were focused on syllabus outcomes and the Board's approach to grading. The R/DP reported that students took the EV test seriously and appeared to enjoy the experience. The school provided no special preparation for it. The school had engaged with VALID10 and were planning to continue with it. The school had not done the proforma or collected artifacts prior to the meeting.

MCWBE3

The metropolitan school did not take up VALID10 in 2015 and it had no plans to do so in 2016. The HT reported that when she had arrived at the school, the science staff had very limited understanding of assessment for learning and had not made use of EV feedback at all. The HT said that she and another new staff member who had arrived at the school in the same year were the only ones who knew anything about SOLO which she characterized as "all about" recognizing "connections."

The focus for now, she said, was on improving teaching and learning practices in junior secondary science and making use of data (from assessment generally and SOLO in particular) to target resources to that end.

There was strong evidence in the artifacts of a focus on scientific literacy and appropriate use of scientific terminology in reports. However nothing was said by her to link this to second cycle SOLO responses. This emphasis appeared only to be recent (i.e., after the HT's arrival at the school and after the period of interest).

B. Grouping for instruction

MCAE2

The school was promoting itself as a school with a special interest in STEM broadly and biosciences in particular. Each Year the school provides a 'selective entry' test for local Y6 students that includes science questions as well as general ability and literacy and numeracy skills. That class is provided with an accelerated program and complete the four year science course by the end of Year 9. The other four classes are unstreamed and students remain in their class until the end of Year 8. The R/DP indicated that this was consistent with a deliberate 'middle school' approach to the first few years of high school aimed at providing support and stability for students to assist them with transfer from primary to secondary education. Some twenty students each year are provided with additional learning support assistance.

MCWBE3

The metropolitan school established six or seven classes (depending on numbers to be enrolled) in Y7 each year. The established classes are the same for the four core subjects and remain relatively unchanged until the end of Y8. Students are allocated to classes based on student data from the feeder primary schools. Two parallel high achiever classes and four or five mixed ability classes are created by the Y7 adviser and other staff (not science) at the school. Changes when they are made are negotiated across the faculties using a diversity of criteria but typically they are unrelated to science assessment results.

C. Use of learning intentions and success criteria

MCAE2

The R/DP did show me some tasks students in Years 7 and 8 were given. Learning intentions and success criteria based on the syllabus were a major focus in the research and other tasks students engaged with and they informed the assessment rubrics used by teachers to mark them. Evidence of learning was primarily gathered from these tasks and used as the basis for reporting to parents twice yearly. The science teachers provide a 300 word report on science achievement twice a year to parents. The reports include specific references to teacher observations of students work to illustrate aspects of achievement relevant to the reporting categories addressed in the rubrics.

The school retained a Year 7 annual test, but most of the evidence of learning comes from 4-5 tasks students do each year, one of which is a practical task. The tasks put a strong focus on science processes and try to engage the students by making them relevant to student interests, including making models (eg parachute and egg drop activity) and explaining them to other members of the class. Teachers provide assessment feedback. Teacher judgment of learning is conveyed in marks which are then translated into grades for the purpose of reporting to parents.

No sample programs were provided, but the description provided was of four to five STEM / Bioscience topics in each of Years 7 and 8 “are identifiers of the school.” These are cross curriculum courses including PDHPE, HSIE and R/DP wrote these programs. Each department provides two hours in a fortnightly cycle for this program. Recognised a need to strengthen understanding and awareness of the scientific method and ability to investigate scientifically and this has led to the shift to inquiry / project based learning emphasis. Success seen in terms of growing number of students taking up senior science courses. Students are taking up school courses in Year 9 & 10 (courses in forensics and zoology) in good numbers.

MCWBE3

The HT described the faculty culture she had inherited as “traditional and resistant to change”. When she arrived at the school she observed a “wide spectrum of learners” at the school but few differentiation strategies in science programs for meeting those needs. As the interview progressed, her grasp of what those strategies could be was elaborated by reference to resources she had developed with staff at the school. Artifacts of this new work were clearly aligned with syllabus.

At the time of her arrival staff were not keen to “do more than required” and none of them had marked either HSC or EV extended response tasks. She observed that when she first arrived staff worked individually to produce assessment tasks which were then individually marked. She also said that at the time of her arrival staff had a poor understanding of the BOS Common Grading Scale and their processes for translating marks into grades were unrelated to syllabus standards and thus inconsistently arrived at across the classes.

She said that it was her observation that students started science in Y7 looking forward to and liking science but were “disengaged” by the end of Y8.

Teaching programs in the period of interest were for topics lasting five weeks (now they are ten weeks). The HT was not happy with the school science programs she had inherited, which in her view were “all over the shop” and had been developed as a joint project with several other schools. Staff did not have enough understanding or willingness to do a scope and sequence for new syllabus topics. One of her first actions was to persuade staff to work on creating / collecting resources for new “scope and sequences” (teaching program outlines mapped to syllabus intentions) which she and the other new teacher had developed for years 7 and 9 soon after their arrival at the school.

There was very little use of ICT in Years 7 and 8 science classes. Prior to 2015, worksheets provided by teachers and textbooks were the main resources used to support teaching and learning she said. Very little work was done outside the classroom then and there were no science specific excursions before she arrived.

The HT has prioritized getting more students to think (in science classes) and to take senior science courses and she is doing that by building the teaching and assessment skills of her staff. There was no mention of linking of SOLO levels to the discussion about what the teaching of thinking might involve.

Literacy and writing in particular are school priorities which the HT says they are embracing now in science and making good use of EV extended response tasks to that end.

D. Classroom discourse and evidence of learning

MCAE2

Reportedly, learning tasks are assigned in class and worked on in students own time as well. Groupwork is encouraged and supported. Teacher observation of student teamwork skills as well as their individually written reports provide evidence of learning. Whole class discussion is strongly encouraged; the use of think-pair-share-report like strategies appear to be used in some classes and some reflective writing by students is encouraged. Students were provided with a diversity of tasks, most draw from a wide range of contexts and other learning areas (see above). The school engages with the Young Scientist competition and provides some students with the ICAS science tests as well, but little was done with the feedback apart from providing the certificates to students when they were returned to the school. The school engages with National Science Week and puts on activities for feeder primary schools.

MCWBE3

Information about classroom practice before 2014 was anecdotal but the HT referred to heavy use of textbooks, worksheets with limited opportunities for extended written responses and conventional laboratory practical work designed to confirm syllabus prioritized theories. Practical work was conducted in groups, but the HT reported that she had little evidence of purposeful use of group work for peer supported learning.

The school has a learning support unit and a number of students are receiving support from its teachers.

The timetabling software used delivers a number of split classes in Years 7 and 8 (typically one class shared between two teachers) and a number of classes were and still are taught in the junior secondary years by PDHPE teachers.

E. Feedback

MCAE2

The R/DP reported recent engagement with Hattie's *Visible Learning* (2009). It was not clear how far back into the period of interest was influenced by this. Marks were given related to rubrics based on syllabus outcomes, but teachers also gave written feedback explaining why the mark was given and suggestions for better answers were also provided. Feedback was provided against rubric criteria in the context of classroom work either one-to-one with the teacher or whole class managed by the teacher.

MCWBE3

The HT reported that when she had arrived two years ago, feedback to students from assessment and other tasks was basic and involved reporting back of marks with, as far as she could ascertain, little discussion or diagnosis. Assessment then was dominated by end of topic tests and marks were recorded and used for reporting summatively. Tasks completed in class were simply marked and handed back with rudimentary discussion (if at all).

EAA/ICAS tests were and still are offered to the top two classes but results are not used for diagnostic purposes.

F. Activating students as instructional resources for others

MCAE2

Opportunities to provide feedback to peers appeared to be limited to groupwork during classwork. Some students were also involved in demonstrating to Year 6 students during Science Week activities.

MCWBE3

The HT was not aware of any peer assessment opportunities being provided in science classes prior to her time at the school.

G. Activating students (and teachers) as learners

MCAE2

Some opportunities were provided in some classes for self-reflective writing, but no information about follow-up was provided. The R/DP indicated that there were regular science department meetings (once a fortnight) and that assessment and programing were discussed. The science department had a STANSW membership and staff participated in marking of external exams (HSC) and participated in other professional learning activities related (most recently) to Hattie's *Visible Learning* program. From that it could be inferred that staff modelled good learning behaviours with each other, but the extent of that modelling for students was unclear from the information provided in the interview. Also, there was awareness of the need for differentiated curriculum to meet the diverse needs of talented students and different approaches for the twenty (estimated) students in the junior secondary years who were on a modified program (which were not individual life-skills programs). Characteristics of that differentiation were not provided.

MCWBE3

Again, at the time of her arrival, the HT reported that there was little evidence of any self-assessment activities or strategies in use. Staff tended to work independently and it has been a slow process upskilling them in assessment literacy since then. Teams have been established within the science faculty to facilitate cooperative development of programs and related resources. This collaboration, she reported, had been effective in raising staff awareness and understanding of assessment issues.

H. Comparative summative comments

MCAE2 was a new school that had at the time of the interview only had its full complement of students from Year 7 to 12 for a few years. In that time, it had deliberately sought to establish a STEM/bioscience identity for itself and provided students with a learning program that reflected that emphasis. MCWBE3, on the other hand was a well-established school that provided its students with what was reported by the new head teacher as a "traditional" program.

For prediction one, when achievement at the end of Years 8 was compared, MCAE2's results across all four result categories were positively biased toward the top band of achievement more so that those at MCWBE3. Both schools had a top stream of students, which may account for the positive skew in both sets of results compared to the state. However, MCAE2 results at the end of Year 10 were slightly positively biased, but could not be compared with MCWBE3 because that school did not provide any Year 10 results to be compared.

In relation to engagement at the end of Year 8, given the priority given to STEM at MCAE2, the level of student engagement (as measured by the combined scores for Items D and E), compared to MCWBE3 were not that different. Top band students at the AE school were only two places higher than the WBE school (9th and 11th respectively out of sixteen schools. The state score was counted as a school; both ranked below the state score (2nd out of 16). The rankings (out of 12) for the total school results were the same (9th and 11th respectively and compared to the state which ranked 5th). Based on the assessment narratives derived from the interviews at both schools, this was an unexpectedly close result, particularly for the AE2 school, which should have returned a more positive result.

Engagement was assessed by looking at Year 12 completions relative to the state. In this comparison (see Table K.4 in Appendix J), at the AE school, Biology completions were 200%, Chemistry just over 100% and Physics was 63% (neither school offered Senior Science). By comparison with the state, the WBE school Biology proportion was 74%, Chemistry was 39% and Physics was 56%.

The figures above support the first prediction; no conclusion could be drawn in relation to prediction two and prediction three was supported.

It was impossible to identify from the assessment narratives why students at both schools had such poor perceptions of their school experience of science at the end of Y8.

Pair THREE: Assessment narratives compared for PCWAE2 and MCWBE5

A. Engagement with EV feedback, resources and SOLO

PCWAE2

The provincial school was ambivalent about the EV program. On the one hand the HT said it was useful for both diagnostic purposes and comparative purposes but did not elaborate on how. On the other it transpired in the interview that the science staff had a negative view of its contribution to the assessment practices at the school (apart from its value in showing the comparative strength of their EV results compared to NAPLAN).

The HT was concerned about the validity of VALID10 because they had recognised "rehashed" Y8 ESSA questions it. They thought that school marking reduced its value for comparative purposes. They would not be doing VALID10 this year (2016) saying that the school had computer access issues, that she would not be at the school in Term 4 and the science teachers did not see the value in it.

The teacher who was to be relieving HT (for the next twelve months) had joined the interview toward the end. She reported that neither she nor the other staff could see the benefits of using SOLO as a basis for assessment because it conflicted with the Board's grading system and students had found it confusing to deal with both systems.

EV results are handed to parents at the first parent-teacher night of the year. The HT reported that:

- parents don't ask questions about the EV test;
- there is no special preparation for the test;
- students like the online science test and take it seriously, as they do NAPLAN;
- the school is focused on results and the principal is "happy" with science results generally and their EV results in particular.

The HT was aware from her own analysis that the school's EV results were better than the schools NAPLAN results but did not elaborate how she had arrived at that conclusion.

MCWBE5

The HT said he agreed to participate in the case study to have a say about the EV program which he saw as problematic. Reasons given included giving science a special status which he was personally uncomfortable with, issues with access to computers (since it went online), science staff not keen to supervise it and a school executive which he said was not interested in the results.

He acknowledged that the test provided good questions which he said were used in their own school tests. He said that the faculty was not given any time by the school to digest EV feedback (compared to NAPLAN results).

He expressed regret at the loss of the Y10 statewide science test (stopped after 2011) because it provided a target (grade pattern) to aim for at the end of Year 10 but also said that they would not be taking up VALID10.

B. Grouping for instruction

PCWAE2

The HT science at the provincial school is responsible for managing the composition of classes for Science, PDHPE and Social Sciences. Students are initially placed in three mixed ability classes using primary school literacy and numeracy data. After six months, students are reorganized into separately graded classes for English, Mathematics and Science based on summative assessment results in each of the subjects for semester one.

The top class in Science has close to thirty students in it; the bottom class has around twenty students in it and is provided with learning support. There is a six monthly review of class placements in Science and students are moved if their performance changes warranted it (either up or down). This potential for changing classes continues up to and including Y10. The process is supported by the HT Science who describes herself as a traditional science teacher.

MCWBE5

From day one in Y7 the metropolitan school places its new students in classes according to four different sets of criteria. A top class of "gifted and talented" students, a second class of "independent learners", two or three classes (depending on numbers) of mixed ability students and a bottom class of students with learning disabilities and otherwise poor learning histories. These classes are the same for English, Mathematics, Science and Social Sciences and they remain in those classes up until the end of Y8. At the end of Y8 all students are graded on the basis of their results from a common assessment task (typically a test) and put into a class based on their rank in the year. They typically stay in that class for Years 9 & 10.

Students are invited to join the "gifted and talented" class on the basis of their results in a test they applied to sit for in Y6. The test was set by the secondary school and did not include any items related to scientific literacy. Students are allocated to the "independent learners" class on the basis of advice from their Y6 teachers. An interesting feature of this school is that it only admits 12-15 students to the top and bottom classes each year. They stay in the class for two years and they have the same Science teacher for the two years.

No explanation or commentary about the merits or otherwise of setting up classes in this way was offered by the HT science. He did say that the bottom class was provided with additional support from time to time by learning support teachers to improve literacy and numeracy levels.

C. Use of learning intentions and success criteria

The science faculties from both schools have high profiles in their local communities and how this is achieved by each of them will be described below.

PCWAE2

The HT explained that from Y7 the policy is to expose students to high expectations in relation to using the language of science and there is close alignment between syllabus intentions, teaching and assessment in the work of the faculty. There are consequences for students who perform very well or not so well. They may be promoted or demoted a class at the middle or end of the year (for the new school year).

The provincial students perform consistently well in local, high profile community agricultural events such as region-based "Hoof and Hook" competitions which are well publicized in the local press. In any given year, teachers of Agriculture are very busy with activities such as the above that take them outside the school during the school day, after school and on weekends. The HT reported that the Science, Agriculture and PDHPE faculty was the "strongest" performing faculty group in the school. There is a strong emphasis in the faculty on competition as the way to get the best out of the students.

In recent times, with the exception of local agricultural events, the school has been withdrawing from general science and technology based excursions and activities beyond the school due to the costs (one example mentioned was the withdrawal from the University of Newcastle's Engineering Challenge). Instead, local resources are increasingly being relied upon (such as having a local Aboriginal elder in to talk to students). According to the HT the school was not overtly responding to the recent STEM initiative by the Department as staff at the school have for some time been using agricultural contexts to create interest in science based careers (Artificial Insemination for cattle and Genetic Modification for Canola seed were given as examples).

The science faculty is heavily invested in the schools literacy program and contributes a period a week (as do the other three core learning areas) to generic literacy activities provided by staff from all faculties in the school. The HT science said she is lobbying for more report writing to be included in the program.

The HT said her priority was to maximize participation in science in the senior school. To that end programming had been pared down to four topics a year with titles such as Biology7, Chemistry8 and Physics9 so that students know what the content of the senior subjects is when they choose them in the second half of Y10.

The faculty has four assessment tasks per semester, two of which are formal tests (down from more than a dozen over the year when she had first arrived). The intention behind the reduction in assessment tasks was to provide more time for teaching and she reported that since doing that, results have improved.

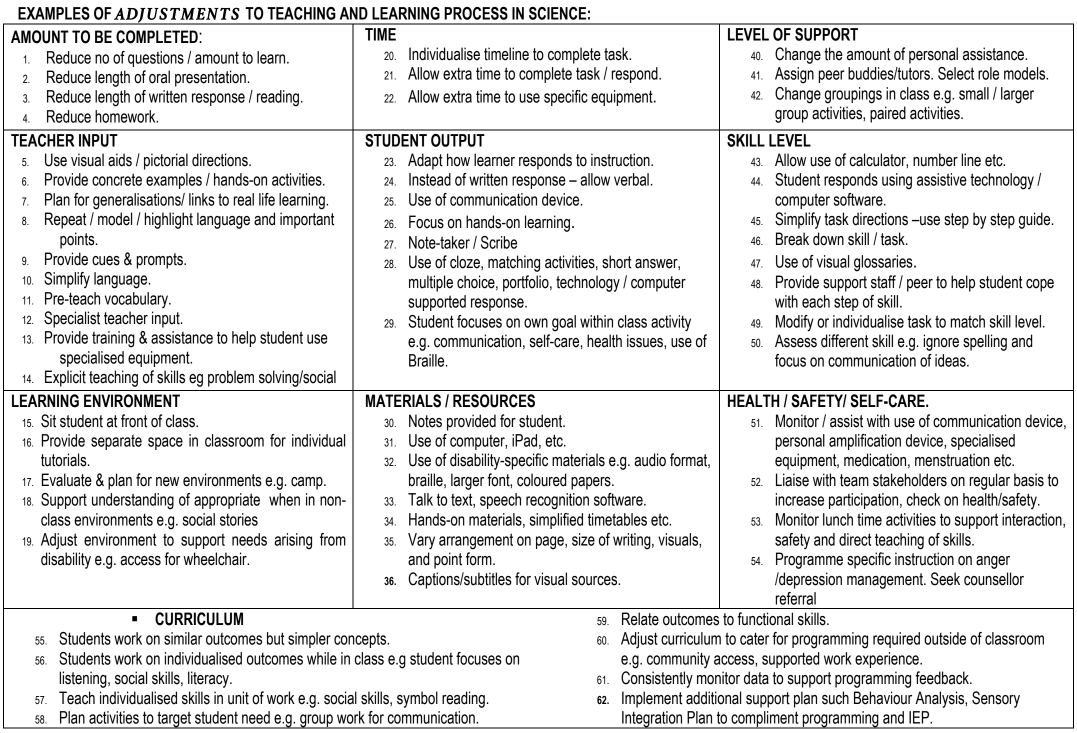
Students do a research project each year which is allocated both school and home time to be worked on. This is more than the syllabus requires (it suggests at least two be done in the four-year program). None of the other case study schools ran a major research project each year. The student research projects are heavily scaffolded to ensure that a traditional report involving an aim, problem, variables, method, results of observations (tables and graphs), conclusion, discussion and bibliography is produced as the expected product.

Teaching programs are organized around syllabus knowledge and understanding outcomes and related content. Investigating and communication skills are addressed in the rubrics for the various tasks embedded in the program. Those tasks are both teaching and learning activities as well as assessment tasks. Among the artifacts provided was a Y7 task requiring students to produce a poster showing how to separate a mixture (one chosen from a number of actual examples within the experiences of students) and another task requiring students to produce a written report on a topic (eg heart transplants) relevant to the Y8 Biology and Society topic.

Teaching and learning programs also list the resources available to do the task which includes traditional text books and worksheets. A column is provided for teachers to add any adjustments they have made to the listed program. To assist with this task, teachers are provided with a one page summary of suggestions for adjusting teaching to ensure that students have access to syllabus outcomes (see Figure 5.2).

Assessment tasks are supported by rubrics that spell out expected learning and how responses will be marked. They are based on the Board's *Common Grade Scale* and marks are awarded in line with rubric criteria and discussed with students. Collated marks are aggregated and recorded and teacher judgment is used to convert marks to grades for the purposes of reporting to parents. Staff are given time to work through the criteria to ensure some consistency of judgment and subsequent marking is shared to further support that.

Students use notebooks to keep a record of their learning activities. Worksheets are expected to be stuck into their notebooks which are expected to be brought to every lesson. Monitoring of bookwork by teachers is not a high priority but they do encourage assist students to peer assess each others bookwork (see below).



EXAMPLES OF *ADJUSTMENTS* TO TEACHING AND LEARNING PROCESS IN SCIENCE:

AMOUNT TO BE COMPLETED: 1. Reduce no of questions / amount to learn. 2. Reduce length of oral presentation. 3. Reduce length of written response / reading. 4. Reduce homework.	**TIME** 20. Individualise timeline to complete task. 21. Allow extra time to complete task / respond. 22. Allow extra time to use specific equipment.	**LEVEL OF SUPPORT** 40. Change the amount of personal assistance. 41. Assign peer buddies/tutors. Select role models. 42. Change groupings in class e.g. small / larger group activities, paired activities.
TEACHER INPUT 5. Use visual aids / pictorial directions. 6. Provide concrete examples / hands-on activities. 7. Plan for generalisations/ links to real life learning. 8. Repeat / model / highlight language and important points. 9. Provide cues & prompts. 10. Simplify language. 11. Pre-teach vocabulary. 12. Specialist teacher input. 13. Provide training & assistance to help student use specialised equipment. 14. Explicit teaching of skills eg problem solving/social	**STUDENT OUTPUT** 23. Adapt how learner responds to instruction. 24. Instead of written response – allow verbal. 25. Use of communication device. 26. Focus on hands-on learning. 27. Note-taker / Scribe 28. Use of cloze, matching activities, short answer, multiple choice, portfolio, technology / computer supported response. 29. Student focuses on own goal within class activity e.g. communication, self-care, health issues, use of Braille.	**SKILL LEVEL** 43. Allow use of calculator, number line etc. 44. Student responds using assistive technology / computer software. 45. Simplify task directions –use step by step guide. 46. Break down skill / task. 47. Use of visual glossaries. 48. Provide support staff / peer to help student cope with each step of skill. 49. Modify or individualise task to match skill level. 50. Assess different skill e.g. ignore spelling and focus on communication of ideas.
LEARNING ENVIRONMENT 15. Sit student at front of class. 16. Provide separate space in classroom for individual tutorials. 17. Evaluate & plan for new environments e.g. camp. 18. Support understanding of appropriate when in non-class environments e.g. social stories 19. Adjust environment to support needs arising from disability e.g. access for wheelchair.	**MATERIALS / RESOURCES** 30. Notes provided for student. 31. Use of computer, iPad, etc. 32. Use of disability-specific materials e.g. audio format, braille, larger font, coloured papers. 33. Talk to text, speech recognition software. 34. Hands-on materials, simplified timetables etc. 35. Vary arrangement on page, size of writing, visuals, and point form. **36.** Captions/subtitles for visual sources.	**HEALTH / SAFETY/ SELF-CARE.** 51. Monitor / assist with use of communication device, personal amplification device, specialised equipment, medication, menstruation etc. 52. Liaise with team stakeholders on regular basis to increase participation, check on health/safety. 53. Monitor lunch time activities to support interaction, safety and direct teaching of skills. 54. Programme specific instruction on anger /depression management. Seek counsellor referral
▪ **CURRICULUM** 55. Students work on similar outcomes but simpler concepts. 56. Students work on individualised outcomes while in class e.g student focuses on listening, social skills, literacy. 57. Teach individualised skills in unit of work e.g. social skills, symbol reading. 58. Plan activities to target student need e.g. group work for communication.	59. Relate outcomes to functional skills. 60. Adjust curriculum to cater for programming required outside of classroom e.g. community access, supported work experience. 61. Consistently monitor data to support programming feedback. **62.** Implement additional support plan such Behaviour Analysis, Sensory Integration Plan to compliment programming and IEP.	

Figure 5.2 Advice on adjustments to teaching to accommodate student differences

MCWBE5

Science teachers at the metropolitan school promoted science at the school by running some of their assessment tasks as "shows" in the playground at lunchtime and in the lead up to National Science Week. The Science Faculty also put on displays using students as demonstrators at the school's annual open night for parents of prospective students. High performing students are also involved in putting on science shows for students in the local feeder primary schools as well. The school has a strong reputation in the community for science according to the HT science which he supported by reference to EV student survey feedback (see Table 5.11 and related analysis) and results from a Y11 student survey conducted by the principal.

The HT's priority for science is that students enjoy the subject. The way he says this is achieved is by giving a priority to practical activities both inside and outside the classroom and reducing assessment pressure. The science program takes students into the playground and local bush from Y7 to Y10. Activities include observations using data loggers and sample collection for further examination and analysis back in the lab. Also, science takes students away for day-long excursions at the end of the year to Taronga Zoo (Y7), Physics is Fun at Luna Park (Y8) and the Aquarium at Darling Harbour (Y9). Learning / assessment tasks include model making and investigations as well as traditional practical tests, research tasks, problem solving and communication tasks.

The sample programs provided to me were written into the Board's programming template and included science knowledge and understanding outcomes and related content but none of the syllabus skill outcomes were explicitly referenced in the programs. Assessment tasks were identified by a title and some additional information about content and skill expectations was provided in the second column under the heading Teaching, Learning and Assessment to assist with developing criteria for assessment.

The faculty programs were used by teachers to plan teaching programs for their classes. The sample programs provided were both for five week topics, suggesting that there were eight topics for the year. Progress is reported separately for the top and bottom classes. The independent learners and other classes are separately assessed. Progress for all groups is reported in terms of grades. However the grade referencing is not done using Course Performance Descriptors. Instead they are referenced to different criteria for the top class, the independent learners and mixed ability groups and the bottom class.

This was the case up to the end of Y8 after which classes are created based on achievement assessed by a common test and task at the end of Y8 and progress thereafter is reported in terms of a grade and place in the year.

Artifacts provided included rubrics with criteria for awarding marks. The criteria included references to science syllabus knowledge, understanding, skills and scientific literacy expectations. There was no evidence that the Board's Course Performance Descriptors or *Common Grade Scale* were used to assign grades and there was no mention about processes used to ensure consistency of teacher judgment in the awarding of grades (for the three or four classes where this was relevant).

There was no mention in the interview of SOLO being used for assessment purposes and it was not evident in any of the artifacts provided. SOLO was not mentioned in the context of the ongoing faculty program review that began several years ago with the introduction of the new syllabus.

In relation to class tasks and the research project, there was no scope for student choice in what they would do or how it would be presented. It was not clear to me whether these tasks were used by all classes or only the middle group (excluding the top and bottom classes).

D. Classroom discourse and evidence of learning

PCWAE2

Underpinning the teaching at the provincial school is a coherent approach to improving general literacy (a strong school priority) and the scientific literacy skills of students.

The bottom Year 7 class receives extra attention from learning support teachers. Oral discussion is a core activity and learning activities are structured to allow students to respond in different ways according to their level of skill. Marking rubrics are related to syllabus outcomes and related content indicators which are shared with students and used to inform oral and written feedback. In this school, streamed classes are used to differentiate teaching and to challenge students at all stages to do better.

According to the HT, there is a strong emphasis on groupwork and students are supported to do this in productive ways through role differentiation and rotation of roles in practical work. The HT gave extended examples of what that differentiation looked like across classes. Classroom activities are differentiated to provide students of all skills and capabilities with a chance to succeed. Worksheets provide scaffolding that ranges from cloze passages to open ended tasks where explanations are expected. Students respond as they can and are assessed by their teachers accordingly.

Oral discussion is the initial go to activity, but it can be used for pre-testing and to engage students who have difficulty accessing and constructing written texts. The HT uses oral reading as a strategy to get students to engage with written text. She encourages students to stop and ask when they don’t understand what they are reading and she constantly probes to ensure understanding. Pauses are opportunities for discussion and sharing, but there are strict protocols observed in the process to ensure no one is humiliated. She argues that having graded classes helps in this because students in the class have similar issues and it is easier to manage when the differences in ability are not so marked.

MCWBE5

The HT here is not so hands on with junior classes and spends most of his teaching time in senior physics classes. He strongly encourages practical activity in junior classes and commented that two recent staff changes have been helpful in having that further implemented. Staff are given freedom to teach their classes as they see fit.

He explained that assessment evidence was being taken from a greater diversity of tasks now than in the past including practical exams (stations set up and students move from one to the other and record in a worksheet what they observe and find), communication and problem solving tasks. Communication tasks involve engaging students with videos on the school Intranet and getting them to provide both oral and written reports. He was particularly proud of model making tasks (a plant cell for Y7 and a toy car for Y8 that goes fastest or farthest and plans for a bungy-jumping “barbie doll” for Y9) because of the opportunities it provides for student engagement in the assessment process (see later section). Model making and related activities have been a feature for many years in the science faculty.

The student research projects (one in Y8 and the other in Y10) are mostly done individually and at home and they are highly scaffolded with a rubric provided by teachers that emphasizes aspects of scientific reports and method. Little information was provided in the interview about the follow-up or support provided to students whilst they were expected to be working on these tasks.

The extent to which support teachers were used to assist learning in the lowest class was not explained in interview. Descriptions of activities used both in and out of the classroom were reported and evidenced in the artifacts provided.

E. Feedback

PCWAE2

In addition to the extensive use of oral feedback during classwork, feedback on written work is provided to students in the form of ticks and crosses to indicate aspects of tasks addressed well or inadequately (or incorrectly). Other feedback is in terms of the Board’s *Common Grade Scale* the language of which students are introduced to in class task and assessment rubrics. It is used to provide feedback to students for both teaching and assessment purposes. The intention is that students are very familiar with it and can use it to self-assess by the time they get to the senior years.

MCWBE5

Teachers are encouraged to provide student with a diversity of activities to support enjoyment and spontaneity in science. A great deal of professional judgement is exercised in assigning grades for reporting in the first few years of science at the school. There are effectively three separate reporting streams based on class placements from primary school assessment of student ability (see above). Content coverage and misconceptions encountered seem to be the basis for feedback to students rather than strict adherence to syllabus outcomes. A creativity / originality mark is also available for models that are made in class.

Student research tasks are for the most part undertaken independently by students working at home. Scaffolds set out expectations in relation to doing the activities in the tasks which are strongly aligned to the syllabus working / communicating scientifically outcomes. How teachers support students as they work on these tasks was not explained.

The HT expressed a concern that senior students did not do very well in the HSC extended response questions because they could not “write a paragraph”. He used the term “backwards mapping” to explain that students needed to be taught to write early on in science. He went on to explain how he was actively working now with his teachers to do more about this in Y7 science. He referred to two literacy programs (TEEEC and the Super Six) that informed the science faculty work in this area. This focus on assessment for learning and literacy appeared to be recent and as a response to new school priorities.

He and his staff as far as I could ascertain had not engaged with the extended response tasks in the EV program, but did freely use the short response items.

F. Activating students as instructional resources for others

PCWAE2

The researcher approached this by asking a direct question about opportunities being provided for peer assessment. The HT said that it was not a formal practice in the early years of secondary schools due to student's natural reticence and lack of confidence related to low literacy abilities. One activity that was used by the HT was to engage students in joint construction on the white board of notes summarising science work. Year 7 and 8 students are invited to write, say, their conclusion on the white board and the class engages in teacher managed discussion to reach a consensus view on what should be recorded.

Students were encouraged to work in groups on practical tasks and support was provided to assist in this process. There was some peer feedback encouraged on student record keeping in their note books too. Students provided each other with a ticked checklist based on their assessment of each other's notebooks (criteria were categorized as positive such as neatness and completeness and negative including graffiti, torn pages and uncorrected spelling errors. At this point the HT spoke about the high absentee rates of students and the fact that some of that was due to suspension from school for inappropriate behaviour. For some students continuity in their school record was an issue that she said impacted over time on achievement. Student involvement in assessing bookwork for each other was an attempt to underscore the importance of having a continuous record of work to study from.

MCWBE5

The HT explained that peer feedback on oral presentations related to 3D models produced by students was encouraged and supported.

Some guidance was given in relation to criteria that should be used (evidence of same in artifacts provided). The teacher retained control over the mark awarded, but there was some discussion with peers about what that should be. His comment was that kids were, on the whole, pretty good at it once they had the criteria provided and they were consistent as well as fair with each other.

G. Activating students (and teachers) as learners

PCWAE2

The science programs in the early years here were very teacher driven. Students were given few opportunities to choose what they studied. They could choose from a range of industrial processes when it came to researching separating mixtures (a Y7 task) but they had to produce a poster. A biology topic task provided a list of three procedures that could be researched, but it had to be presented in the form of a written report (Y8). Student research projects were tightly constrained both in topic (seed germination for Y7) and expectations for presentation (scaffold for the written report).

In terms of teachers being activated as learners, the HT was full of praise for her staff (four full time teachers and one casual who was not science trained). She said of them that they were the "most cohesive collaborative staff [she had] ever worked with." They had engaged willingly with the tasks involved in redoing programs for the new syllabus, took on VALID10 but found it wanting, were fully committed to getting the best from their students and engaged frequently in professional dialogue on teaching, student and assessment issues. When asked about what they thought "progression in learning science" meant, both the HT and soon to be relieving HT were able to give a good account each using a different example. The HT elaborated using investigation skills and described how that might look for different "ability" students. Both demonstrated a good understanding of differentiation in relation to syllabus outcomes.

MCWBE5

The HT reported that in recent years there has been more willingness by staff to meet to discuss professional issues such as assessment. There had been time spent collaborating on the development of new programs as well. Sample programs from

2013 and 2016 were provided showing changes but it was not clear to me whether these were written by staff other than the HT. I was not provided with specific outcomes from any of these reported recent meetings. Artifacts provided included the following scaffold for Y7 students to self-assess (Figure 5.3). This too appeared to be a recent initiative (post 2014).

Student Self Evaluation	Rate each statement out of 10
This is my best work	10 9 8 7 6 5 4 3 2 1 0
I understood this task	10 9 8 7 6 5 4 3 2 1 0
All criteria have been met	10 9 8 7 6 5 4 3 2 1 0
I am proud of my work	10 9 8 7 6 5 4 3 2 1 0

Figure 5.3 Self-assessment rating scale

H. Comparative summative comments

Two things stood out in this comparison. The first was the strong focus on instruction aimed at improving the literacy skills of the students at the provincial school which the science department strongly supported in their science programming and lesson delivery. There was apparently no such emphasis at the metropolitan school. The focus there was on engaging students with a diversity of science activities designed to engage and interest students. The goal at the provincial school was to prepare students for senior science options.

The second was the high stakes assessment policy that graded provincial students in science from the end of semester one in Year 7 and moved students at the end of every semester thereafter either up or down a class if performance warranted it. Semester tests for all classes played a role in that. However, the attention to differentiated curriculum delivery was most thoroughly demonstrated by the provincial school here compared to all the other case study schools. The metropolitan school also established two high achieving classes on the basis of Year 6 information about achievement (one class) and demonstrated capacity for independent learning (a second class). Both classes once established remained largely unchanged until the end of Year 8. Summative assessment was low key and evidence of learning was collected from a wider range of activities.

In relation to prediction one, PCWAE2's achievement profile was more positively skewed to the top band achievers than MCWBE5's (Table K.1 in Appendix J). The bias was most obvious for the extended response component of the EV results. This was evidence of the effectiveness of the strong focus on improving students literacy skills in those early years of secondary schooling. There was no insight provided during the interview about how science teachers responded to the class of independent learners at the metropolitan school.

However, when looking at engagement (Table K.5D in Appendix J), students reported very different levels of support for their experience of science at the school. The lower achieving (overall) metropolitan school's top band students rated their experience (Items D and E on the student survey) 4th out of 16 (the number of case study schools plus the state figure counted as one school) and above the state figure compared to the provincial school's 14th which was below the state figure. Taking all three achievement bands into account, at the end of Year 8, MCWBE5 students ranked 3rd and PCWAE2 students ranked 12th which was the lowest of all the case study schools.

Engagement with science as measured by the proportions of students completing Year 12 science courses was stronger at the metropolitan school for the more demanding Chemistry and Physics courses (see Table K.4 in Appendix J). Both PCWAE2 and MCWBE5 (compared to the state) had more students completing Biology (both had 133%); in Chemistry, both schools had about the same proportions completing as in the state, but MCWBE5 had slightly more than PCWAE2 (100% versus 89%); in Physics the proportions relative to the state were slightly better for MCWBE5 (106% versus 81%). In the Senior Science course, more students at PCWAE2 completed the course than at MCWBE5 (288% versus 192%). However, when one looks at the large number of Senior Science course completions at the provincial school compared to the metropolitan school, either students at the provincial school had become more positive about science in the two years after Year 8 or they had no (or less attractive) options to choose from in Years 11 and 12. Science is optional after Year 10.

Pair FOUR: Assessment narratives compared for MGFSAE2 and MGFSWBE1

A. Engagement with EV feedback, resources and SOLO

MCFSWAE1

The HT from the coeducational WAE selective school (MCFSWAE1) participated in order to support research such as was represented by this project and to provide feedback about the EV program which was said to be a "high quality" program because its tasks and items "set high expectations" and "provide a basis for discriminating between responses from high ability students". Science teachers at this school use items and tasks from the EV tests in their own assessment programs but do not use SOLO-based rubrics to assess responses. The school is not planning to take up VALID10. Student survey results are not looked at nor discussed with staff or students. The HT thought that the test provided quality feedback to teachers and "liked" that it was mandatory.

The HT said that students enjoyed doing the test online and took it as seriously as they did NAPLAN. Some even used their own devices to do the test. No special preparation for the test is undertaken apart from registration and working through the sample items. There has been no feedback from parents about the test or results (when given to parents) and it receives no attention in annual school reports. The principal takes an interest in the results.

MGFSAE2

The HT at the girls AE selective school (MGFSAE2) participated to find out more about SOLO. The HT reported that the school's science assessment program involved a "SOLO based approach to assessment" by providing stimulus material with test items. SOLO-based rubrics are increasingly being used to mark responses to tasks and to inform feedback to students. It was reported that most of the staff at the school support SOLO as a basis for their own professional learning and for its usefulness in assessing student's work.

EV feedback is discussed at the time it is provided to the school. The HT does an analysis of achievement to identify strengths and weaknesses overall and between classes and this analysis is discussed with staff. The HT reported that EV results inform teachers' ongoing development of teaching programs and teacher assessment of student work.

The girls enjoy doing the test online and take it seriously. Some use their own devices to do the test. Staff will continue with VALID10 and are interested in the feedback on student growth from Y8 to Y10, particularly in relation to middle band students (only one or two students were assessed as low band).

The HT describes the SOLO rubric as about rewarding student responses that show appropriate "connections" between science concepts. Differences between SOLO marking and Board marking were described to me but were not seen as problematic. The HT had completed the proforma and acknowledged that she found the responses to the student survey confronting but useful. The concern was that student attitude responses were below state figures but no immediate thoughts about how to improve attitudes were offered. The HT nominated items F & E (from the student survey) as the most useful feedback from the perspective of science faculty priorities…that students learn their science and enjoy it.

MGFSWBE1

The head teacher from the girls WBE selective school (MGFSWBE1) participated to provide a professional learning session for science teachers about assessment. My project aligned with the focus at the school and in science on assessment for learning. An external consultant had been employed to improve their understanding of "differentiating assessment" and how to obtain and better use assessment data to improve teaching and learning. Science teachers use EV stimulus and related items and extended response tasks in their assessment program but rubrics that reflect syllabus intentions rather than SOLO thinking levels are used to assess student achievement. The staff view of SOLO was that the test items and tasks bring context and skills together so that responses can be assessed to reveal different levels of thinking using science content knowledge. Staff were critical of their current tests that focused very much on the acquisition of knowledge and understanding and they acknowledged that they did not sufficiently discriminate between levels of achievement.

Some science teachers reported that students did not take the EV test seriously because results are not counted in assessment. By contrast NAPLAN is taken seriously. Teachers reported that students were "stressed" because they were not sure the school's computers would work and that EV test questions were different to those in other science tests done at the school. The school plans to continue with the VALID10 program and see value in continuing their learning about SOLO.

B. Grouping for instruction

In the WAE school, students with the weakest literacy results are allocated to one class. In the AE school, students are put into classes on the basis of their choice of foreign language to be studied and in the WBE school, they try to spread students from feeder OC schools across the five classes formed. This is done to provide all students with the opportunity to broaden their friendship base. Thus Y7 classes in all three schools are effectively mixed ability classes from the perspective of science.

Essentially, all three schools retain the same classes from Years 7 to 10. An exception to this general approach is found in the WAE school where students with exceptional results are invited to join a gifted and talented class which is established from Y8. Acceptance into the class is conditional on the students agreeing to do chemistry in the senior years. The class is accelerated but no details were provided as to what that meant.

C. Use of learning intentions and success criteria

MCFSWAE1

The HT at the coeducational WAE school reported that the school placed a high priority on literacy. Extra assistance is given to the one class where students with weaker literacy skills were placed. The HT has expertise in literacy and an emphasis on literacy skills is evidenced in the artifacts provided to me. The HT reported a high science faculty priority for teaching scientific literacy skills valued in the world beyond school, for teaching critical thinking rather than rote learning and for greater student engagement with science at school and beyond.

Two topics, one each from Y7 and Y8, from school program provided to me demonstrated the priority for skill development. The program organized content into five columns, the first described content (science contexts and content to use and learn), the second skills (what students were to do with that content / the third contained references to pages in a science text book), a fourth included references to faculty and other resources relevant to the activities. A final column listed in syllabus outcomes shorthand (e.g. SC4-CW-2e / WS 6.3-6.4 AB 8) provides the link between school activities and syllabus intentions. The Y7 program topics in 2012 numbered 16. From 2015 this was reduced to 13, the year after the period of interest.

Learning / assessment tasks are accompanied by marking rubrics showing in great detail how marks are to be allocated. One Y7 literacy assessment provided to me targeted the writing of scientific explanations. The marking criteria for the five related tasks in that assignment allocate marks for completion of aspects of the task as well as for more sophisticated demonstrations of those aspects. Figure 5.4 shows part of the rubric for that assignment. The success criteria appear to be derived mostly from syllabus intentions but they also include literacy criteria as well.

Task 4 Re-writes two paragraphs in own words and uses the scaffold for structuring each explanation	
Identifies the phenomenon being addressed in the first paragraph /1	

Employs the explanation sequence, and, as appropriate: “action verbs, technical words, time connectives, cause-and-effect connectives” in order to explain the phenomenon /2	
Identifies the phenomenon being addressed in the second paragraph /1	
Employs the explanation sequence, and, as appropriate: “action verbs, technical words, time connectives, cause-and-effect connectives” in order to explain the phenomenon /2	
Total /6	

Task 5 Identifies the following language features of the text for one of the explanations and uses the correct symbol in so doing	
action verbs rises /1	
technical language or terms $\boxed{\text{evaporates}}$ /1	
time connectives when /1	
cause-and-effect connectives As a result /1	
Total /4	

Figure 5.4 Sample marking criteria for scientific explanations (part only)

MGFSAE2

The HT at the AE girls selective school said that the faculty priorities for junior secondary science were to prepare girls for a career in science, to ensure they were scientifically literate, able to creatively problem solve and to enjoy planning and conducting scientific activities. Learning programs for science in Years 7-10 at this school were organized into 4 X 10 week topics. Each topic was comprised of activities to be completed by students. The activities combined syllabus content and syllabus defined skills. The overall assessment plan showed that by the end of the year students overall grade would reflect the acquisition of both skills and knowledge and understandings.

The activities for a Y9 topic titled *The Complex Human* were organized into “booklets” and related scaffolds directed students to work in groups and to individually record specified outputs from those activities. This appeared to be a model for teaching and learning science that had been in place for some years. The scope of the activities I reviewed in the artifacts provided were consistent with syllabus expectations for knowledge and understanding and skills for Stage 5 students (topics in most non selective schools visited targeted Y8 or Stage 4 content), but some of the activities went beyond that.

Outputs to be provided included the construction of tables, graphs, procedures, risk assessments, descriptions, explanations, generalisations, conclusions and justifications. Assessment rubrics to be used by teachers to score the tasks described the features of outputs to be rewarded with marks. The features described for reward were both indicators of breadth of coverage and depth of understanding / level of skill demonstrated.

Also a SOLO based scaffold was provided. The scaffold was being trialed with school intranet science quizzes. It was based on SOLO level descriptors for comparison with student outputs to selected activities themselves based on content in the school's Y8 program.

The science faculty assessment policy document (provided with the artifacts) described the procedures to be followed when marks were transformed into grades for the purposes of reporting to parents. These appeared to be consistent with BOS *Common Grade Scale* requirements.

A sample Stage 4 activity titled Energy tranformations included a marking rubric that collated marks for two components of the syllabus working scientifically strand (planning and conducting an investigation and processing and analyzing data and information).

The Y9 student research project booklet provided included the steps to be followed in the development of a proposal for research, including opportunities for feedback from teachers, Turnitin software and student peers (see Figure 6.3) and assessors from the scientific community at a school based event. Students were encouraged to submit their project to the STANSW Young Scientists Competition as well.

MGFSWBE1

The HT for the WBE selective girls school explained that one of the school priorities for the year was assessment for learning and that an external consultant had been employed to provide professional learning to teachers. Science teachers had attended workshops provided by the consultant.

The science faculty priorities included working on their assessment tasks to improve their quality, moving the girls from rote learning and memorizing to thinking, improving their scientific literacy, building their understanding of the role of science in society and encouraging greater levels of enthusiasm for science.

The science learning program provides for four topics per year. No sample programs were provided. The assessment schemes for each of Years 7-10 were provided. There are four formal assessment tasks per year. Each task provides for a final equally weighted assessment of knowledge and understanding and working scientifically outcomes expressed as a grade. The syllabus outcomes targeted by the task are provided in full as part of the task notification. The rubric for assigning marks was included with the tasks and the links between marks and grades was also provided (from Year 8 onwards). The BOS *Common Grade Scale* appeared to be the basis for the award of grades.

Y7 tasks included a formal test, a task involving developing a game (a cross curriculum project…see next paragraph), a multi-media presentation and a "VALID Style test". In Y8 the tasks included a practical test, a mid-course test, a student research project and a "Yearly Exam VALID style".

The Y7 game task provides opportunities for students to demonstrate outcomes from the Art and PDH & PE and Science syllabuses. The multimedia task involves students in peer assessment of group work (see later section on feedback).

A Y8 "VALID style test" was provided in the set of artifacts. The short items in the test were similar in format to those used by the BOS in its external tests (both past and current ones). EV tests typically provide a stimulus text and a related set of 3-5 items about that text (Appendix 1.X includes an EV test booklet). Extended response tasks from previous EV tests were appropriated into their tests also, but the response scaffolds were modified to conform with BOS test formats. There was no evidence provided that SOLO concepts were used to mark responses.

Students also sit the ICAS science tests. Results are not used by the school for diagnostic purposes. Certificates are presented to students as an affirmation of their high ranking in the state for achievement of the five sets of scientific skills assessed by the test. The HT had declined an offer from the EAA team to show science teachers at the school how to use the tests to track school and individual progress using the results.

D. Classroom discourse and evidence of learning

MCFSWAE1

The assessment narrative for the WAE selective school reveals that students in their first two years of science are provided with learning activities based heavily of textbook, classroom worksheets and conventional school laboratory activities. The programs provided describe activities that combine both scientific skills and content, including opportunities for students to plan and design the laboratory activities. There is some evidence in tasks and tests provided that science rich contexts in line with syllabus expectations are provided as a stimulus for teaching and assessment activities (consistent with the EV assessment model of providing stimulus material and a group of related items the responses to which are dependent on comprehension of the text in the stimulus material).

Excursions are rare and science visitors to the school are on an infrequent “ad hoc” basis. Science teachers do not appear to make much use of resources beyond the classroom to enrich their teaching (such as the school grounds or local creeks and reserves) or engage their students in science investigations sponsored by external agencies such as BHP, Rio Tinto or the Young Scientist Competition (run by the Science Teachers Association of NSW). ICT use by science teachers is not a strong component in the teaching of science at the school according to the HT. National Science Week is not exploited for its celebration of science. ICAS tests of science thinking processes are mandatory but no attempt is made to use the results other than to affirm the high competence of the students.

The HT said that groupwork is not actively taught in the early years of science education at this school. The first major research project which the syllabus described as an opportunity for groupwork is an individual project (using plants) for Y8 students at this school. No artifacts relating to the Student Research Project (SRP) were provided.

Literacy based tasks, formal tests, written assignments, research projects and practical tasks appear to be the most valued sources of evidence for science learning. Practical tests are introduced no earlier than Y9 as a result of students in Y7 & 8 being stressed by the novelty and complexity of these assessments when they were introduced there a number of years ago.

Students at this school in Years 7 and 8 are frequently asked and given support to write scientific explanations. Also, they are challenged to use those skills in tasks well beyond their everyday experience. An example provided is a Y9 task (first introduced in 2013?) where students are asked, as a scientist, to prepare resources including a 3D model that could be used in a three minute TED presentation to evaluate strategies being used to reduce ozone depletion. Actually using the resources in a presentation was not required.

MGFSAE2

The artifacts provided by the HT at the AE girls' school reveal that teaching at the girls schools provides many structured opportunities for the girls to work cooperatively on a wide variety of tasks. The girls are encouraged to discuss the results of these activities, according to the HT, with both their teacher and peers before recording what they have learned. The structure of the set of activities provides a pathway that the girls can follow at their own pace rather than one set by the teacher who is freed up from presenting to the class to being able to work with small groups or one on one with students needing support. Thus evidence of learning is provided to teachers in the course of informal oral discussions and formally via the texts produced in response to prompts provided by the activity scaffold.

MGFSWBE1

The HT at the WBE girls school only provided examples of the assessment tasks used at the school. These artifacts combined with answers to questions provided by both the HT and teachers at the school revealed a willingness to work outside the science faculty with other faculties and to provide excursions to science rich environments including the zoo in Y7, a shoreline environment in Y9 and the Powerhouse Museum in Year 8.

There was some mention of a strong commitment to project based learning in previous years which is now confined to the SRP in Y8 and a cross curriculum project in Y7 (mentioned above) to produce a game that addresses outcomes from Art, PDH & PE as well as science.

Teacher willingness to work beyond the science classroom provides opportunities for devising authentic tasks through which to both teach science and to assess what was learned according to the HT science at the WBE girls school. Teacher talk at the interview about the tasks used and the various forms of evidence of that learning including student presentations, models (a game with science content) as well as more conventional tests, assignments and student research project reports describes the breadth of activities used to teach and assess evidence of learning at the school. In the junior secondary years, there does not appear to be a strong focus on improving student writing skills in the context of science beyond addressing the conventional sections of a traditional school scientific report which is the common assessment task for Term 3 in Y8. There appears to be little evidence of teaching to help students develop the expressive language skills using scientific vocabulary prior to that. The report was constructed by students working within a highly structured, teacher-provided scaffold.

Whilst the student research project involves groupwork, no persuasive evidence of formal teaching in the skills of groupwork was presented. Two rubrics for assessing the task and related report are provided. One to teachers and a second one to students which they use to self assess. Both rubrics appear to be modeled on the BOS *Common Grade Scale*. Only three grades are possible (A, B or C) which appears to be based on the historic evidence from Year 10 external science testing that ended in 2011 and perhaps (but not stated anywhere) the pattern of levels awarded in the EV results package. It was not clear how or whether students were coached in the use of their self-assessment rubric.

E. Feedback

MCFSWAE1

The HT at the coed school stated that science teachers at the school provide considerable informal feedback to students in the normal course of day to day teaching. They also provide students with formal feedback on performance in tests and tasks on a "look and listen" basis (my characterisation). Students are provided with the test/task and their individually teacher marked feedback sheets. A whole class presentation is made by the teacher to the class about the overall strengths and weaknesses in responses. Students are then expected to reflect on their individual feedback in their own time.

Teachers record marks awarded for tests / tasks and they are converted to grades for the purpose of reporting to parents twice a year. No insights about how the conversion was done was provided and no evidence was provided that the Board's *Common Grade Scale* (or SOLO levels for that matter) was the basis for that conversion either. No evidence was provided that students are given access to syllabus outcomes or the BOS *Common Grade Scale* in the early years of secondary school. Reporting up to the end of Year 9 was in terms of grades only. Place in the year is provided in Year 10 as well as grades.

MGFSAE2

The activity "booklets" used by the AE school provide more time for the teacher to work with students to provide feedback on individual issues as they arise. The growing use of SOLO provides another dimension to the type of feedback a teacher is able to provide as well.

MGFSWBE1

The HT science at the WBE girls school provided the rubrics used to convey feedback to students about their learning in the four formal assessment tasks. That feedback took the form of marks assigned according to what appeared to be syllabus based criteria. According to the HT science, consistency of marking was ensured by discussion of rubrics and sample responses at meetings of relevant teachers convened by the science coordinators for each Year group. Marks were subsequently converted to grades for the purpose of reporting using a version of the BOS *Common Grade Scale* model.

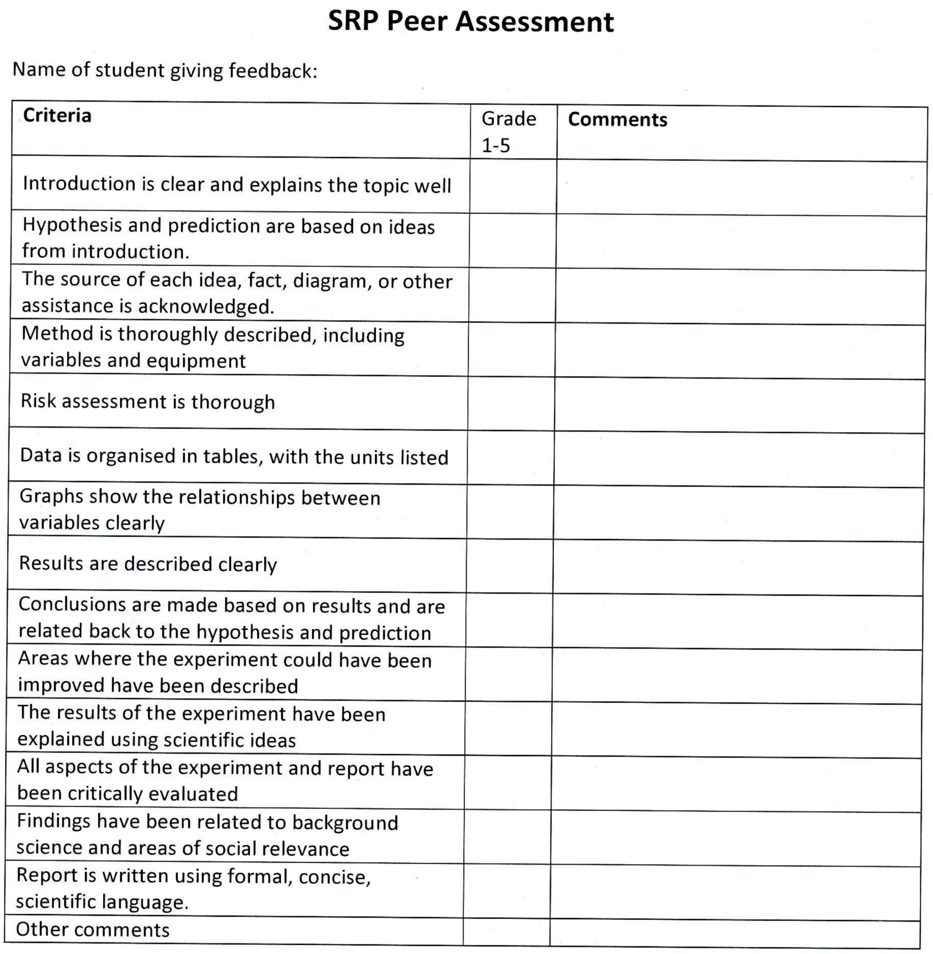
F. Activating students as instructional resources for others

MCFSWAE1

The HT science at the WAE school acknowledged that making the most of groupwork and the range of strategies associated with it (such as think-pair-share and report activities) was not a high priority amongst science teachers at the school. Nor was any evidence provided about opportunities students have to provide feedback to peers about their performance or achievement.

MGFSAE2

Discussion with peers in the context of groupwork is strongly supported by teachers at the AE school according to the HT science there. Informal discussion provides opportunities for joint construction with peers of responses, individually recorded, to the diversity of required outputs presented to students in the activity "booklets". There is a formal opportunity at the AE school in Y9 for students to provide feedback to peers about the quality of their reports and related explanations in terms of specific success criteria (Figure 5.5).



SRP Peer Assessment

Name of student giving feedback:

Criteria	Grade 1-5	Comments
Introduction is clear and explains the topic well		
Hypothesis and prediction are based on ideas from introduction.		
The source of each idea, fact, diagram, or other assistance is acknowledged.		
Method is thoroughly described, including variables and equipment		
Risk assessment is thorough		
Data is organised in tables, with the units listed		
Graphs show the relationships between variables clearly		
Results are described clearly		
Conclusions are made based on results and are related back to the hypothesis and prediction		
Areas where the experiment could have been improved have been described		
The results of the experiment have been explained using scientific ideas		
All aspects of the experiment and report have been critically evaluated		
Findings have been related to background science and areas of social relevance		
Report is written using formal, concise, scientific language.		
Other comments		

Figure 5.5 Peer assessment scaffold for a Y9 task at the AE school

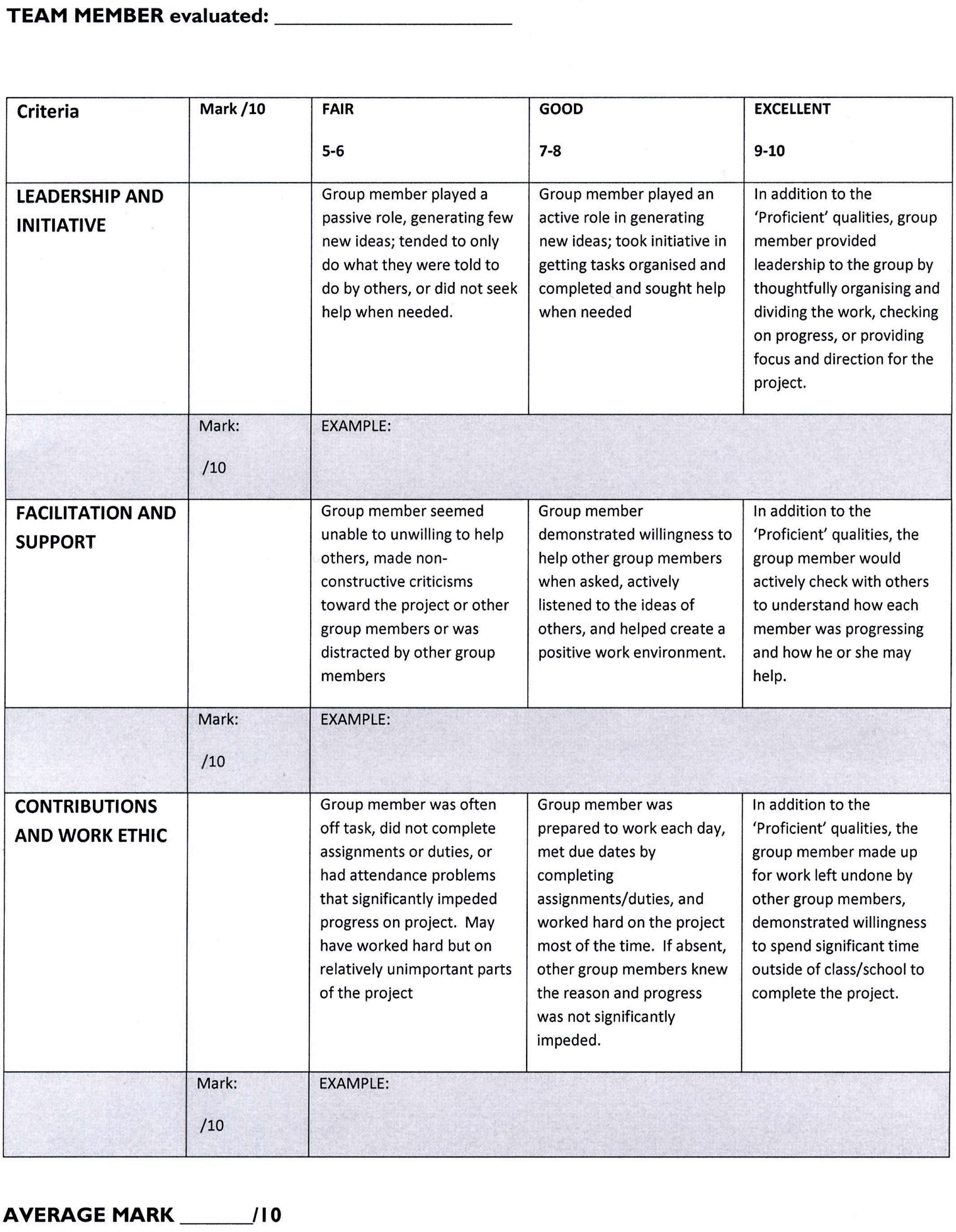
MGFSWBE1

At the WBE girls school, the one formal opportunity to provide feedback to peers using a skill based generic scaffold (Figure 5.6) was mediated by the class teacher who would only pass it on if he/she approved the contents (the person making the assessment was anonymous). It was not clear whether more than one teacher (who responded to that question very convincingly at the interview) used groupwork to provide opportunities for students to act as instructional resources for their peers.

The topic for the presentation was animal classification and an excursion to the zoo was involved. Information collected there was expected to be used back at school to prepare and deliver a multimedia report to the class.

It was not clear about the extent to which this was/is used and for how many years it may have been used.

The structure of the student presentation was organized using a teacher provided scaffold that doubled as a rubric (for the teacher to use) to assign marks for aspects of the group's preparation and presentation.



TEAM MEMBER evaluated: ______________________

Criteria	Mark /10	FAIR 5-6	GOOD 7-8	EXCELLENT 9-10
LEADERSHIP AND INITIATIVE		Group member played a passive role, generating few new ideas; tended to only do what they were told to do by others, or did not seek help when needed.	Group member played an active role in generating new ideas; took initiative in getting tasks organised and completed and sought help when needed	In addition to the 'Proficient' qualities, group member provided leadership to the group by thoughtfully organising and dividing the work, checking on progress, or providing focus and direction for the project.
	Mark: /10	EXAMPLE:		
FACILITATION AND SUPPORT		Group member seemed unable to unwilling to help others, made non-constructive criticisms toward the project or other group members or was distracted by other group members	Group member demonstrated willingness to help other group members when asked, actively listened to the ideas of others, and helped create a positive work environment.	In addition to the 'Proficient' qualities, the group member would actively check with others to understand how each member was progressing and how he or she may help.
	Mark: /10	EXAMPLE:		
CONTRIBUTIONS AND WORK ETHIC		Group member was often off task, did not complete assignments or duties, or had attendance problems that significantly impeded progress on project. May have worked hard but on relatively unimportant parts of the project	Group member was prepared to work each day, met due dates by completing assignments/duties, and worked hard on the project most of the time. If absent, other group members knew the reason and progress was not significantly impeded.	In addition to the 'Proficient' qualities, the group member made up for work left undone by other group members, demonstrated willingness to spend significant time outside of class/school to complete the project.
	Mark: /10	EXAMPLE:		

AVERAGE MARK _______/10

Figure 5.6 Peer assessment rubric for Y7 multi-media presentation task

G. Activating students (and teachers) as learners

MCFSWAE1

Many opportunities are provided to students in the coed school in the junior secondary years to develop good learning behaviours in the context of laboratory based activity worksheets that support the development of skills in comprehension, analysis, evaluation and justification of choices (using expressive, oral and written language).

In relation to teacher modeling of good learning behaviours, the HT reported that the science faculty met every two weeks and that the agenda often involved shared professional work such as development of program resources, assessment tasks, marking rubrics and joint marking of student work. There was no mention in the interviews about how teachers worked with their classes to help students achieve control over their learning apart from their work with language skills (explanations).

MGFSAE2

The use of a peer assessment scaffold by students at the AE girls school as described in the previous section provides teachers with a means for promoting good learning behaviours in all students. The skill of assessing your own work against criteria is an important step to self-regulated learning or learning how to learn. If you can recognise gaps or weakness in your own work, then you can devise strategies to address them. The time taken to explain how to do that self-assessment is an example of teachers modeling good learning behaviours.

The embrace by science teachers at the AE girls school of SOLO and their preparedness to work with it to improve their own professional competence and the learning outcomes for girls at their school was also evident.

MGFSWBE1

One concrete example of support for self-assessment was provided by the HT at the WBE school. In the context of their student research project (Task 3, Term 3 of Year 8) students are provided with the rubric teachers would use to assess the plan for their investigation and encouraged to use it for themselves prior to submitting their proposal. It was not clear to me whether this had been used earlier than last year once the new programs had been put in place.

According to the HT, science teachers at the WBE school meet regularly. Some of the meetings are devoted to collaborative work on programming but more recently on developing better assessment tasks to improve the quality of information about student learning including how to better discriminate between achievement. No evidence of support for and use by teachers of the SOLO model was mentioned in the interview apart from the appropriation of extended response tasks for use in their own tests modeled on Board formats.

H. Comparative summative comments

Comparisons between these three schools are fraught because of their differences in SEA scores and by the fact that one is a coeducational school, the other two girls schools. On the assumption that gender differences are not statistically significant, at least in the first few years of secondary school (PISA and TIMSS results for Australia support that conclusion), it is very obvious that at the end of Year 8, the comprehensive school is doing better in terms of achievement and engagement than either of the two girls schools. The focus on writing at the WAE and AE school shows up in the lower extended response score for the WBE school, despite them doing best in the communicating scientifically category of results (see Table K.1 in Appendix J).

In terms of engagement at the end of Year 8, the WAE fully selective entry school ranked above (8th) the other two fully selective entry schools in terms of student's enjoyment of their school science experience (see Table K.5D in Appendix J). Students at the AE school did not enjoy their science experience coming in at the bottom of the rankings 16th by top students on Items D and E combined. The WBE school did better at 13th.

Pair FIVE: Assessment narratives compared for PCWAE2 and PCWAE3

The narrative for PCWAE2 was presented above in the context of pair TWO. Thus only the information for PCWAE3 will be provided here.

A. Engagement with EV feedback, resources and SOLO

Only the head teacher attended the interview. Participation was on the basis of wanting to know how they were doing in terms of assessment practices which he did not think were any different to other schools. The school had not done VALID10 in 2015 and had no plans to do so going forward. The reason given for that was consistent with policy decision to keep formal assessments to a minimum and manage in a low key way because of its perceived negative impact on students motivation to learn. There was a large Indigenous population at the school (the largest of the three WAE provincial schools compared here) and the SEA score was very low. There was no special preparation for the EV test which he reported students enjoyed doing. No parent had asked about the report on results when it was sent home. The proforma had been completed for the interview and assessment-related artifacts were available as well.

B. Grouping for instruction

The school establishes 4-5 classes in Year 7 each year depending on numbers from feeder primary schools. Classes are streamed on the basis of feeder school achievement and other data. Whilst not graded from a science perspective, the top Year 7 class receives a more "challenging" program in science than is provided for the other classes. The bottom class receives additional support from learning support teachers who work with the science teachers in the class. These classes are largely retained going into Year 8 with some changes based on end of year test results and "behaviour" issues.

C. Use of learning intentions and success criteria

Learning programs are based on syllabus learning intentions (outcomes and related content) and traditional content organisers (Introduction to Laboratory / Forces / Solids, Liquids and Gases / Earth, Sun and Moon / Skills—Preparation for the SRP /Cells and Classification and Working with Nature are the topic headings for Year 7). The 2nd, 3rd, 5th and 6th topics each have 7 weeks allocated to them. The last topic includes a focus on "patterns in nature…respiration and photosynthesis…ecology…plant systems and structures and human (fire) and natural disasters…scientific and indigenous knowledge to extract resources from the environment." Many of the activities associated with the topics are literacy focused (correct use of appropriate vocabulary…adaptation not adaption) and separate pages list spelling and other literacy resources for each topic. Each topic has specific assessment tasks and there is a common assessment task each term (four in a year). There don't appear to be any formal exams or tests. The priority is for student engagement and enjoyment. Students are provided with a diversity of activities using a wide range of resources from within the school including Agriculture, which science manages. Students visit a local science fair each year. Relevance is important (eg diabetes in the context of work on disease). The students do a major research project each year which is done mostly in class time. Textbooks and worksheets are important components of classroom work. Students use school ICT, but it is not a large part of their work.

D. Classroom discourse and evidence of learning

Class discourse focuses explicitly on science language use including oral (first) and then written work. Research projects are scaffolded to help students learn the components of a scientific report; the scaffolding is progressively reduced from Year 7 to Year 10. Written responses to common assessment tasks is an important component of the assessment decisions and subsequent reporting to parents.

E. Feedback

This is largely provided by the teacher in the context of whole class discussion (oral) and to individuals and small groups during practical work in the lab. Students are provided with feedback sheets from common tasks and advice as to how they went in terms of grades based on the Board's common grade descriptors.

F. Activating students as instructional resources for others

Peer assessment was not a priority for Years 7 and 8. There was some use of think-pair-share-report strategy, but not widespread (according to HT). Groupwork was encouraged, but no evidence of teaching students the skills of working in groups was provided.

G. Activating students (and teachers) as learners

The focus was on teacher managed learning, students were not given opportunities to generate learning expectations or success criteria, but in feedback on tests, some teachers explained how feedback could be used to improve learning. The teachers at the school modelled good learning behaviours in class and with each other in meetings to discuss and develop assessment criteria which were then used individually to assess their own students. In commentary on the proforma, the HT saw the connection between liking science and better results.

H. Comparative summative comments

PCWAE2 and PCWAE3 had much in common. They had relatively large numbers (compared to PCWAE1) of indigenous students. The two schools set up graded classes, but both produced evidence of differentiated teaching in response to student skills. A major difference between the two schools was the approach taken to summative assessment. Like MCWBE5 (the school compared to PCWAE2 above), PC WAE3 had a low key approach to summative assessment.

At the end of Year 8, prediction one was satisfied in terms of both achievement and engagement (the extended response differential indicated that PCWAE2 was the more successful in terms of teaching writing skills).

Prediction two was about the extrapolation of results from Year 8 to Year 10. The evidence of results was not directly comparable, but the indication here was that, despite PCWAE3 having a higher proportion of its students absent on any day from Years 7 to 10 (see analysis in Chapter 5) than was so at PCWAE2, their Y10 result pattern was biased slightly more to the higher grades than PCWAE3's results were (see Table K.3 in Appendix J).

In terms of prediction three, PCWAE2 had proportionately more of its students completing science courses at the end of Year 12 when compared to PCWAE3 (relative to the state numbers). The most marked difference was in Biology where the difference was 133% versus 67%). Again, this finding needs to be qualified by unknowns about school resources and student demand for senior science courses.

Appendix I: Data tables for paired school comparisons

Table K.1

Achievement results for comparable school pairs (Year 8 EV reporting categories)

		EV %		ERT %		WSCI %		CSCI %	
School	AB	sch	sta	sch	sta	sch	sta	sch	sta
MCWAE1	5-6	7	18.6	12	20.3	9	19.4	8	22.4
x̄ = 1.85 ± 0.48	3-4	66	67.9	57	63.4	56	63.3	56	60.3
SEAS = 2.8 ± 0.46	1-2	27	13.5	32	16.3	35	17.3	36	17.3
MCAE2	5-6	16	18.6	18	20.3	17	19.4	25	22.4
x̄ = .03 ± 0.42	3-4	77	67.9	72	63.4	72	63.3	62	60.3
SEAS = 3.9 ± 0.30	1-2	7	13.5	10	16.3	11	17.3	13	17.3
MCWBE3	5-6	12	18.6	17	20.3	13	19.4	21	22.4
x̄ = -1.69 ± 0.13	3-4	76	67.9	68	63.4	70	63.3	61	60.3
SEAS = 4.0 ± 0.25	1-2	12	13.5	15	16.3	17	17.3	18	17.3
PCWAE2	5-6	12	18.6	18	20.3	16	19.4	14	22.4
x̄ = 1.69 ± 0.21	3-4	76	67.9	66	63.4	69	63.3	71	60.3
SEAS = 1.8 ± 0.45	1-2	12	13.5	16	16.3	15	17.3	15	17.3
MCWBE5	5-6	13	18.6	12	20.3	17	19.4	16	22.4
x̄ = -1.48 ± 0.28	3-4	69	67.9	66	63.4	61	63.3	66	60.3
SEAS = 2.1 ± 0.11	1-2	18	13.5	22	16.3	22	17.3	19	17.3
MCFSWAE1 x̄ = 1.19 ± 0.29 SEAS = 8.6 ± 0.16	5-6	95	18.6	85	20.3	80	19.4	87	22.4
	3-4	5	67.9	15	63.4	20	63.3	13	60.3
MGFSAE2 x̄ = -0.09 ± 0.44 SEAS = 8.3 ± 0.16	5-6	95	18.6	85	20.3	76	19.4	89	22.4
	3-4	5	67.9	15	63.4	24	63.3	11	60.3
MGFSWBE1 x̄ = -1.42 ± 0.02 SEAS = 8.9 ± 0.14	5-6	94	18.6	70	20.3	78	19.4	93	22.4
	3-4	6	67.9	30	63.4	22	63.3	7	60.3
PCWAE1	5-6	29	18.6	32	20.3	45	19.4	38	22.4
x̄ = 2.68 ± 0.38	3-4	69	67.9	62	63.4	51	63.3	58	60.3
SEAS = 2.7 ± 0.22	1-2	2	13.5	6	16.3	4	17.3	4	17.3
PCWAE2	SEE FOURTH DATA SET ABOVE								
PCWAE3	5-6	12	18.6	15	20.3	15	19.4	14	22.4
x̄ = 1.43 ± 0.25	3-4	75	67.9	66	63.4	68	63.3	66	60.3
SEAS = 2.0 ± 0.27	1-2	13	13.5	19	16.3	17	17.3	20	17.3

Note. SEAS = socio-educational advantage score / x̄ = mean school residual / AB = achievement band / EV % = proportions of students at each level of EV score (sch = school & sta = state) / ERT % = proportions for extended response tasks / WSCI = proportions for working scientifically / CSCI = proportions for communicating scientifically

Table K.2

Engagement measures at the end of Year 8

		Item A/4		Item B/4		Item C/4		Item D/4		Item E/%		Item F/%	
School	**AB**	sch	sta	sch	sta	sch	sta	sch	sta	sch	sta	sch	sta
MCWAE1	**5-6**	3.21	2.78	1.43	1.56	3.69	2.76	3.21	2.83	26.81	13.50	34.64	25.13
	3-4	2.49	1.76	2.33	1.69	2.81	2.35	2.63	2.23	12.53	6.65	22.88	16.50
	1-2	2.00	1.37	2.59	2.03	2.44	2.01	2.47	1.91	6.88	4.58	12.81	9.71
MCAE2	**5-6**	2.33	2.78	1.33	1.56	2.75	2.76	3.00	2.83	12.07	13.50	22.14	25.13
	3-4	1.61	1.76	1.80	1.69	2.27	2.35	2.26	2.23	5.12	6.65	14.27	16.50
	1-2	1.10	1.37	2.12	2.03	1.39	2.01	1.22	1.91	0.75	4.58	7.10	9.71
MCWBE3	**5-6**	2.87	2.78	1.53	1.56	2.82	2.76	2.57	2.83	11.99	13.50	19.57	25.13
	3-4	1.41	1.76	1.90	1.69	2.16	2.35	1.75	2.23	3.70	6.65	10.03	16.50
	1-2	1.09	1.37	1.90	2.03	1.87	2.01	1.30	1.91	0.63	4.58	4.47	9.71
PCWAE2	**5-6**	2.75	2.78	1.48	1.56	2.85	2.76	2.74	2.83	9.26	13.50	24.90	25.13
	3-4	1.69	1.76	2.06	1.69	1.84	2.35	2.08	2.23	2.98	6.65	16.17	16.50
	1-2	1.44	1.37	1.86	2.03	2.04	2.01	1.93	1.91	2.92	4.58	10.02	9.71
MCWBE5	**5-6**	2.90	2.78	1.58	1.56	3.12	2.76	3.00	2.83	19.60	13.50	25.35	25.13
	3-4	1.97	1.76	1.74	1.69	2.51	2.35	2.39	2.23	9.18	6.65	19.24	16.50
	1-2	1.32	1.37	2.02	2.03	2.26	2.01	2.03	1.91	6.06	4.58	10.17	9.71
MCFS WAE1*	**5-6**	3.22	2.78	1.95	1.56	2.92	2.76	2.81	2.83	13.71	13.50	28.51	25.13
	3-4	2.28	1.76	2.61	1.69	2.59	2.35	2.38	2.23	2.14	6.65	24.44	16.50
MGFS AE2*	**5-6**	2.73	2.78	1.65	1.56	2.53	2.76	2.32	2.83	6.58	13.50	20.95	25.13
	3-4	ns	1.76	ns	1.69	ns	2.35	ns	2.23	ns	6.65	ns	16.50
MGFS WBE1*	**5-6**	3.01	2.78	1.72	1.56	2.81	2.76	2.80	2.83	9.81	13.50	25.40	25.13
	3-4	2.66	1.76	2.02	1.69	2.65	2.35	2.68	2.23	6.44	6.65	18.94	16.50
PCWAE1	**5-6**	2.55	2.78	1.50	1.56	2.65	2.76	2.46	2.83	12.44	13.50	19.89	25.13
	3-4	1.35	1.76	1.44	1.69	1.55	2.35	1.99	2.23	3.64	6.65	12.70	16.50
	1-2	1.75	1.37	2.00	2.03	1.75	2.01	1.25	1.91	nil	4.58	8.33	9.71
PCWAE2	**SEE FOURTH DATA SET ABOVE**												
PCWAE3	**5-6**	2.60	2.78	1.35	1.56	1.25	2.76	2.19	2.83	8.61	13.50	20.19	25.13
	3-4	1.91	1.76	1.50	1.69	1.84	2.35	2.19	2.23	8.0	6.65	14.61	16.50
	1-2	1.09	1.37	2.10	2.03	1.4	2.01	1.67	1.91	nil	4.58	1.85	9.71

Note. Scores for Items A to D range from 0-4; Items E & F are the proportions (as a %) of students from that achievement band at that school (sch = school & sta = state).
AB = achievement band / Item A = intend to study science in senior years / Item B = science is the hardest subject I learn / Item C = enjoyed primary school science / Item D = enjoy secondary science lessons / Item E = number choosing science (as one of three favourite subjects) / Item F = number choosing science (as one of the three subjects they learn most in)
* These three schools had no students in the bottom achievement band / ns = no results supplied

Table K.3

Year 10 results

Grade (%)	MCWAE1 (%)				MCAE2[1] (%)				
	2011	2014	2015	MEAN	2012	2013	2014	2015	MEAN
A (13)	3	1	1	**2**	7	5	4	8	**6**
B (25)	10	6	9	**8**	15	22	18	31	**22**
C (36)	29	19	16	**21**	37	52	52	51	**47**
D (19)	32	38	22	**31**	37	18	26	7	**22**
E (7)	26	36	52	**38**	4	3	nil	3	**3**

[1] MCWBE3 did not provide any Year 10 results and MCAE2's results were used here instead.

	PCWAE2 (%)					MCWBE5 (%)				
	2012	2013	2014	2015	MEAN	2012	2013	2014	2015	MEAN
A (13)	7	5	2	6	**5**	6	11	9	11	**9**
B (25)	19	11	9	10	**12**	17	21	19	23	**20**
C (36)	47	56	46	33	**46**	53	46	44	46	**47**
D (19)	15	25	29	44	**28**	21	11	21	18	**18**
E (7)	12	3	14	7	**9**	3	11	7	2	**6**

	Grade	2012	2013	2014	2015	MEAN
MCFSWAE1	A (13)	59	64	65	63	63
	B (25)	36	31	30	33	33
	C (36)	5	5	6	4	5
	D (19)	nil	nil	nil	nil	0
	E (7)	nil	nil	nil	nil	0
MGFSAE2	A (13)	89	86	82	83	85
	B (25)	8	13	18	15	14
	C (36)	3	1	nil	2	2
	D (19)	nil	nil	nil	nil	0
	E (7)	nil	nil	nil	nil	0
MGFSWBE1	A (13)	80	65	66	85	74
	B (25)	17	35	33	15	25
	C (36)	3	nil	1	nil	1
	D (19)	nil	nil	nil	nil	0
	E (7)	nil	nil	nil	nil	0
PCWAE1	A (13)	13	15	18	11	14
	B (25)	19	21	5	29	19
	C (36)	45	39	46	46	44
	D (19)	23	24	27	14	22
	E (7)	nil	nil	4	nil	1
PCWAE2	A (13)	7	5	2	6	5
	B (25)	19	11	9	10	12
	C (36)	47	56	46	33	46
	D (19)	15	25	29	44	28
	E (7)	12	3	14	7	9

	LEVEL* \ YR**	2009	2010	2011	
PCWAE3	6 (9)	4	3	5	4
	5 (25)	22	21	16	20
	4 (35)	38	39	40	39
	3 (23)	27	30	34	30
	2 (5)	9	7	5	7
	1 (<1)	0	0	0	0

GRADE (Four year average proportions of state population achieving grades A to E as a percentage)

Table K.4

Science course completions at the end of Year 12

Subject (state % in 2015)	MCWAE1 %				MCWBE3 %			
Year	2013	2014	2015	MEAN	2013	2014	2015	MEAN
Biology (28.5)	25	22	48	32	15	31	18	21
Chemistry (18)	17	6	14	12	3	8	9	7
Earth & Env. Sc. (2.4)	n/a	n/a	n/a	N/A	n/a	n/a	n/a	N/A
Physics (16)	15	12	14	14	12	6	8	9
Senior Science (10.4)	27	26	23	25	n/a	n/a	n/a	N/A

Subject (state % in 2015)	PCWAE1 %				MCAE2 %			
Year	2013	2014	2015	MEAN	2013	2014	2015	MEAN
Biology (28.5)	41	50	30	40	n/a	55	58	57
Chemistry (18)	24	n/a	20	22	n/a	21	17	19
Earth & Env. Sc. (2.4)	n/a	n/a	n/a	N/A	n/a	n/a	n/a	N/A
Physics (16)	n/a	10	35	22	n/a	12	8	10
Senior Science (10.4)	59	40	n/a	50	n/a	n/a	n/a	N/A

Subject (state % in 2015)	PCWAE2 %				MCWBE5 %			
Year	2013	2014	2015	MEAN	2013	2014	2015	MEAN
Biology (28.5)	31	46	38	38	35	37	42	38
Chemistry (18)	15	14	18	16	13	21	20	18
Earth & Env. Sc. (2.4)	n/a	n/a	n/a	N/A	n/a	n/a	n/a	N/A
Physics (16)	8	n/a	18	13	26	12	14	17
Senior Science (10.4)	46	n/a	13	30	22	12	26	20

Subject (state % in 2015)	MGFSAE2 %				MGFSWBE1 %			
Year	2013	2014	2015	MEAN	2013	2014	2015	MEAN
Biology (28.5)	21	22	17	20	18.6	23.6	24.8	22
Chemistry (18)	55	53	54	54	58.4	61.8	54.8	58
Earth & Env. Sc. (2.4)	n/a	n/a	n/a	N/A	n/a	n/a	n/a	N/A
Physics (16)	18	21	30	23	30	29.3	23.6	28
Senior Science (10.4)	n/a	n/a	n/a	N/A	n/a	n/a	n/a	N/A

Subject (state % in 2015)	MCFSWAE1				PCWAE3 %			
Year	2013	2014	2015	MEAN	2013	2014	2015	MEAN
Biology (28.5)	31.9	40	30.2	34	21	19	17	19
Chemistry (18)	65.9	74.3	71.2	70	17	15	19	17
Earth & Env. Sc. (2.4)	n/a	n/a	n/a	N/A	n/a	10	n/a	10
Physics (16)	46.4	45.7	46	46	21	3	9	11
Senior Science (10.4)	5.8	10.7	11.5	9	8	32	26	22

Note. The proportions (%) reported are relative to the total English candidature for the state in 2015 and at the school for each year. n/a = subject not offered that year

Table K.5A

Student survey item scores, ranks and relative to the state

School*	A	A	A	A	A	B**	B**	B**	B**	B**
	TOP	TRNK	TMB	RTMB	STA	TOP	TRNK	TMB	RTMB	STA
PCWAE1	-0.23	12	-0.26	7	B	-0.06	8	-0.34	10	B
MCWAE1	0.43	2	1.79	1	A	-0.13	12	1.07	1	A
PCWAE2	-0.03	9	-0.03	4	B	-0.08	9	0.12	6	A
MCWAE2	-0.01	8	-1.49	13	B	-0.27	15	-2.25	13	B
PCWAE3	-0.18	11	-0.31	8	B	-0.21	13	-0.33	11	B
MCFSWAE1	0.44	1	#NULL!	NA	NA	0.39	1	#NULL!	NA	NA
MCAE2	-0.45	14	-0.87	10	B	-0.23	14	-0.03	9	B
MCAE3	-0.33	13	-0.24	6	B	-0.12	10	0.31	4	A
MGFSAE2	-0.05	10	#NULL!	NA	NA	0.09	3	#NULL!	NA	NA
MCAE6	-0.49	15	-0.89	11	B	-0.71	16	-0.77	12	B
MGFSWBE1	0.23	4	#NULL!	NA	NA	0.16	2	#NULL!	NA	NA
MCWBE5	0.12	5	0.28	2	A	0.02	4	0.06	5	A
MCWBE4	-1.45	16	-1.15	12	B	-0.12	11	0.37	2	A
MCWBE3	0.09	6	-0.54	9	B	-0.03	7	0.05	7	A
MCPSWBE2	0.28	3	-0.20	5	B	0.00	5=	0.20	3	A
STATE	2.78	7	5.91	3	=	1.56	5=	5.28	8	=

Schools* in residual order / A & B = Item number of survey / TOP = top band scores relative to state (see bottom row) / TRNK = top band rank (n = 16) / TMB = sum of scores for all three achievement bands relative to the state (see bottom row) / STA = above (A) or below (B) the state score. Item A = I want to study a science subject in years 11 &12. Item B** = Science is the hardest subject I learn (disagreement with that was taken as a good thing)

Table K.5B

Student survey item scores, ranks and relative to the state

School*	C	C	C	C	C	D	D	D	D	D
	TOP	TRNK	TMB	RTMB	STA	TOP	TRNK	TMB	RTMB	STA
PCWAE1	-0.11	12	-1.28	11	B	-0.37	14	-1.27	11	B
MCWAE1	0.93	1	2.75	1	A	0.38	2	1.34	1	A
PCWAE2	0.09	7	-0.30	8	B	-0.09	11	-0.22	7	B
MCWAE2	-0.12	13	-2.27	12	B	0.03	6	-2.00	13	B
PCWAE3	-1.51	16	-4.15	13	B	-0.64	16	-0.92	10	B
MCFSWAE1	0.16	6	#NULL!	NA	NA	-0.02	9	#NULL!	NA	NA
MCAE2	-0.01	11	-0.72	10	B	0.17	4	-0.49	9	B
MCAE3	-0.15	14	-0.36	9	B	-0.16	12	0.14	4	A
MGFSAE2	-0.23	15	#NULL!	NA	NA	-0.51	15	#NULL!	NA	NA
MCAE6	0.47	2	1.17	2	A	0.15	5	0.06	5	A
MGFSWBE1	0.05	9	#NULL!	NA	NA	-0.03	10	#NULL!	NA	NA
MCWBE5	0.36	3	1.13	3	A	0.17	3	0.45	3	A
MCWBE4	0.35	4	0.97	4	A	0.50	1	1.13	2	A
MCWBE3	0.06	8	-0.21	7	B	-0.26	13	-1.35	12	B
MCPSWBE2	0.18	5	-0.07	6	B	0.01	7	-0.45	8	B
STATE	2.76	10	7.12	5	=	2.83	8	6.97	6	=

Schools* in residual order / C & D = Item number of survey / TOP = top band scores relative to state (see bottom row) / TRNK = top band rank (n = 16) / TMB = sum of scores for all three achievement bands relative to the state (see bottom row) / STA = above (A) or below (B) the state score. Item C = In primary school, I enjoyed lessons that were about science / Item D = In secondary school, I enjoy science lessons

Table K.5C

Student survey item scores, ranks and relative to the state

School*	E	E	E	E	E	F	F	F	F	F
	TOP	TRNK	TMB	RTMB	STA	TOP	TRNK	TMB	RTMB	STA
PCWAE1	-1.06	9	-8.65	12	B	-5.24	15	-15.66	11	B
MCWAE1	13.31	2	21.49	1	A	9.51	1	28.50	1	A
PCWAE2	-4.24	14	-9.57	13	B	-0.23	9	-0.48	7	B
MCWAE2	1.90	4	-5.01	8	B	2.14	5	-4.35	9	B
PCWAE3	-4.89	15	-8.12	10	B	-4.94	14	-19.63	12	B
MCFSWAE1	0.21	7	#NULL!	NA	NA	3.38	3	#NULL!	NA	NA
MCAE2	-1.43	10	-6.79	9	B	-2.99	11	-10.82	10	B
MCAE3	0.75	6	2.40	4	A	-4.13	12	-0.22	6	B
MGFSAE2	-6.92	16	#NULL!	NA	NA	-4.18	13	#NULL!	NA	NA
MCAE6	-2.39	12	-4.40	7	B	-0.84	10	-3.41	8	B
MGFSWBE1	-3.69	13	#NULL!	NA	NA	0.27	6	#NULL!	NA	NA
MCWBE5	6.10	3	10.11	3	A	0.22	7	3.64	3	A
MCWBE4	16.13	1	17.18	2	A	8.20	2	14.66	2	A
MCWBE3	-1.51	11	-8.41	11	B	-5.56	16	-22.83	13	B
MCPSWBE2	1.01	5	-0.74	6	B	2.46	4	3.25	4	A
STATE	13.50	8	24.73	5	=	25.13	8	51.34	5	=

Schools* in residual order / E & F = Item number of survey / TOP = top band scores relative to state (see bottom row) / TRNK = top band rank (n = 16) / TMB = sum of scores for all three achievement bands relative to the state (see bottom row) / STA = above (A) or below (B) the state score. Item E = my three favourite subjects (15 to choose from) / Item F = the three subjects I thought I learned most in (15 to choose from)

Table K.5D

Student survey items (D + E) scores, ranks and relative to the state

School*	D+E TOP	D+E TRNK	D+E TMB	D+E RTMB	D+E STA	NSALL RANK /12	TALL RANK /16
PCWAE1	-1.43	10	-9.92	13	B	10	13
MCWAE1	13.69	3	22.83	1	A	1	1
PCWAE2	-4.33	14	-9.79	12	B	7	11
MCWAE2	1.93	5	-7.01	8	B	9	6
PCWAE3	-5.53	15	-9.04	10	B	11	16
MCFSWAE1	0.19	8	#NULL!	NA	NA	N/A	4
MCAE2	-1.26	9	-7.28	9	B	8	12
MCAE3	0.59	7	2.54	4	A	4	10
MGFSAE2	-7.43	16	#NULL!	NA	NA	N/A	15
MCAE6	-2.24	12	-4.34	7	B	6	9
MGFSWBE1	-3.72	13	#NULL!	NA	NA	N/A	8
MCWBE5	6.27	4	10.56	3	A	3	3
MCWBE4	16.63	1	18.31	2	A	2	2
MCWBE3	-1.77	11	-9.76	11	B	12	14
MCPSWBE2	1.02	6	-1.19	6	B	N/A	5
STATE	16.33	2	31.70	5	=	5	7

Schools* in residual order / D + E = Item numbers of survey / TOP = top band scores relative to state (see bottom row) / TRNK = top band rank (n = 16) / TMB = sum of scores for all three achievement bands relative to the state (see bottom row) / STA = above (A) or below (B) the state score. Items D + E = the sum of the scores for Items D and E from the student survey (see above for what they are) NSALL RANK = sum of all achievement band survey scores (rank out of 12) / TALL RANK = sum of top achievement band survey scores (rank out of 16).

Appendix J: Survey descriptive statistics

The reference to residual quintile group is about the three groups (WAE, AE and WBE and the groups separating WAE from AE and AE from WBE…five groups in all differentiated by their residuals).

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Residual quintile group * EV1A	85	100.0%	0	0.0%	85	100.0%
Residual quintile group * EV1B	85	100.0%	0	0.0%	85	100.0%
Residual quintile group * EV1C	84	98.8%	1	1.2%	85	100.0%
Residual quintile group * EV1D	84	98.8%	1	1.2%	85	100.0%
Residual quintile group * EV1E	85	100.0%	0	0.0%	85	100.0%
Residual quintile group * EV1F	85	100.0%	0	0.0%	85	100.0%
Residual quintile group * EV1G	83	97.6%	2	2.4%	85	100.0%
Residual quintile group * EV1H	84	98.8%	1	1.2%	85	100.0%
Residual quintile group * EV1I	84	98.8%	1	1.2%	85	100.0%
Residual quintile group * EV2A	85	100.0%	0	0.0%	85	100.0%
Residual quintile group * EV2B	85	100.0%	0	0.0%	85	100.0%
Residual quintile group * EV2C	84	98.8%	1	1.2%	85	100.0%
Residual quintile group * EV2D	85	100.0%	0	0.0%	85	100.0%
Residual quintile group * EV2E	85	100.0%	0	0.0%	85	100.0%
Residual quintile group * EV2F	84	98.8%	1	1.2%	85	100.0%
Residual quintile group * EV2G	84	98.8%	1	1.2%	85	100.0%
Residual quintile group * EV2H	85	100.0%	0	0.0%	85	100.0%
Residual quintile group * EV2I	84	98.8%	1	1.2%	85	100.0%
Residual quintile group * EV2J	85	100.0%	0	0.0%	85	100.0%
Residual quintile group * EV2K	85	100.0%	0	0.0%	85	100.0%
Residual quintile group * EV2L	83	97.6%	2	2.4%	85	100.0%
Residual quintile group * EV2M	84	98.8%	1	1.2%	85	100.0%
Residual quintile group * EV3	85	100.0%	0	0.0%	85	100.0%

Residual quintile group * EV5	84	98.8%	1	1.2%	85	100.0%
Residual quintile group * S6A	84	98.8%	1	1.2%	85	100.0%
Residual quintile group * S6B	83	97.6%	2	2.4%	85	100.0%
Residual quintile group * S6C	84	98.8%	1	1.2%	85	100.0%
Residual quintile group * S6D	84	98.8%	1	1.2%	85	100.0%
Residual quintile group * S6E	84	98.8%	1	1.2%	85	100.0%
Residual quintile group * S6F	83	97.6%	2	2.4%	85	100.0%
Residual quintile group * S6G	84	98.8%	1	1.2%	85	100.0%
Residual quintile group * S6H	83	97.6%	2	2.4%	85	100.0%
Residual quintile group * S6I	83	97.6%	2	2.4%	85	100.0%
Residual quintile group * S6J	84	98.8%	1	1.2%	85	100.0%
Residual quintile group * S7	84	98.8%	1	1.2%	85	100.0%

CODE:1.00 = YES and 2.00 = NO

Residual quintile group * EV1A Crosstabulation

			EV1A		
			1.00	2.00	Total
Residual quintile group	WBE	Count	17	15	32
		% within Residual quintile group	53.1%	46.9%	100.0%
	AE	Count	20	8	28
		% within Residual quintile group	71.4%	28.6%	100.0%
	WAE	Count	20	5	25
		% within Residual quintile group	80.0%	20.0%	100.0%
Total		Count	57	28	85
		% within Residual quintile group	67.1%	32.9%	100.0%

Residual quintile group * EV1B Crosstabulation

			EV1B		
			1.00	2.00	Total
Residual quintile group	WBE	Count	17	15	32
		% within Residual quintile group	53.1%	46.9%	100.0%
	AE	Count	22	6	28
		% within Residual quintile group	78.6%	21.4%	100.0%
	WAE	Count	18	7	25
		% within Residual quintile group	72.0%	28.0%	100.0%
Total		Count	57	28	85
		% within Residual quintile group	67.1%	32.9%	100.0%

Residual quintile group * EV1C Crosstabulation

			EV1C		
			1.00	2.00	Total
Residual quintile group	WBE	Count	13	18	31
		% within Residual quintile group	41.9%	58.1%	100.0%
	AE	Count	21	7	28
		% within Residual quintile group	75.0%	25.0%	100.0%
	WAE	Count	17	8	25
		% within Residual quintile group	68.0%	32.0%	100.0%
Total		Count	51	33	84
		% within Residual quintile group	60.7%	39.3%	100.0%

Residual quintile group * EV1D Crosstabulation

			EV1D		
			1.00	2.00	Total
Residual quintile group	WBE	Count	10	22	32
		% within Residual quintile group	31.3%	68.8%	100.0%
	AE	Count	13	14	27
		% within Residual quintile group	48.1%	51.9%	100.0%
	WAE	Count	17	8	25
		% within Residual quintile group	68.0%	32.0%	100.0%
Total		Count	40	44	84
		% within Residual quintile group	47.6%	52.4%	100.0%

Residual quintile group * EV1E Crosstabulation

			EV1E		
			1.00	2.00	Total
Residual quintile group	WBE	Count	17	15	32
		% within Residual quintile group	53.1%	46.9%	100.0%
	AE	Count	23	5	28
		% within Residual quintile group	82.1%	17.9%	100.0%
	WAE	Count	16	9	25
		% within Residual quintile group	64.0%	36.0%	100.0%
Total		Count	56	29	85
		% within Residual quintile group	65.9%	34.1%	100.0%

Residual quintile group * EV1F Crosstabulation

			EV1F		
			1.00	2.00	Total
Residual quintile group	WBE	Count	5	27	32
		% within Residual quintile group	15.6%	84.4%	100.0%
	AE	Count	6	22	28
		% within Residual quintile group	21.4%	78.6%	100.0%
	WAE	Count	8	17	25
		% within Residual quintile group	32.0%	68.0%	100.0%
Total		Count	19	66	85
		% within Residual quintile group	22.4%	77.6%	100.0%

Residual quintile group * EV1G Crosstabulation

			EV1G		
			1.00	2.00	Total
Residual quintile group	WBE	Count	12	20	32
		% within Residual quintile group	37.5%	62.5%	100.0%
	AE	Count	15	11	26
		% within Residual quintile group	57.7%	42.3%	100.0%
	WAE	Count	14	11	25
		% within Residual quintile group	56.0%	44.0%	100.0%
Total		Count	41	42	83
		% within Residual quintile group	49.4%	50.6%	100.0%

Residual quintile group * EV1H Crosstabulation

			EV1H		
			1.00	2.00	Total
Residual quintile group	WBE	Count	8	24	32
		% within Residual quintile group	25.0%	75.0%	100.0%
	AE	Count	11	16	27
		% within Residual quintile group	40.7%	59.3%	100.0%
	WAE	Count	9	16	25
		% within Residual quintile group	36.0%	64.0%	100.0%
Total		Count	28	56	84
		% within Residual quintile group	33.3%	66.7%	100.0%

Residual quintile group * EV1I Crosstabulation

			EV1I		
			1.00	2.00	Total
Residual quintile group	WBE	Count	3	28	31
		% within Residual quintile group	9.7%	90.3%	100.0%
	AE	Count	6	22	28
		% within Residual quintile group	21.4%	78.6%	100.0%
	WAE	Count	6	19	25
		% within Residual quintile group	24.0%	76.0%	100.0%
Total		Count	15	69	84
		% within Residual quintile group	17.9%	82.1%	100.0%

Residual quintile group * EV2A Crosstabulation

			EV2A		
			1.00	2.00	Total
Residual quintile group	WBE	Count	7	25	32
		% within Residual quintile group	21.9%	78.1%	100.0%
	AE	Count	9	19	28
		% within Residual quintile group	32.1%	67.9%	100.0%
	WAE	Count	11	14	25
		% within Residual quintile group	44.0%	56.0%	100.0%
Total		Count	27	58	85
		% within Residual quintile group	31.8%	68.2%	100.0%

Residual quintile group * EV2B Crosstabulation

			EV2B		
			1.00	2.00	Total
Residual quintile group	WBE	Count	15	17	32
		% within Residual quintile group	46.9%	53.1%	100.0%
	AE	Count	24	4	28
		% within Residual quintile group	85.7%	14.3%	100.0%
	WAE	Count	21	4	25
		% within Residual quintile group	84.0%	16.0%	100.0%
Total		Count	60	25	85
		% within Residual quintile group	70.6%	29.4%	100.0%

Residual quintile group * EV2C Crosstabulation

			EV2C		
			1.00	2.00	Total
Residual quintile group	WBE	Count	11	21	32
		% within Residual quintile group	34.4%	65.6%	100.0%
	AE	Count	18	9	27
		% within Residual quintile group	66.7%	33.3%	100.0%
	WAE	Count	9	16	25
		% within Residual quintile group	36.0%	64.0%	100.0%
Total		Count	38	46	84
		% within Residual quintile group	45.2%	54.8%	100.0%

Residual quintile group * EV2D Crosstabulation

			EV2D		
			1.00	2.00	Total
Residual quintile group	WBE	Count	16	16	32
		% within Residual quintile group	50.0%	50.0%	100.0%
	AE	Count	15	13	28
		% within Residual quintile group	53.6%	46.4%	100.0%
	WAE	Count	16	9	25
		% within Residual quintile group	64.0%	36.0%	100.0%
Total		Count	47	38	85
		% within Residual quintile group	55.3%	44.7%	100.0%

Residual quintile group * EV2E Crosstabulation

			EV2E		
			1.00	2.00	Total
Residual quintile group	WBE	Count	20	12	32
		% within Residual quintile group	62.5%	37.5%	100.0%
	AE	Count	22	6	28
		% within Residual quintile group	78.6%	21.4%	100.0%
	WAE	Count	19	6	25
		% within Residual quintile group	76.0%	24.0%	100.0%
Total		Count	61	24	85
		% within Residual quintile group	71.8%	28.2%	100.0%

Residual quintile group * EV2F Crosstabulation

			EV2F		
			1.00	2.00	Total
Residual quintile group	WBE	Count	16	15	31
		% within Residual quintile group	51.6%	48.4%	100.0%
	AE	Count	22	6	28
		% within Residual quintile group	78.6%	21.4%	100.0%
	WAE	Count	18	7	25
		% within Residual quintile group	72.0%	28.0%	100.0%
Total		Count	56	28	84
		% within Residual quintile group	66.7%	33.3%	100.0%

Residual quintile group * EV2G Crosstabulation

			EV2G		
			1.00	2.00	Total
Residual quintile group	WBE	Count	19	13	32
		% within Residual quintile group	59.4%	40.6%	100.0%
	AE	Count	20	7	27
		% within Residual quintile group	74.1%	25.9%	100.0%
	WAE	Count	18	7	25
		% within Residual quintile group	72.0%	28.0%	100.0%
Total		Count	57	27	84
		% within Residual quintile group	67.9%	32.1%	100.0%

Residual quintile group * EV2H Crosstabulation

			EV2H		
			1.00	2.00	Total
Residual quintile group	WBE	Count	14	18	32
		% within Residual quintile group	43.8%	56.3%	100.0%
	AE	Count	21	7	28
		% within Residual quintile group	75.0%	25.0%	100.0%
	WAE	Count	14	11	25
		% within Residual quintile group	56.0%	44.0%	100.0%
Total		Count	49	36	85
		% within Residual quintile group	57.6%	42.4%	100.0%

Residual quintile group * EV2I Crosstabulation

			EV2I		
			1.00	2.00	Total
Residual quintile group	WBE	Count	7	24	31
		% within Residual quintile group	22.6%	77.4%	100.0%
	AE	Count	21	7	28
		% within Residual quintile group	75.0%	25.0%	100.0%
	WAE	Count	12	13	25
		% within Residual quintile group	48.0%	52.0%	100.0%
Total		Count	40	44	84
		% within Residual quintile group	47.6%	52.4%	100.0%

Residual quintile group * EV2J Crosstabulation

			EV2J		
			1.00	2.00	Total
Residual quintile group	WBE	Count	0	32	32
		% within Residual quintile group	0.0%	100.0%	100.0%
	AE	Count	2	26	28
		% within Residual quintile group	7.1%	92.9%	100.0%
	WAE	Count	0	25	25
		% within Residual quintile group	0.0%	100.0%	100.0%
Total		Count	2	83	85
		% within Residual quintile group	2.4%	97.6%	100.0%

Residual quintile group * EV2K Crosstabulation

			EV2K		
			1.00	2.00	Total
Residual quintile group	WBE	Count	0	32	32
		% within Residual quintile group	0.0%	100.0%	100.0%
	AE	Count	2	26	28
		% within Residual quintile group	7.1%	92.9%	100.0%
	WAE	Count	2	23	25
		% within Residual quintile group	8.0%	92.0%	100.0%
Total		Count	4	81	85
		% within Residual quintile group	4.7%	95.3%	100.0%

Residual quintile group * EV2L Crosstabulation

			EV2L		
			1.00	2.00	Total
Residual quintile group	WBE	Count	10	22	32
		% within Residual quintile group	31.3%	68.8%	100.0%
	AE	Count	9	18	27
		% within Residual quintile group	33.3%	66.7%	100.0%
	WAE	Count	13	11	24
		% within Residual quintile group	54.2%	45.8%	100.0%
Total		Count	32	51	83
		% within Residual quintile group	38.6%	61.4%	100.0%

Residual quintile group * EV2M Crosstabulation

			EV2M		
			1.00	2.00	Total
Residual quintile group	WBE	Count	7	25	32
		% within Residual quintile group	21.9%	78.1%	100.0%
	AE	Count	9	18	27
		% within Residual quintile group	33.3%	66.7%	100.0%
	WAE	Count	9	16	25
		% within Residual quintile group	36.0%	64.0%	100.0%
Total		Count	25	59	84
		% within Residual quintile group	29.8%	70.2%	100.0%

Residual quintile group * EV3 Crosstabulation (see questionnaire for the key)

			EV3					
			1.00	2.00	3.00	4.00	5.00	Total
Residual quintile group	WBE	Count	5	4	12	9	2	32
		% within Residual quintile group	15.6%	12.5%	37.5%	28.1%	6.3%	100.0%
	AE	Count	0	0	8	11	9	28
		% within Residual quintile group	0.0%	0.0%	28.6%	39.3%	32.1%	100.0%
	WAE	Count	0	2	6	11	6	25
		% within Residual quintile group	0.0%	8.0%	24.0%	44.0%	24.0%	100.0%
Total		Count	5	6	26	31	17	85
		% within Residual quintile group	5.9%	7.1%	30.6%	36.5%	20.0%	100.0%

Residual quintile group * EV5 Crosstabulation (see questionnaire for the key)

			EV5			
			1.00	2.00	3.00	Total
Residual quintile group	WBE	Count	15	6	11	32
		% within Residual quintile group	46.9%	18.8%	34.4%	100.0%
	AE	Count	14	8	5	27
		% within Residual quintile group	51.9%	29.6%	18.5%	100.0%
	WAE	Count	18	1	6	25
		% within Residual quintile group	72.0%	4.0%	24.0%	100.0%
Total		Count	47	15	22	84
		% within Residual quintile group	56.0%	17.9%	26.2%	100.0%

Residual quintile group * S6A Crosstabulation

			S6A		
			1.00	2.00	Total
Residual quintile group	WBE	Count	17	15	32
		% within Residual quintile group	53.1%	46.9%	100.0%
	AE	Count	15	13	28
		% within Residual quintile group	53.6%	46.4%	100.0%
	WAE	Count	13	11	24
		% within Residual quintile group	54.2%	45.8%	100.0%
Total		Count	45	39	84
		% within Residual quintile group	53.6%	46.4%	100.0%

Residual quintile group * S6B Crosstabulation

			S6B		
			1.00	2.00	Total
Residual quintile group	WBE	Count	6	26	32
		% within Residual quintile group	18.8%	81.3%	100.0%
	AE	Count	7	20	27
		% within Residual quintile group	25.9%	74.1%	100.0%
	WAE	Count	9	15	24
		% within Residual quintile group	37.5%	62.5%	100.0%
Total		Count	22	61	83
		% within Residual quintile group	26.5%	73.5%	100.0%

Residual quintile group * S6C Crosstabulation

			S6C		
			1.00	2.00	Total
Residual quintile group	WBE	Count	13	19	32
		% within Residual quintile group	40.6%	59.4%	100.0%
	AE	Count	10	18	28
		% within Residual quintile group	35.7%	64.3%	100.0%
	WAE	Count	12	12	24
		% within Residual quintile group	50.0%	50.0%	100.0%
Total		Count	35	49	84
		% within Residual quintile group	41.7%	58.3%	100.0%

Residual quintile group * S6D Crosstabulation

			S6D		
			1.00	2.00	Total
Residual quintile group	WBE	Count	6	26	32
		% within Residual quintile group	18.8%	81.3%	100.0%
	AE	Count	4	24	28
		% within Residual quintile group	14.3%	85.7%	100.0%
	WAE	Count	8	16	24
		% within Residual quintile group	33.3%	66.7%	100.0%
Total		Count	18	66	84
		% within Residual quintile group	21.4%	78.6%	100.0%

Residual quintile group * S6E Crosstabulation

			S6E		
			1.00	2.00	Total
Residual quintile group	WBE	Count	7	25	32
		% within Residual quintile group	21.9%	78.1%	100.0%
	AE	Count	8	20	28
		% within Residual quintile group	28.6%	71.4%	100.0%
	WAE	Count	9	15	24
		% within Residual quintile group	37.5%	62.5%	100.0%
Total		Count	24	60	84
		% within Residual quintile group	28.6%	71.4%	100.0%

Residual quintile group * S6F Crosstabulation

			S6F		
			1.00	2.00	Total
Residual quintile group	WBE	Count	4	27	31
		% within Residual quintile group	12.9%	87.1%	100.0%
	AE	Count	5	23	28
		% within Residual quintile group	17.9%	82.1%	100.0%
	WAE	Count	6	18	24
		% within Residual quintile group	25.0%	75.0%	100.0%
Total		Count	15	68	83
		% within Residual quintile group	18.1%	81.9%	100.0%

Residual quintile group * S6G Crosstabulation

			S6G		
			1.00	2.00	Total
Residual quintile group	WBE	Count	6	26	32
		% within Residual quintile group	18.8%	81.3%	100.0%
	AE	Count	7	21	28
		% within Residual quintile group	25.0%	75.0%	100.0%
	WAE	Count	8	16	24
		% within Residual quintile group	33.3%	66.7%	100.0%
Total		Count	21	63	84
		% within Residual quintile group	25.0%	75.0%	100.0%

Residual quintile group * S6H Crosstabulation

			S6H		
			1.00	2.00	Total
Residual quintile group	WBE	Count	3	28	31
		% within Residual quintile group	9.7%	90.3%	100.0%
	AE	Count	3	25	28
		% within Residual quintile group	10.7%	89.3%	100.0%
	WAE	Count	5	19	24
		% within Residual quintile group	20.8%	79.2%	100.0%
Total		Count	11	72	83
		% within Residual quintile group	13.3%	86.7%	100.0%

Residual quintile group * S6I Crosstabulation

			S6I		
			1.00	2.00	Total
Residual quintile group	WBE	Count	2	29	31
		% within Residual quintile group	6.5%	93.5%	100.0%
	AE	Count	1	27	28
		% within Residual quintile group	3.6%	96.4%	100.0%
	WAE	Count	1	23	24
		% within Residual quintile group	4.2%	95.8%	100.0%
Total		Count	4	79	83
		% within Residual quintile group	4.8%	95.2%	100.0%

Residual quintile group * S6J Crosstabulation

			S6J		
			1.00	2.00	Total
Residual quintile group	WBE	Count	0	32	32
		% within Residual quintile group	0.0%	100.0%	100.0%
	AE	Count	2	26	28
		% within Residual quintile group	7.1%	92.9%	100.0%
	WAE	Count	3	21	24
		% within Residual quintile group	12.5%	87.5%	100.0%
Total		Count	5	79	84
		% within Residual quintile group	6.0%	94.0%	100.0%

Residual quintile group * S7 Crosstabulation (see questionnaire for the key)

			S7					
			1.00	2.00	3.00	4.00	5.00	Total
Residual quintile group	WBE	Count	9	5	12	6	0	32
		% within Residual quintile group	28.1%	15.6%	37.5%	18.8%	0.0%	100.0%
	AE	Count	6	6	9	7	0	28
		% within Residual quintile group	21.4%	21.4%	32.1%	25.0%	0.0%	100.0%
	WAE	Count	8	5	3	5	3	24
		% within Residual quintile group	33.3%	20.8%	12.5%	20.8%	12.5%	100.0%
Total		Count	23	16	24	18	3	84
		% within Residual quintile group	27.4%	19.0%	28.6%	21.4%	3.6%	100.0%

SECTION THREE: ASSESSMENT FOR LEARNING (AFL) DESCRIPTIVE STATISTICS

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Residual quintile group * AFL9A	83	97.6%	2	2.4%	85	100.0%
Residual quintile group * AFL9B	83	97.6%	2	2.4%	85	100.0%
Residual quintile group * AFL9C	83	97.6%	2	2.4%	85	100.0%
Residual quintile group * AFL9D	83	97.6%	2	2.4%	85	100.0%
Residual quintile group * AFL9E	82	96.5%	3	3.5%	85	100.0%
Residual quintile group * AFL9F	82	96.5%	3	3.5%	85	100.0%
Residual quintile group * AFL9G	81	95.3%	4	4.7%	85	100.0%
Residual quintile group * AFL9H	83	97.6%	2	2.4%	85	100.0%
Residual quintile group * AFL10A	82	96.5%	3	3.5%	85	100.0%
Residual quintile group * AFL10B	83	97.6%	2	2.4%	85	100.0%
Residual quintile group * AFL10C	83	97.6%	2	2.4%	85	100.0%
Residual quintile group * AFL10D	83	97.6%	2	2.4%	85	100.0%
Residual quintile group * AFL10E	82	96.5%	3	3.5%	85	100.0%
Residual quintile group * AFL10F	83	97.6%	2	2.4%	85	100.0%
Residual quintile group * AFL10G	82	96.5%	3	3.5%	85	100.0%
Residual quintile group * AFL10H	82	96.5%	3	3.5%	85	100.0%
Residual quintile group * AFL11A	83	97.6%	2	2.4%	85	100.0%
Residual quintile group * AFL11B	83	97.6%	2	2.4%	85	100.0%
Residual quintile group * AFL11C	82	96.5%	3	3.5%	85	100.0%
Residual quintile group * AFL11D	82	96.5%	3	3.5%	85	100.0%
Residual quintile group * AFL11E	83	97.6%	2	2.4%	85	100.0%
Residual quintile group * Exemplary or model answers	85	100.0%	0	0.0%	85	100.0%
Residual quintile group * Success criteria	85	100.0%	0	0.0%	85	100.0%
Residual quintile group * Misconceptions	85	100.0%	0	0.0%	85	100.0%
Residual quintile group * SOLO levels	85	100.0%	0	0.0%	85	100.0%
Residual quintile group * QT model	85	100.0%	0	0.0%	85	100.0%
Residual quintile group * Bloom categories	85	100.0%	0	0.0%	85	100.0%

Residual quintile group * Syllabus outcomes / standards	85	100.0%	0	0.0%	85	100.0%
Residual quintile group * AFL13A	81	95.3%	4	4.7%	85	100.0%
Residual quintile group * AFL13B	81	95.3%	4	4.7%	85	100.0%
Residual quintile group * AFL13C	81	95.3%	4	4.7%	85	100.0%
Residual quintile group * AFL13D	82	96.5%	3	3.5%	85	100.0%
Residual quintile group * AFL13E	80	94.1%	5	5.9%	85	100.0%
Residual quintile group * AFL13F	81	95.3%	4	4.7%	85	100.0%
Residual quintile group * AFL14A	81	95.3%	4	4.7%	85	100.0%
Residual quintile group * AFL14B	81	95.3%	4	4.7%	85	100.0%
Residual quintile group * AFL14C	83	97.6%	2	2.4%	85	100.0%
Residual quintile group * AFL14D	83	97.6%	2	2.4%	85	100.0%
Residual quintile group * AFL14E	81	95.3%	4	4.7%	85	100.0%
Residual quintile group * AFL14F	83	97.6%	2	2.4%	85	100.0%
Residual quintile group * AFL14G	83	97.6%	2	2.4%	85	100.0%
Residual quintile group * AFL14H	83	97.6%	2	2.4%	85	100.0%
Residual quintile group * AFL15A	83	97.6%	2	2.4%	85	100.0%
Residual quintile group * AFL15B	82	96.5%	3	3.5%	85	100.0%
Residual quintile group * AFL15C	82	96.5%	3	3.5%	85	100.0%
Residual quintile group * AFL15D	79	92.9%	6	7.1%	85	100.0%
Residual quintile group * AFL15E	83	97.6%	2	2.4%	85	100.0%

N = 47 items

SEE QUESIONNAIRE FOR KEY EXPLAINING RESPONSE AND RELATED NUMBER

Residual quintile group * AFL9A Crosstabulation

			AFL9A				
			1.00	3.00	4.00	5.00	Total
Residual quintile group	WBE	Count	2	1	8	20	31
		% within Residual quintile group	6.5%	3.2%	25.8%	64.5%	100.0%
	AE	Count	0	0	7	21	28
		% within Residual quintile group	0.0%	0.0%	25.0%	75.0%	100.0%
	WAE	Count	0	0	11	13	24
		% within Residual quintile group	0.0%	0.0%	45.8%	54.2%	100.0%
Total		Count	2	1	26	54	83
		% within Residual quintile group	2.4%	1.2%	31.3%	65.1%	100.0%

Residual quintile group * AFL9B Crosstabulation

			AFL9B					
			1.00	2.00	3.00	4.00	5.00	Total
Residual quintile group	WBE	Count	2	4	8	17	0	31
		% within Residual quintile group	6.5%	12.9%	25.8%	54.8%	0.0%	100.0%
	AE	Count	1	2	8	15	2	28
		% within Residual quintile group	3.6%	7.1%	28.6%	53.6%	7.1%	100.0%
	WAE	Count	0	4	5	14	1	24
		% within Residual quintile group	0.0%	16.7%	20.8%	58.3%	4.2%	100.0%
Total		Count	3	10	21	46	3	83
		% within Residual quintile group	3.6%	12.0%	25.3%	55.4%	3.6%	100.0%

Residual quintile group * AFL9C Crosstabulation

			AFL9C					
			1.00	2.00	3.00	4.00	5.00	Total
Residual quintile group	WBE	Count	1	1	3	10	16	31
		% within Residual quintile group	3.2%	3.2%	9.7%	32.3%	51.6%	100.0%
	AE	Count	0	0	3	4	21	28
		% within Residual quintile group	0.0%	0.0%	10.7%	14.3%	75.0%	100.0%
	WAE	Count	0	0	1	8	15	24
		% within Residual quintile group	0.0%	0.0%	4.2%	33.3%	62.5%	100.0%
Total		Count	1	1	7	22	52	83
		% within Residual quintile group	1.2%	1.2%	8.4%	26.5%	62.7%	100.0%

Residual quintile group * AFL9D Crosstabulation

			AFL9D					
			1.00	2.00	3.00	4.00	5.00	Total
Residual quintile group	WBE	Count	1	5	14	9	2	31
		% within Residual quintile group	3.2%	16.1%	45.2%	29.0%	6.5%	100.0%

	AE	Count	1	1	10	15	1	28
		% within Residual quintile group	3.6%	3.6%	35.7%	53.6%	3.6%	100.0%
	WAE	Count	0	3	7	12	2	24
		% within Residual quintile group	0.0%	12.5%	29.2%	50.0%	8.3%	100.0%
Total		Count	2	9	31	36	5	83
		% within Residual quintile group	2.4%	10.8%	37.3%	43.4%	6.0%	100.0%

Residual quintile group * AFL9E Crosstabulation

			AFL9E			
			3.00	4.00	5.00	Total
Residual quintile group	WBE	Count	1	11	19	31
		% within Residual quintile group	3.2%	35.5%	61.3%	100.0%
	AE	Count	0	6	21	27
		% within Residual quintile group	0.0%	22.2%	77.8%	100.0%
	WAE	Count	1	10	13	24
		% within Residual quintile group	4.2%	41.7%	54.2%	100.0%
Total		Count	2	27	53	82
		% within Residual quintile group	2.4%	32.9%	64.6%	100.0%

Residual quintile group * AFL9F Crosstabulation

			AFL9F				
			2.00	3.00	4.00	5.00	Total
Residual quintile group	WBE	Count	1	4	19	6	30
		% within Residual quintile group	3.3%	13.3%	63.3%	20.0%	100.0%
	AE	Count	1	4	16	7	28
		% within Residual quintile group	3.6%	14.3%	57.1%	25.0%	100.0%
	WAE	Count	1	5	10	8	24
		% within Residual quintile group	4.2%	20.8%	41.7%	33.3%	100.0%
Total		Count	3	13	45	21	82
		% within Residual quintile group	3.7%	15.9%	54.9%	25.6%	100.0%

Residual quintile group * AFL9G Crosstabulation

			AFL9G					
			1.00	2.00	3.00	4.00	5.00	Total
Residual quintile group	WBE	Count	1	1	5	18	5	30
		% within Residual quintile group	3.3%	3.3%	16.7%	60.0%	16.7%	100.0%
	AE	Count	0	1	4	18	4	27
		% within Residual quintile group	0.0%	3.7%	14.8%	66.7%	14.8%	100.0%
	WAE	Count	0	1	5	11	7	24
		% within Residual quintile group	0.0%	4.2%	20.8%	45.8%	29.2%	100.0%
Total		Count	1	3	14	47	16	81

		% within Residual quintile group	1.2%	3.7%	17.3%	58.0%	19.8%	100.0%

Residual quintile group * AFL9H Crosstabulation

			AFL9H					
			1.00	2.00	3.00	4.00	5.00	Total
Residual quintile group	WBE	Count	4	12	10	5	0	31
		% within Residual quintile group	12.9%	38.7%	32.3%	16.1%	0.0%	100.0%
	AE	Count	2	12	5	8	1	28
		% within Residual quintile group	7.1%	42.9%	17.9%	28.6%	3.6%	100.0%
	WAE	Count	3	13	6	0	2	24
		% within Residual quintile group	12.5%	54.2%	25.0%	0.0%	8.3%	100.0%
Total		Count	9	37	21	13	3	83
		% within Residual quintile group	10.8%	44.6%	25.3%	15.7%	3.6%	100.0%

Residual quintile group * AFL10A Crosstabulation

			AFL10A				
			2.00	3.00	4.00	5.00	Total
Residual quintile group	WBE	Count	0	5	18	8	31
		% within Residual quintile group	0.0%	16.1%	58.1%	25.8%	100.0%
	AE	Count	1	4	16	6	27
		% within Residual quintile group	3.7%	14.8%	59.3%	22.2%	100.0%
	WAE	Count	0	5	14	5	24
		% within Residual quintile group	0.0%	20.8%	58.3%	20.8%	100.0%
Total		Count	1	14	48	19	82
		% within Residual quintile group	1.2%	17.1%	58.5%	23.2%	100.0%

Residual quintile group * AFL10B Crosstabulation

			AFL10B		
			4.00	5.00	Total
Residual quintile group	WBE	Count	13	18	31
		% within Residual quintile group	41.9%	58.1%	100.0%
	AE	Count	12	16	28
		% within Residual quintile group	42.9%	57.1%	100.0%
	WAE	Count	11	13	24
		% within Residual quintile group	45.8%	54.2%	100.0%
Total		Count	36	47	83
		% within Residual quintile group	43.4%	56.6%	100.0%

Residual quintile group * AFL10C Crosstabulation

			AFL10C				
			1.00	3.00	4.00	5.00	Total
	WBE	Count	1	1	14	15	31

Residual quintile group		% within Residual quintile group	3.2%	3.2%	45.2%	48.4%	100.0%
	AE	Count	0	0	12	16	28
		% within Residual quintile group	0.0%	0.0%	42.9%	57.1%	100.0%
	WAE	Count	0	0	5	19	24
		% within Residual quintile group	0.0%	0.0%	20.8%	79.2%	100.0%
Total		Count	1	1	31	50	83
		% within Residual quintile group	1.2%	1.2%	37.3%	60.2%	100.0%

Residual quintile group * AFL10D Crosstabulation

			AFL10D				
			2.00	3.00	4.00	5.00	Total
Residual quintile group	WBE	Count	1	2	11	17	31
		% within Residual quintile group	3.2%	6.5%	35.5%	54.8%	100.0%
	AE	Count	0	2	8	18	28
		% within Residual quintile group	0.0%	7.1%	28.6%	64.3%	100.0%
	WAE	Count	0	0	6	18	24
		% within Residual quintile group	0.0%	0.0%	25.0%	75.0%	100.0%
Total		Count	1	4	25	53	83
		% within Residual quintile group	1.2%	4.8%	30.1%	63.9%	100.0%

Residual quintile group * AFL10E Crosstabulation

			AFL10E					
			1.00	2.00	3.00	4.00	5.00	Total
Residual quintile group	WBE	Count	2	2	5	19	3	31
		% within Residual quintile group	6.5%	6.5%	16.1%	61.3%	9.7%	100.0%
	AE	Count	1	2	3	17	5	28
		% within Residual quintile group	3.6%	7.1%	10.7%	60.7%	17.9%	100.0%
	WAE	Count	2	2	4	10	5	23
		% within Residual quintile group	8.7%	8.7%	17.4%	43.5%	21.7%	100.0%
Total		Count	5	6	12	46	13	82
		% within Residual quintile group	6.1%	7.3%	14.6%	56.1%	15.9%	100.0%

Residual quintile group * AFL10F Crosstabulation

			AFL10F				
			2.00	3.00	4.00	5.00	Total
Residual quintile group	WBE	Count	0	8	19	4	31
		% within Residual quintile group	0.0%	25.8%	61.3%	12.9%	100.0%
	AE	Count	0	2	17	9	28
		% within Residual quintile group	0.0%	7.1%	60.7%	32.1%	100.0%
	WAE	Count	1	1	14	8	24
		% within Residual quintile group	4.2%	4.2%	58.3%	33.3%	100.0%

Total		Count	1	11	50	21	83
		% within Residual quintile group	1.2%	13.3%	60.2%	25.3%	100.0%

Residual quintile group * AFL10G Crosstabulation

			AFL10G				
			2.00	3.00	4.00	5.00	Total
Residual quintile group	WBE	Count	2	13	12	3	30
		% within Residual quintile group	6.7%	43.3%	40.0%	10.0%	100.0%
	AE	Count	3	8	13	4	28
		% within Residual quintile group	10.7%	28.6%	46.4%	14.3%	100.0%
	WAE	Count	2	2	16	4	24
		% within Residual quintile group	8.3%	8.3%	66.7%	16.7%	100.0%
Total		Count	7	23	41	11	82
		% within Residual quintile group	8.5%	28.0%	50.0%	13.4%	100.0%

Residual quintile group * AFL10H Crosstabulation

			AFL10H					
			1.00	2.00	3.00	4.00	5.00	Total
Residual quintile group	WBE	Count	1	1	1	16	12	31
		% within Residual quintile group	3.2%	3.2%	3.2%	51.6%	38.7%	100.0%
	AE	Count	0	0	0	14	13	27
		% within Residual quintile group	0.0%	0.0%	0.0%	51.9%	48.1%	100.0%
	WAE	Count	0	0	0	6	18	24
		% within Residual quintile group	0.0%	0.0%	0.0%	25.0%	75.0%	100.0%
Total		Count	1	1	1	36	43	82
		% within Residual quintile group	1.2%	1.2%	1.2%	43.9%	52.4%	100.0%

Residual quintile group * AFL11A Crosstabulation

			AFL11A			
			3.00	4.00	5.00	Total
Residual quintile group	WBE	Count	4	8	19	31
		% within Residual quintile group	12.9%	25.8%	61.3%	100.0%
	AE	Count	1	7	20	28
		% within Residual quintile group	3.6%	25.0%	71.4%	100.0%
	WAE	Count	1	5	18	24
		% within Residual quintile group	4.2%	20.8%	75.0%	100.0%
Total		Count	6	20	57	83
		% within Residual quintile group	7.2%	24.1%	68.7%	100.0%

Residual quintile group * AFL11B Crosstabulation

			AFL11B			
			3.00	4.00	5.00	Total
Residual quintile group	WBE	Count	3	7	21	31

		% within Residual quintile group	9.7%	22.6%	67.7%	100.0%
	AE	Count	1	11	16	28
		% within Residual quintile group	3.6%	39.3%	57.1%	100.0%
	WAE	Count	0	6	18	24
		% within Residual quintile group	0.0%	25.0%	75.0%	100.0%
Total		Count	4	24	55	83
		% within Residual quintile group	4.8%	28.9%	66.3%	100.0%

Residual quintile group * AFL11C Crosstabulation

			AFL11C				
			2.00	3.00	4.00	5.00	Total
Residual quintile group	WBE	Count	4	8	13	6	31
		% within Residual quintile group	12.9%	25.8%	41.9%	19.4%	100.0%
	AE	Count	3	4	13	8	28
		% within Residual quintile group	10.7%	14.3%	46.4%	28.6%	100.0%
	WAE	Count	2	4	8	9	23
		% within Residual quintile group	8.7%	17.4%	34.8%	39.1%	100.0%
Total		Count	9	16	34	23	82
		% within Residual quintile group	11.0%	19.5%	41.5%	28.0%	100.0%

Residual quintile group * AFL11D Crosstabulation

			AFL11D		
			4.00	5.00	Total
Residual quintile group	WBE	Count	10	20	30
		% within Residual quintile group	33.3%	66.7%	100.0%
	AE	Count	8	20	28
		% within Residual quintile group	28.6%	71.4%	100.0%
	WAE	Count	5	19	24
		% within Residual quintile group	20.8%	79.2%	100.0%
Total		Count	23	59	82
		% within Residual quintile group	28.0%	72.0%	100.0%

Residual quintile group * AFL11E Crosstabulation

			AFL11E		
			4.00	5.00	Total
Residual quintile group	WBE	Count	12	19	31
		% within Residual quintile group	38.7%	61.3%	100.0%
	AE	Count	6	22	28
		% within Residual quintile group	21.4%	78.6%	100.0%
	WAE	Count	2	22	24
		% within Residual quintile group	8.3%	91.7%	100.0%

Total	Count	20	63	83
	% within Residual quintile group	24.1%	75.9%	100.0%

Residual quintile group * Exemplary or model answers Crosstabulation

			Exemplary or model answers					
			.00	2.00	3.00	4.00	5.00	Total
Residual quintile group	WBE	Count	2	1	3	19	7	32
		% within Residual quintile group	6.3%	3.1%	9.4%	59.4%	21.9%	100.0%
	AE	Count	3	0	2	15	8	28
		% within Residual quintile group	10.7%	0.0%	7.1%	53.6%	28.6%	100.0%
	WAE	Count	1	0	1	15	8	25
		% within Residual quintile group	4.0%	0.0%	4.0%	60.0%	32.0%	100.0%
Total		Count	6	1	6	49	23	85
		% within Residual quintile group	7.1%	1.2%	7.1%	57.6%	27.1%	100.0%

Residual quintile group * Success criteria Crosstabulation

			Success criteria				
			.00	3.00	4.00	5.00	Total
Residual quintile group	WBE	Count	2	5	14	11	32
		% within Residual quintile group	6.3%	15.6%	43.8%	34.4%	100.0%
	AE	Count	3	1	6	18	28
		% within Residual quintile group	10.7%	3.6%	21.4%	64.3%	100.0%
	WAE	Count	1	1	9	14	25
		% within Residual quintile group	4.0%	4.0%	36.0%	56.0%	100.0%
Total		Count	6	7	29	43	85
		% within Residual quintile group	7.1%	8.2%	34.1%	50.6%	100.0%

Residual quintile group * Misconceptions Crosstabulation

			Misconceptions				
			.00	3.00	4.00	5.00	Total
Residual quintile group	WBE	Count	3	3	14	12	32
		% within Residual quintile group	9.4%	9.4%	43.8%	37.5%	100.0%
	AE	Count	3	2	10	13	28
		% within Residual quintile group	10.7%	7.1%	35.7%	46.4%	100.0%
	WAE	Count	1	3	8	13	25
		% within Residual quintile group	4.0%	12.0%	32.0%	52.0%	100.0%
Total		Count	7	8	32	38	85
		% within Residual quintile group	8.2%	9.4%	37.6%	44.7%	100.0%

Residual quintile group * SOLO levels Crosstabulation

			SOLO levels					
			.00	2.00	3.00	4.00	5.00	Total
	WBE	Count	9	10	7	4	2	32

Residual quintile group		% within Residual quintile group	28.1%	31.3%	21.9%	12.5%	6.3%	100.0%
	AE	Count	9	4	8	7	0	28
		% within Residual quintile group	32.1%	14.3%	28.6%	25.0%	0.0%	100.0%
	WAE	Count	8	7	3	3	4	25
		% within Residual quintile group	32.0%	28.0%	12.0%	12.0%	16.0%	100.0%
Total		Count	26	21	18	14	6	85
		% within Residual quintile group	30.6%	24.7%	21.2%	16.5%	7.1%	100.0%

Residual quintile group * QT model Crosstabulation

			QT model					
			.00	2.00	3.00	4.00	5.00	Total
Residual quintile group	WBE	Count	3	7	9	10	3	32
		% within Residual quintile group	9.4%	21.9%	28.1%	31.3%	9.4%	100.0%
	AE	Count	5	2	5	8	8	28
		% within Residual quintile group	17.9%	7.1%	17.9%	28.6%	28.6%	100.0%
	WAE	Count	3	3	5	7	7	25
		% within Residual quintile group	12.0%	12.0%	20.0%	28.0%	28.0%	100.0%
Total		Count	11	12	19	25	18	85
		% within Residual quintile group	12.9%	14.1%	22.4%	29.4%	21.2%	100.0%

Residual quintile group * Bloom categories Crosstabulation

			Bloom categories					
			.00	2.00	3.00	4.00	5.00	Total
Residual quintile group	WBE	Count	3	3	12	10	4	32
		% within Residual quintile group	9.4%	9.4%	37.5%	31.3%	12.5%	100.0%
	AE	Count	3	1	3	13	8	28
		% within Residual quintile group	10.7%	3.6%	10.7%	46.4%	28.6%	100.0%
	WAE	Count	2	2	4	12	5	25
		% within Residual quintile group	8.0%	8.0%	16.0%	48.0%	20.0%	100.0%
Total		Count	8	6	19	35	17	85
		% within Residual quintile group	9.4%	7.1%	22.4%	41.2%	20.0%	100.0%

Residual quintile group * Syllabus outcomes / standards Crosstabulation

			Syllabus outcomes / standards					
			.00	2.00	3.00	4.00	5.00	Total
Residual quintile group	WBE	Count	2	3	3	15	9	32
		% within Residual quintile group	6.3%	9.4%	9.4%	46.9%	28.1%	100.0%
	AE	Count	4	2	1	8	13	28
		% within Residual quintile group	14.3%	7.1%	3.6%	28.6%	46.4%	100.0%
	WAE	Count	1	1	1	7	15	25
		% within Residual quintile group	4.0%	4.0%	4.0%	28.0%	60.0%	100.0%

Total		Count	7	6	5	30	37	85
		% within Residual quintile group	8.2%	7.1%	5.9%	35.3%	43.5%	100.0%

Residual quintile group * AFL13A Crosstabulation

			AFL13A					
			1.00	2.00	3.00	4.00	5.00	Total
Residual quintile group	WBE	Count	0	12	12	6	0	30
		% within Residual quintile group	0.0%	40.0%	40.0%	20.0%	0.0%	100.0%
	AE	Count	1	6	7	12	2	28
		% within Residual quintile group	3.6%	21.4%	25.0%	42.9%	7.1%	100.0%
	WAE	Count	0	5	9	7	2	23
		% within Residual quintile group	0.0%	21.7%	39.1%	30.4%	8.7%	100.0%
Total		Count	1	23	28	25	4	81
		% within Residual quintile group	1.2%	28.4%	34.6%	30.9%	4.9%	100.0%

Residual quintile group * AFL13B Crosstabulation

			AFL13B					
			1.00	2.00	3.00	4.00	5.00	Total
Residual quintile group	WBE	Count	0	15	12	4	0	31
		% within Residual quintile group	0.0%	48.4%	38.7%	12.9%	0.0%	100.0%
	AE	Count	1	7	10	8	1	27
		% within Residual quintile group	3.7%	25.9%	37.0%	29.6%	3.7%	100.0%
	WAE	Count	1	6	10	5	1	23
		% within Residual quintile group	4.3%	26.1%	43.5%	21.7%	4.3%	100.0%
Total		Count	2	28	32	17	2	81
		% within Residual quintile group	2.5%	34.6%	39.5%	21.0%	2.5%	100.0%

Residual quintile group * AFL13C Crosstabulation

			AFL13C					
			1.00	2.00	3.00	4.00	5.00	Total
Residual quintile group	WBE	Count	1	2	2	13	13	31
		% within Residual quintile group	3.2%	6.5%	6.5%	41.9%	41.9%	100.0%
	AE	Count	0	2	3	5	17	27
		% within Residual quintile group	0.0%	7.4%	11.1%	18.5%	63.0%	100.0%
	WAE	Count	0	1	0	8	14	23
		% within Residual quintile group	0.0%	4.3%	0.0%	34.8%	60.9%	100.0%
Total		Count	1	5	5	26	44	81
		% within Residual quintile group	1.2%	6.2%	6.2%	32.1%	54.3%	100.0%

Residual quintile group * AFL13D Crosstabulation

			AFL13D					
			1.00	2.00	3.00	4.00	5.00	Total
	WBE	Count	0	3	11	14	3	31

Residual quintile group		% within Residual quintile group	0.0%	9.7%	35.5%	45.2%	9.7%	100.0%
	AE	Count	1	1	5	15	6	28
		% within Residual quintile group	3.6%	3.6%	17.9%	53.6%	21.4%	100.0%
	WAE	Count	0	1	2	16	4	23
		% within Residual quintile group	0.0%	4.3%	8.7%	69.6%	17.4%	100.0%
Total		Count	1	5	18	45	13	82
		% within Residual quintile group	1.2%	6.1%	22.0%	54.9%	15.9%	100.0%

Residual quintile group * AFL13E Crosstabulation

			AFL13E					
			1.00	2.00	3.00	4.00	5.00	Total
Residual quintile group	WBE	Count	0	8	12	11	0	31
		% within Residual quintile group	0.0%	25.8%	38.7%	35.5%	0.0%	100.0%
	AE	Count	2	11	6	6	1	26
		% within Residual quintile group	7.7%	42.3%	23.1%	23.1%	3.8%	100.0%
	WAE	Count	2	8	7	5	1	23
		% within Residual quintile group	8.7%	34.8%	30.4%	21.7%	4.3%	100.0%
Total		Count	4	27	25	22	2	80
		% within Residual quintile group	5.0%	33.8%	31.3%	27.5%	2.5%	100.0%

Residual quintile group * AFL13F Crosstabulation

			AFL13F					
			1.00	2.00	3.00	4.00	5.00	Total
Residual quintile group	WBE	Count	1	17	8	2	3	31
		% within Residual quintile group	3.2%	54.8%	25.8%	6.5%	9.7%	100.0%
	AE	Count	1	16	3	6	1	27
		% within Residual quintile group	3.7%	59.3%	11.1%	22.2%	3.7%	100.0%
	WAE	Count	0	9	7	5	2	23
		% within Residual quintile group	0.0%	39.1%	30.4%	21.7%	8.7%	100.0%
Total		Count	2	42	18	13	6	81
		% within Residual quintile group	2.5%	51.9%	22.2%	16.0%	7.4%	100.0%

Residual quintile group * AFL14A Crosstabulation

			AFL14A				
			2.00	3.00	4.00	5.00	Total
Residual quintile group	WBE	Count	3	15	9	3	30
		% within Residual quintile group	10.0%	50.0%	30.0%	10.0%	100.0%
	AE	Count	2	6	15	5	28
		% within Residual quintile group	7.1%	21.4%	53.6%	17.9%	100.0%
	WAE	Count	1	5	10	7	23
		% within Residual quintile group	4.3%	21.7%	43.5%	30.4%	100.0%

Total	Count	6	26	34	15	81
	% within Residual quintile group	7.4%	32.1%	42.0%	18.5%	100.0%

Residual quintile group * AFL14B Crosstabulation

			AFL14B					
			1.00	2.00	3.00	4.00	5.00	Total
Residual quintile group	WBE	Count	1	3	7	10	10	31
		% within Residual quintile group	3.2%	9.7%	22.6%	32.3%	32.3%	100.0%
	AE	Count	1	1	4	15	7	28
		% within Residual quintile group	3.6%	3.6%	14.3%	53.6%	25.0%	100.0%
	WAE	Count	0	0	3	8	11	22
		% within Residual quintile group	0.0%	0.0%	13.6%	36.4%	50.0%	100.0%
Total		Count	2	4	14	33	28	81
		% within Residual quintile group	2.5%	4.9%	17.3%	40.7%	34.6%	100.0%

Residual quintile group * AFL14C Crosstabulation

			AFL14C			
			3.00	4.00	5.00	Total
Residual quintile group	WBE	Count	3	16	12	31
		% within Residual quintile group	9.7%	51.6%	38.7%	100.0%
	AE	Count	1	9	18	28
		% within Residual quintile group	3.6%	32.1%	64.3%	100.0%
	WAE	Count	1	8	15	24
		% within Residual quintile group	4.2%	33.3%	62.5%	100.0%
Total		Count	5	33	45	83
		% within Residual quintile group	6.0%	39.8%	54.2%	100.0%

Residual quintile group * AFL14D Crosstabulation

			AFL14D				
			2.00	3.00	4.00	5.00	Total
Residual quintile group	WBE	Count	2	6	19	4	31
		% within Residual quintile group	6.5%	19.4%	61.3%	12.9%	100.0%
	AE	Count	1	3	10	14	28
		% within Residual quintile group	3.6%	10.7%	35.7%	50.0%	100.0%
	WAE	Count	0	5	7	12	24
		% within Residual quintile group	0.0%	20.8%	29.2%	50.0%	100.0%
Total		Count	3	14	36	30	83
		% within Residual quintile group	3.6%	16.9%	43.4%	36.1%	100.0%

Residual quintile group * AFL14E Crosstabulation

			AFL14E			
			3.00	4.00	5.00	Total
Residual quintile group	WBE	Count	3	16	12	31

		% within Residual quintile group	9.7%	51.6%	38.7%	100.0%
	AE	Count	1	8	18	27
		% within Residual quintile group	3.7%	29.6%	66.7%	100.0%
	WAE	Count	0	9	14	23
		% within Residual quintile group	0.0%	39.1%	60.9%	100.0%
Total		Count	4	33	44	81
		% within Residual quintile group	4.9%	40.7%	54.3%	100.0%

Residual quintile group * AFL14F Crosstabulation

			AFL14F					
			1.00	2.00	3.00	4.00	5.00	Total
Residual quintile group	WBE	Count	2	14	6	9	0	31
		% within Residual quintile group	6.5%	45.2%	19.4%	29.0%	0.0%	100.0%
	AE	Count	1	8	9	9	1	28
		% within Residual quintile group	3.6%	28.6%	32.1%	32.1%	3.6%	100.0%
	WAE	Count	1	5	8	9	1	24
		% within Residual quintile group	4.2%	20.8%	33.3%	37.5%	4.2%	100.0%
Total		Count	4	27	23	27	2	83
		% within Residual quintile group	4.8%	32.5%	27.7%	32.5%	2.4%	100.0%

Residual quintile group * AFL14G Crosstabulation

			AFL14G					
			1.00	2.00	3.00	4.00	5.00	Total
Residual quintile group	WBE	Count	2	7	9	11	2	31
		% within Residual quintile group	6.5%	22.6%	29.0%	35.5%	6.5%	100.0%
	AE	Count	1	3	11	7	6	28
		% within Residual quintile group	3.6%	10.7%	39.3%	25.0%	21.4%	100.0%
	WAE	Count	0	4	6	9	5	24
		% within Residual quintile group	0.0%	16.7%	25.0%	37.5%	20.8%	100.0%
Total		Count	3	14	26	27	13	83
		% within Residual quintile group	3.6%	16.9%	31.3%	32.5%	15.7%	100.0%

Residual quintile group * AFL14H Crosstabulation

			AFL14H					
			1.00	2.00	3.00	4.00	5.00	Total
Residual quintile group	WBE	Count	2	7	8	12	2	31
		% within Residual quintile group	6.5%	22.6%	25.8%	38.7%	6.5%	100.0%
	AE	Count	0	3	7	15	3	28
		% within Residual quintile group	0.0%	10.7%	25.0%	53.6%	10.7%	100.0%
	WAE	Count	0	2	4	12	6	24
		% within Residual quintile group	0.0%	8.3%	16.7%	50.0%	25.0%	100.0%

Total	Count	2	12	19	39	11	83
	% within Residual quintile group	2.4%	14.5%	22.9%	47.0%	13.3%	100.0%

Residual quintile group * AFL15A Crosstabulation

			AFL15A			
			2.00	4.00	5.00	Total
Residual quintile group	WBE	Count	1	6	24	31
		% within Residual quintile group	3.2%	19.4%	77.4%	100.0%
	AE	Count	0	5	23	28
		% within Residual quintile group	0.0%	17.9%	82.1%	100.0%
	WAE	Count	0	8	16	24
		% within Residual quintile group	0.0%	33.3%	66.7%	100.0%
Total		Count	1	19	63	83
		% within Residual quintile group	1.2%	22.9%	75.9%	100.0%

Residual quintile group * AFL15B Crosstabulation

			AFL15B				
			2.00	3.00	4.00	5.00	Total
Residual quintile group	WBE	Count	1	1	9	20	31
		% within Residual quintile group	3.2%	3.2%	29.0%	64.5%	100.0%
	AE	Count	0	1	5	22	28
		% within Residual quintile group	0.0%	3.6%	17.9%	78.6%	100.0%
	WAE	Count	0	1	7	15	23
		% within Residual quintile group	0.0%	4.3%	30.4%	65.2%	100.0%
Total		Count	1	3	21	57	82
		% within Residual quintile group	1.2%	3.7%	25.6%	69.5%	100.0%

Residual quintile group * AFL15C Crosstabulation

			AFL15C				
			2.00	3.00	4.00	5.00	Total
Residual quintile group	WBE	Count	0	3	12	16	31
		% within Residual quintile group	0.0%	9.7%	38.7%	51.6%	100.0%
	AE	Count	1	0	7	20	28
		% within Residual quintile group	3.6%	0.0%	25.0%	71.4%	100.0%
	WAE	Count	0	1	9	13	23
		% within Residual quintile group	0.0%	4.3%	39.1%	56.5%	100.0%
Total		Count	1	4	28	49	82
		% within Residual quintile group	1.2%	4.9%	34.1%	59.8%	100.0%

Residual quintile group * AFL15D Crosstabulation

			AFL15D					
			1.00	2.00	3.00	4.00	5.00	Total
	WBE	Count	1	1	4	11	14	31

Residual quintile group		% within Residual quintile group	3.2%	3.2%	12.9%	35.5%	45.2%	100.0%
	AE	Count	0	0	2	7	18	27
		% within Residual quintile group	0.0%	0.0%	7.4%	25.9%	66.7%	100.0%
	WAE	Count	0	0	0	8	13	21
		% within Residual quintile group	0.0%	0.0%	0.0%	38.1%	61.9%	100.0%
Total		Count	1	1	6	26	45	79
		% within Residual quintile group	1.3%	1.3%	7.6%	32.9%	57.0%	100.0%

Residual quintile group * AFL15E Crosstabulation

			AFL15E					
			1.00	2.00	3.00	4.00	5.00	Total
Residual quintile group	WBE	Count	0	1	2	16	12	31
		% within Residual quintile group	0.0%	3.2%	6.5%	51.6%	38.7%	100.0%
	AE	Count	0	1	3	10	14	28
		% within Residual quintile group	0.0%	3.6%	10.7%	35.7%	50.0%	100.0%
	WAE	Count	1	1	1	7	14	24
		% within Residual quintile group	4.2%	4.2%	4.2%	29.2%	58.3%	100.0%
Total		Count	1	3	6	33	40	83
		% within Residual quintile group	1.2%	3.6%	7.2%	39.8%	48.2%	100.0%

RESPONDENT DATA DISAGGREGATED INTO CASE STUDY (CS) SCHOOLS, SCHOOLS THAT IDENTIFIED THEMSELVES AND REMAINDER (ANONYMOUS)

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Residual quintile group * Gender * Status within residual quintile group	78	91.8%	7	8.2%	85	100.0%
Residual quintile group * Teaching experience * Status within residual quintile group	78	91.8%	7	8.2%	85	100.0%
Residual quintile group * Science teacher by training * Status within residual quintile group	80	94.1%	5	5.9%	85	100.0%
Residual quintile group * Alternative quals * Status within residual quintile group	85	100.0%	0	0.0%	85	100.0%
Residual quintile group * Head teacher or not * Status within residual quintile group	81	95.3%	4	4.7%	85	100.0%
Residual quintile group * Highest qualification * Status within residual quintile group	79	92.9%	6	7.1%	85	100.0%
Residual quintile group * Year highest qual completed * Status within residual quintile group	85	100.0%	0	0.0%	85	100.0%
Residual quintile group * Where trained * Status within residual quintile group	78	91.8%	7	8.2%	85	100.0%
Residual quintile group * Last taught Yr 7-9 classes * Status within residual quintile group	81	95.3%	4	4.7%	85	100.0%
Residual quintile group * Y8 classes at your school * Status within residual quintile group	79	92.9%	6	7.1%	85	100.0%
Residual quintile group * FT science teachers * Status within residual quintile group	78	91.8%	7	8.2%	85	100.0%
Residual quintile group * PT science teachers * Status within residual quintile group	64	75.3%	21	24.7%	85	100.0%

REFER TO THE QUESIONNAIRE FOR THE KEY LINKING NUMBERS TO RESPONSE

Residual quintile group * Gender * Status within residual quintile group Crosstabulation

				Gender		
Status within residual quintile group				1	2	Total
UNKNOWN	Residual quintile group	WBE	Count	13	6	19
			% within Residual quintile group	68.4%	31.6%	100.0%
		AE	Count	12	4	16
			% within Residual quintile group	75.0%	25.0%	100.0%
		WAE	Count	8	5	13
			% within Residual quintile group	61.5%	38.5%	100.0%
	Total		Count	33	15	48
			% within Residual quintile group	68.8%	31.3%	100.0%
IDKNOWN	Residual quintile group	WBE	Count	4	1	5
			% within Residual quintile group	80.0%	20.0%	100.0%
		AE	Count	3	3	6
			% within Residual quintile group	50.0%	50.0%	100.0%
		WAE	Count	3	1	4
			% within Residual quintile group	75.0%	25.0%	100.0%
	Total		Count	10	5	15
			% within Residual quintile group	66.7%	33.3%	100.0%
CSSCHOOL	Residual quintile group	WBE	Count	3	2	5
			% within Residual quintile group	60.0%	40.0%	100.0%
		AE	Count	3	1	4
			% within Residual quintile group	75.0%	25.0%	100.0%
		WAE	Count	5	1	6
			% within Residual quintile group	83.3%	16.7%	100.0%
	Total		Count	11	4	15
			% within Residual quintile group	73.3%	26.7%	100.0%
Total	Residual quintile group	WBE	Count	20	9	29
			% within Residual quintile group	69.0%	31.0%	100.0%
		AE	Count	18	8	26
			% within Residual quintile group	69.2%	30.8%	100.0%
		WAE	Count	16	7	23
			% within Residual quintile group	69.6%	30.4%	100.0%
	Total		Count	54	24	78
			% within Residual quintile group	69.2%	30.8%	100.0%

Residual quintile group * Teaching experience * Status within residual quintile group Crosstabulation

				Teaching experience				
Status within residual quintile group				1	2	3	4	Total
UNKNOWN	Residual quintile group	WBE	Count	8	2	1	8	19
			% within Residual quintile group	42.1%	10.5%	5.3%	42.1%	100.0%
		AE	Count	1	3	3	9	16
			% within Residual quintile group	6.3%	18.8%	18.8%	56.3%	100.0%
		WAE	Count	1	2	3	7	13
			% within Residual quintile group	7.7%	15.4%	23.1%	53.8%	100.0%
	Total		Count	10	7	7	24	48
			% within Residual quintile group	20.8%	14.6%	14.6%	50.0%	100.0%
IDKNOWN	Residual quintile group	WBE	Count		1	0	4	5
			% within Residual quintile group		20.0%	0.0%	80.0%	100.0%
		AE	Count		2	0	4	6
			% within Residual quintile group		33.3%	0.0%	66.7%	100.0%
		WAE	Count		1	1	2	4
			% within Residual quintile group		25.0%	25.0%	50.0%	100.0%
	Total		Count		4	1	10	15
			% within Residual quintile group		26.7%	6.7%	66.7%	100.0%
CSSCHOOL	Residual quintile group	WBE	Count		1	2	2	5
			% within Residual quintile group		20.0%	40.0%	40.0%	100.0%
		AE	Count		0	1	3	4
			% within Residual quintile group		0.0%	25.0%	75.0%	100.0%
		WAE	Count		0	1	5	6
			% within Residual quintile group		0.0%	16.7%	83.3%	100.0%
	Total		Count		1	4	10	15
			% within Residual quintile group		6.7%	26.7%	66.7%	100.0%
Total	Residual quintile group	WBE	Count	8	4	3	14	29
			% within Residual quintile group	27.6%	13.8%	10.3%	48.3%	100.0%
		AE	Count	1	5	4	16	26
			% within Residual quintile group	3.8%	19.2%	15.4%	61.5%	100.0%
		WAE	Count	1	3	5	14	23
			% within Residual quintile group	4.3%	13.0%	21.7%	60.9%	100.0%
	Total		Count	10	12	12	44	78
			% within Residual quintile group	12.8%	15.4%	15.4%	56.4%	100.0%

Residual quintile group * Science teacher by training * Status within residual quintile group Crosstabulation

Status within residual quintile group				Science teacher by training		Total
				1	2	
UNKNOWN	Residual quintile group	WBE	Count	16	3	19
			% within Residual quintile group	84.2%	15.8%	100.0%
		AE	Count	18	0	18
			% within Residual quintile group	100.0%	0.0%	100.0%
		WAE	Count	12	1	13
			% within Residual quintile group	92.3%	7.7%	100.0%
	Total		Count	46	4	50
			% within Residual quintile group	92.0%	8.0%	100.0%
IDKNOWN	Residual quintile group	WBE	Count	5		5
			% within Residual quintile group	100.0%		100.0%
		AE	Count	6		6
			% within Residual quintile group	100.0%		100.0%
		WAE	Count	4		4
			% within Residual quintile group	100.0%		100.0%
	Total		Count	15		15
			% within Residual quintile group	100.0%		100.0%
CSSCHOOL	Residual quintile group	WBE	Count	5		5
			% within Residual quintile group	100.0%		100.0%
		AE	Count	4		4
			% within Residual quintile group	100.0%		100.0%
		WAE	Count	6		6
			% within Residual quintile group	100.0%		100.0%
	Total		Count	15		15
			% within Residual quintile group	100.0%		100.0%
Total	Residual quintile group	WBE	Count	26	3	29
			% within Residual quintile group	89.7%	10.3%	100.0%
		AE	Count	28	0	28
			% within Residual quintile group	100.0%	0.0%	100.0%
		WAE	Count	22	1	23
			% within Residual quintile group	95.7%	4.3%	100.0%
	Total		Count	76	4	80
			% within Residual quintile group	95.0%	5.0%	100.0%

Residual quintile group * Alternative quals * Status within residual quintile group Crosstabulation

				Alternative quals		
Status within residual quintile group					.0	Total
UNKNOWN	Residual quintile group	WBE	Count	18	4	22
			% within Residual quintile group	81.8%	18.2%	100.0%
		AE	Count	17	1	18
			% within Residual quintile group	94.4%	5.6%	100.0%
		WAE	Count	12	3	15
			% within Residual quintile group	80.0%	20.0%	100.0%
	Total		Count	47	8	55
			% within Residual quintile group	85.5%	14.5%	100.0%
IDKNOWN	Residual quintile group	WBE	Count	5	0	5
			% within Residual quintile group	100.0%	0.0%	100.0%
		AE	Count	5	1	6
			% within Residual quintile group	83.3%	16.7%	100.0%
		WAE	Count	4	0	4
			% within Residual quintile group	100.0%	0.0%	100.0%
	Total		Count	14	1	15
			% within Residual quintile group	93.3%	6.7%	100.0%
CSSCHOOL	Residual quintile group	WBE	Count	3	2	5
			% within Residual quintile group	60.0%	40.0%	100.0%
		AE	Count	4	0	4
			% within Residual quintile group	100.0%	0.0%	100.0%
		WAE	Count	5	1	6
			% within Residual quintile group	83.3%	16.7%	100.0%
	Total		Count	12	3	15
			% within Residual quintile group	80.0%	20.0%	100.0%
Total	Residual quintile group	WBE	Count	26	6	32
			% within Residual quintile group	81.3%	18.8%	100.0%
		AE	Count	26	2	28
			% within Residual quintile group	92.9%	7.1%	100.0%
		WAE	Count	21	4	25
			% within Residual quintile group	84.0%	16.0%	100.0%
	Total		Count	73	12	85
			% within Residual quintile group	85.9%	14.1%	100.0%

Residual quintile group * Head teacher or not * Status within residual quintile group Crosstabulation

Status within residual quintile group				Head teacher or not 1	 2	Total
UNKNOWN	Residual quintile group	WBE	Count	4	15	19
			% within Residual quintile group	21.1%	78.9%	100.0%
		AE	Count	6	12	18
			% within Residual quintile group	33.3%	66.7%	100.0%
		WAE	Count	5	9	14
			% within Residual quintile group	35.7%	64.3%	100.0%
	Total		Count	15	36	51
			% within Residual quintile group	29.4%	70.6%	100.0%
IDKNOWN	Residual quintile group	WBE	Count	4	1	5
			% within Residual quintile group	80.0%	20.0%	100.0%
		AE	Count	4	2	6
			% within Residual quintile group	66.7%	33.3%	100.0%
		WAE	Count	4	0	4
			% within Residual quintile group	100.0%	0.0%	100.0%
	Total		Count	12	3	15
			% within Residual quintile group	80.0%	20.0%	100.0%
CSSCHOOL	Residual quintile group	WBE	Count	4	1	5
			% within Residual quintile group	80.0%	20.0%	100.0%
		AE	Count	4	0	4
			% within Residual quintile group	100.0%	0.0%	100.0%
		WAE	Count	4	2	6
			% within Residual quintile group	66.7%	33.3%	100.0%
	Total		Count	12	3	15
			% within Residual quintile group	80.0%	20.0%	100.0%
Total	Residual quintile group	WBE	Count	12	17	29
			% within Residual quintile group	41.4%	58.6%	100.0%
		AE	Count	14	14	28
			% within Residual quintile group	50.0%	50.0%	100.0%
		WAE	Count	13	11	24
			% within Residual quintile group	54.2%	45.8%	100.0%
	Total		Count	39	42	81
			% within Residual quintile group	48.1%	51.9%	100.0%

Residual quintile group * Highest qualification * Status within residual quintile group Crosstabulation

				Highest qualification					
Status within residual quintile group				1	2	3	4	5	Total
UNKNOWN	Residual quintile group	WBE	Count	10	5	1	1	0	17
			% within Residual quintile group	58.8%	29.4%	5.9%	5.9%	0.0%	100.0%
		AE	Count	11	2	3	0	2	18
			% within Residual quintile group	61.1%	11.1%	16.7%	0.0%	11.1%	100.0%
		WAE	Count	9	1	2	0	2	14
			% within Residual quintile group	64.3%	7.1%	14.3%	0.0%	14.3%	100.0%
	Total		Count	30	8	6	1	4	49
			% within Residual quintile group	61.2%	16.3%	12.2%	2.0%	8.2%	100.0%
IDKNOWN	Residual quintile group	WBE	Count	5	0				5
			% within Residual quintile group	100.0%	0.0%				100.0%
		AE	Count	5	1				6
			% within Residual quintile group	83.3%	16.7%				100.0%
		WAE	Count	2	2				4
			% within Residual quintile group	50.0%	50.0%				100.0%
	Total		Count	12	3				15
			% within Residual quintile group	80.0%	20.0%				100.0%
CSSCHOOL	Residual quintile group	WBE	Count	4	1	0			5
			% within Residual quintile group	80.0%	20.0%	0.0%			100.0%
		AE	Count	4	0	0			4
			% within Residual quintile group	100.0%	0.0%	0.0%			100.0%
		WAE	Count	5	0	1			6
			% within Residual quintile group	83.3%	0.0%	16.7%			100.0%
	Total		Count	13	1	1			15
			% within Residual quintile group	86.7%	6.7%	6.7%			100.0%
Total	Residual quintile group	WBE	Count	19	6	1	1	0	27
			% within Residual quintile group	70.4%	22.2%	3.7%	3.7%	0.0%	100.0%
		AE	Count	20	3	3	0	2	28

			% within Residual quintile group	71.4%	10.7%	10.7%	0.0%	7.1%	100.0%
		WAE	Count	16	3	3	0	2	24
			% within Residual quintile group	66.7%	12.5%	12.5%	0.0%	8.3%	100.0%
	Total		Count	55	12	7	1	4	79
			% within Residual quintile group	69.6%	15.2%	8.9%	1.3%	5.1%	100.0%

Residual quintile group * Where trained * Status within residual quintile group Crosstabulation

Status within residual quintile group				Where trained			Total
				1	2	3	
UNKNOWN	Residual quintile group	WBE	Count	0	2	15	17
			% within Residual quintile group	0.0%	11.8%	88.2%	100.0%
		AE	Count	4	2	12	18
			% within Residual quintile group	22.2%	11.1%	66.7%	100.0%
		WAE	Count	2	1	11	14
			% within Residual quintile group	14.3%	7.1%	78.6%	100.0%
	Total		Count	6	5	38	49
			% within Residual quintile group	12.2%	10.2%	77.6%	100.0%
IDKNOWN	Residual quintile group	WBE	Count			5	5
			% within Residual quintile group			100.0%	100.0%
		AE	Count			6	6
			% within Residual quintile group			100.0%	100.0%
		WAE	Count			4	4
			% within Residual quintile group			100.0%	100.0%
	Total		Count			15	15
			% within Residual quintile group			100.0%	100.0%
CSSCHOOL	Residual quintile group	WBE	Count	0		5	5
			% within Residual quintile group	0.0%		100.0%	100.0%
		AE	Count	1		3	4
			% within Residual quintile group	25.0%		75.0%	100.0%
		WAE	Count	1		4	5
			% within Residual quintile group	20.0%		80.0%	100.0%
	Total		Count	2		12	14
			% within Residual quintile group	14.3%		85.7%	100.0%
Total	Residual quintile group	WBE	Count	0	2	25	27
			% within Residual quintile group	0.0%	7.4%	92.6%	100.0%
		AE	Count	5	2	21	28

			% within Residual quintile group	17.9%	7.1%	75.0%	100.0%
		WAE	Count	3	1	19	23
			% within Residual quintile group	13.0%	4.3%	82.6%	100.0%
	Total		Count	8	5	65	78
			% within Residual quintile group	10.3%	6.4%	83.3%	100.0%

REFERENCES

AAS, Australian Academy of Sciences. (2016). *Primary Connections: Linking science with literacy.* Retrieved from https://primaryconnections.org.au/about

AAS, Australian Academy of Science. (2017). *Science by Doing -- Home*. Retrieved from https://www.sciencebydoing.edu.au/

ABS, Australian Burea of Statistics. (2018). 4221.0 Schools Australia, 2017. Retrieved from http://www.abs.gov.au/ausstats/abs@.nsf/PrimaryMainFeatures/4221.0?OpenDocument

ACACA, Australian Curriculum, Assessment and Certification Authorities. (2018). Home. Retrieved from http://www.acaca.edu.au/index.php/schooling/assessment-and-reporting/

ACARA, Australian Curriculum Assessment and Reporting Authority. (2013a). Guide to understanding ICSEA (Index of Community Socio-educational Advantage) values from 2013 onwards. Retrieved from https://acaraweb.blob.core.windows.net/resources/Guide_to_understanding_ICSEA_values.pdf

ACARA, Australian Curriculum Assessment and Reporting Authority. (2013b). NAPLAN. Retrieved from http://www.nap.edu.au/naplan/naplan.html

ACARA, Australian Curriculum Assessment and Reporting Authority. (2014a). *2012* NAP SL Public Report. Retrieved from http://www.nap.edu.au/results-and-reports/national-reports.html

ACARA, Australian Curriculum Assessment and Reporting Authority. (2014b). ICSEA 2013: Technical Report. Retrieved from https://www.myschool.edu.au/MoreInformation

ACARA, Australian Curriculum Assessment and Reporting Authority. (2014c). The Australian Curriculum: Science F-10. Retrieved from http://www.australiancurriculum.edu.au/Download/F10

ACARA, Australian Curriculum Assessment and Reporting Authority. (2015). ICSEA 2014: Technical Report. Retrieved from https://www.myschool.edu.au/media/1033/icsea_2014_technical_report.pdf

ACARA, Australian Curriculum Assessment and Reporting Authority. (2016a). About us. Retrieved from https://www.acara.edu.au/about-us

ACARA, Australian Curriculum Assessment and Reporting Authority. (2016b). My School website / About. Retrieved from http://www.myschool.edu.au/about/

ACARA, Australian Curriculum Assessment and Reporting Authority. (2016c). Reporting. Retrieved from https://www.acara.edu.au/reporting

ACARA, Australian Curriculum Assessment and Reporting Authority. (2016d). NAP website: Welcome. Retrieved from http://www.nap.edu.au/about/why-nap.html

ACARA, Australian Curriculum Assessment and Reporting Authority. (2017). NAP Sample Assessments-Science Literacy. Retrieved from http://www.nap.edu.au/nap-sample-assessments/science-literacy

ACARA, Australian Curriculum Assessment and Reporting Authority. (2018). Science: Sequence of Achievement: 7-10. Retrieved from http://docs.acara.edu.au/resources/Science_Sequence_of_achievement.pdf

ACER, Australian Council for Educational Research. (2004a). Science Education Assessment Resource (SEAR): Final Report. Retrieved from http://cms.curriculum.edu.au/sear/newcms/view_page.asp?page_id=3526

ACER, Australian Council for Educational Research. (2004b). Scientific Literacy Progress Map (pp 4). Melbourne, Victoria: Australian Council for Educational Research.

AECRC, Australian Education Council Review Committee. (1992). *Key Competencies. Report of the Committee to advise the Australian Education Council and Ministers of Vocational Education, Employment and Training on employment-related Key Competencies for postcompulsory education and training*. Retrieved from http://www.voced.edu.au/content/ngv%3A28045

ARG, Assessment Reform Group. (2002a). Assessment for Learning: 10 principles. Retrieved from http://www.nuffieldfoundation.org/assessment-reform-group

ARG, Assessment Reform Group. (2002b). Testing, Motivation and Learning. Retrieved from http://www.nuffieldfoundation.org/assessment-reform-group

ARG, Assessment Reform Group. (2006). The role of teachers in the assessment of learning. Retrieved from http://www.nuffieldfoundation.org/assessment-reform-group

Au, W. (2007). High-Stakes Testing and Curricular Control: A Qualitative Metasynthesis. *Educational Researcher, 36*, 11.

Australia, Commonwealth of. (2001). Backing Australia's Ability: An action plan for the future. Canberra: Commonwealth of Australia. Retrieved from https://trove.nla.gov.au/work/34335833.

Ball, S., Rae, I., & Tognolini, J. (2000). A report for the National Education Performance Monitoring Taskforce: options for the assessment and reporting of primary students in the key learning area of science to be used in the reporting of nationally comparable outcomes of schooling within the context of the National Goals for Schooling in the Twenty-First Century. Retrieved from http://educationcouncil.edu.au/site/DefaultSite/filesystem/documents/Reports%20and%20publications/Archive%20Publications/Measuring%20and%20Reporting%20Student%20Performance/Assessment_Primary_Students_Science-Context_National_Goals.pdf

Batterham, R. (2000). *The Chance to Change: Final Report*. Canberra: Department of Science, Industry and Resources.

Bell, B., & Cowie, B. (2002). *Formative Assessment and Science Education* (Vol. 12). Dordrecht: Kluwer.

Beveridge, M. (1985). The development of young childrens' understanding of the process of evaporation. *British Journal of Educational psychology, 55*, 84-90.

Biggs, J. (1995). Assessing for Learning: Some dimensions underlying new approaches to educational assessment. *Alberta Journal of Educational Research, 41*(1), 18.

Biggs, J. (1998). Assessment and Classroom Learning: a role for summative assessment? *Assessment in Education: Principles, Policy & Practice, 5*(1), 103-110. doi:10.1080/0969595980050106

Biggs, J. (1999). What the Student Does: teaching for enhanced learning. *Higher Education Research & Development, 18*(1), 57-75.

Biggs, J., & Collis, K. (1982). *Evaluating the Quality of Learning: The SOLO (Structure of the Observed Learning Outcome) Taxonomy*. New York: Academic Press.

Biggs, J., & Collis, K. (1991). Multimodal learning and the quality of intelligent behaviour. In Helga Rowe (Ed.), *Intelligence: Reconceptualization and measurement* (pp. 57-76). Melbourne, Victoria: ACER.

Billett, S. (1996). Situated learning: Bridging sociocultural and cognitive theorising. *Learning and Instruction, 6*(3), 263-280.

Black, P. (2007). Full marks for feedback. *Making the Grade (Journal of the Institute of Educational Assessors), Spring 2007*, 18-21.

Black, P. (2013). Formative and Summative Aspects of Assessment: Theoretical and Research Foundations in the Context of Pedagogy. SAGE Handbook of Research on Classroom Assessment. SAGE Publications, Inc. Thousand Oaks, CA: SAGE Publications, Inc.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2004). *Working inside the black box: assessment for learning in the classroom*. London: NFER-Nelson.

Black, P., McCormick, R., James, M., & Pedder, D. (2006). Learning How to Learn and Assessment for Learning: a theoretical inquiry. *Research Papers in Education - Special Issue, 21*(2), 119-132.

Black, P., & Wiliam, D. (1998a). Assessment and Classroom Learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7-74. doi:10.1080/0969595980050102

Black, P., & Wiliam, D. (1998b). *Inside the black box : raising standards through classroom assessment*. London: King's College London. Dept. of Education & Professional Studies.

Black, P., & Wiliam, D. (2005). Changing teaching through formative assessment: Research and practice. *CERI, 2005*, 223-240. Retrieved from http://www.oecd.org/education/ceri/35337920.pdf

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability (formerly: Journal of Personnel Evaluation in Education), 21*(1), 5-31. doi:10.1007/s11092-008-9068-5

Bloom, B., Engelhart, M., Furst, E., Hill, W., & Krathwohl, D. (Eds.). (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain.* New York: David McKay.

Bøe, M. V., Henriksen, E. K., Lyons, T., & Schreiner, C. (2013). Participation in science and technology: young people's achievement-related choices in late-modern socieities. *Studies in Science Education, 47*(1), 37-72. doi:10.1080/03057267.2011.549621

Boekaerts, M., & Corno, L. (2005). Self-Regulation in the Classroom: A Perspective on Assessment and Intervention. *Applied Psychology: An International Review, 54*(2), 199-231.

Boekaerts, M., Maes, S., & Karoly, P. (2005). Self-Regulation Across Domains of Applied Psychology: Is there an Emerging Consensus? *Applied Psychology: An International Review, 54*(2), 15.

BOS, Board of Studies NSW. (2003). *Science Years 7-10 Syllabus* (Vol. 2013). Sydney: Board of Studies NSW.

BOS, Board of Studies NSW. (2011). *School Certificate Review--Discussion Paper*. Sydney: Board of Studies NSW.

BOS, Board of Studies NSW. (2013). *About the Common Grade Scale*. Retrieved from http://arc.nesa.nsw.edu.au/go/7-8/common-grade-scale/.

BOS, Board of Studies NSW. (n.d.). Stage 5 Course Performance Descriptors--Science. Retrieved from http://arc.nesa.nsw.edu.au/go/9-10/stage-5-grading/cpds/index/science

BOSTES, Board of Studies, Teaching and Educational Standards NSW. (2012). NSW *Syllabuses for the Australian Curriculum: Science K-10*. Retrieved from http://syllabus.bostes.nsw.edu.au/science/

Boyle, S., Fahey, E., Loughran, J., & Mitchell, I. (2001). Classroom research into good learning behaviours. *Educational Action Research, 9*(2), 27. doi:10.1080/09650790100200149

Broadfoot, P. (2009). Foreword. In J. Cumming & C. Wyatt-Smith (Eds.), Educational assessment in the 21st century: connecting theory and practice (pp. 309). Dordrecht: Springer. Retrieved from http://www.lib.uts.edu.au/sso/goto.php?url=http://dx.doi.org/10.1007/978-1-4020-9964-9. doi:10.1007/978-1-4020-9964-9_3

Broadfoot, P., & Black, P. (2004). Redefining assessment? The first ten years of Assessment in Education. *Assessment in Education: Principles, Policy & Practice, 11*(1), 7-27. doi:10.1080/0969594042000208976

Brookhart, S. (2003). Developing Measurement Theory for Classroom Assessment Purposes and Uses. *Educational Measurement: Issues and Practice, 22*(4), 5-12.

Bryman, A. (2012). *Social Science Methods*. New York: Oxford University Press.

CC, Curriculum Corporation. (n.d.). Assessment for Learning website. Retrieved from http://www.assessmentforlearning.edu.au/default.asp

CERI, OECD-Centre for Educational Research and Information. (2005). Formative Assessment: Improving learning in secondary classrooms. Retrieved from http://www.oecd.org/education/ceri/35661078.pdf

CERI, OECD-Centre for Educational Research and Information. (2008). 21st Century Learning: Research, Innovation and Policy. Directions from recent OECD analysis (pp. 13). Retrieved from http://www.oecd.org/site/educeri21st/40554299.pdf

CGCS, Council of the Great City Schools. (2015). Student Testing in America's Great City Schools: An Inventory and Preliminary Analysis. Retrieved from http://www.cgcs.org/cms/lib/DC00001581/Centricity/Domain/87/Testing Report.pdf

Chubb, Ian. (2012). Mathematics, Engineering & Science in the National Interest. Retrieved from Canberra:

http://www.chiefscientist.gov.au/category/archives/mathematics-engineering-and-science-report/

Clark, I. (2012). Formative Assessment: Assessment Is for Self-regulated Learning. *Educational Psychology Review, 24*, 205-249. doi:10.1007/s10648-011-9191-6

Commonwealth of Australia Constitution Act (The Constitution). Retrieved from https://www.legislation.gov.au/Details/C2013Q00005

Connell, R. W. (1985). The Competitive Academic Curriculum. In D. Cohen & T. Maxwell (Eds.), *Blocked at the Entrance: Context, Cases and Commentary on Curriculum Change*: Entrance Publications.

Cooney, G. (2006). Review of Assessments in the context of National Developments. Retrieved from https://www.det.nsw.edu.au/media/downloads/dethome/yr2007/cooneyreviewfll.pdf

Corcoran, T., Mosher, F. A., & Rogat, A. (2009). *Learning Progressions in Science: An evidence-based approach to reform*. Retrieved from http://www.cpre.org/ and http://www.ccii-cpre.org/.

Corrigan, D., Gunstone, R., & Jones, A. (Eds.). (2013). *Valuing Assessment in Science Education: Pedagogy, Curriculum, Policy*: Dordrecht: Springer.

Cowie, B. (2005). Student commentary on classroom assessment in science: a sociocultural interpretation. *International Journal of Science Education, 27*(2), 199-214. doi:10.1080/0950069042000276721

Cowie, B. (2013). *Assessment in the Science Classroom: Priorities, Practices, and Prospects. SAGE Handbook of Research on Classroom Assessment. SAGE Publications, Inc*. Thousand Oaks, CA: SAGE Publications, Inc.

Cowie, B., & Bell, B. (1999). A Model of Formative Assessment in Science. *Assessment in Education: Principles, Policy & Practice, 6*(1), 101-116. doi:10.1080/09695949993026

Creswell, J. W. (2012). *Educational Research: Planning, Conducting, and Evaluating Quantitative and Qualitative Research (4e)*. Boston, MA: Pearson.

Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and Conducting Mixed Methods Research (2E)*. Thousand Oaks, CA: SAGE.

CRTTE, Committee for the Review of Teaching and Teacher Education. (2003). *Australia's Teachers: Australia's Future. Advancing Innovation, Science, Technology and Mathematics.(Dow Report)*. Canberra, ACT: Department of Education, Science and Training.

Cumming, Joy, & Wyatt-Smith, Claire. (2009). *Educational assessment in the 21st century: connecting theory and practice* (pp. xxv). Retrieved from http://www.lib.uts.edu.au/sso/goto.php?url=http://dx.doi.org/10.1007/978-1-4020-9964-9 doi:10.1007/978-1-4020-9964-9_3

CURASS, Australian Education Council Curriculum and Assessment Committee. (1994). *Science--a curriculum profile for Australian schools*. Carlton, Victoria, Australia: Curriculum Corporation.

Dann, R. (2002). *Promoting assessment as learning: improving the learning process*. London: Routledge/Falmer.

Darling-Hammond, L. (2003). Standards and Assessments: Where We Are and What We Need. Teachers College Record. # 11109 (2/16/2003). http://www.tcrecord.org.

Deakin-Crick, R., Broadfoot, P., & Claxton, G. (2004). Developing an Effective Lifelong Learning Inventory: the ELLI project. *Assessment in Education, 11*(3), 247-272. doi:10.1080/0969594042000304582

DEC, NSW Department of Education and Communities. (2014). *ESSA test booklet (pp. 31)*. Sydney: Department of Education and Communities

DEC, NSW Department of Education and Communities. (2015). *Essential Secondary Science Assessment 2014 state report*. Sydney: NSW Department of Education and Communities.

Denzin, N. K., & Lincoln, Y. S. (Eds.). (2011). *The SAGE Handbook of Qualitative Research 4E*. Washington DC: SAGE.

DES, Department for Education and Skills. (2003). 21st Century Skills. Realising our Potential. Individuals, Employers, Nation. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/336816/21st_Century_Skills_Realising_Our_Potential.pdf

DET, NSW Department of Education and Training. (2003*). Quality Teaching in NSW Public Schools: Discussion Paper.* Sydney, Australia: NSW DET, Professional Support and Curriculum Directorate.

DET, NSW Department of Education and Training. (2006). *Quality teaching in NSW public schools: An assessment practice guide*. Sydney: DET.

DET, NSW Department of Education and Training. (2007). *ESSA report for parents (2006 example).* Retrieved from http://www.schools.nsw.edu.au/learning/7-12assessments/essa/essasmart.php

DET, NSW Department of Education and Training. (2008). *Principles of Assessment and Reporting in NSW Public Schools*. Sydney: NSW Department of Education and Training. Retrieved from http://www.curriculumsupport.education.nsw.gov.au/timetoteach/assess/princep_ass.htm.

DET, NSW Department of Education and Training. (2011). Essential Secondary Science Assessment 2011 framework. Sydney: EMSAD.

DET, NSW Department of Education and Training. (2015). *DN/15/00033: Critical changes to Essential Secondary Science Assessment (ESSA)* Sydney. Sydney: Department of Education and Training.

Driver, R., & Easley, J. (1978). Pupils and paradigms: A review of literature related to concept development in adolescent science students. *Studies in Science Education, 5*, 61-84.

Dulfer, N., Polesel, J., & Rice, S. (2012). *The Experience of Education: The impacts of high stakes testing on school students and their families. An Educator's Perspective*. Rydalmere: The Whitlam Institute.

Duschl, R. A., Schweingruber, H. A., & Shouse, A. W. (Eds.). (2007). *Taking Science to School: Learning and Teaching Science in Grades K-8.* Washington, DC: The National Academies Press.

EAA, Educational Assessment Australia. (2018). ICAS Science. Retrieved from https://www.eaa.unsw.edu.au/icas/subjects/science

Earl, K., & Giles, D. (2011). An-other Look at Assessment: Assessment in Learning. *New Zealand Journal of Teachers' Work, 8*(1), 11-20.

ESA, Education Services Australia. (n.d.). Home. Retrieved from https://www.esa.edu.au/solutions/our-solutions

ESA, Education Services Australia. (2012). National Digital Learning Resources Network: Science Retrieved from http://www.ndlrn.edu.au/using_digital_resources/australian_curriculum_resources/science.html

Fensham, P. (2013). International Assessments of Science Learning: Their Positive and Negative Contributions to Science Education. In D. Corrigan, R. Gunstone, & A. Jones (Eds.), *Valuing Assessment in Science Education: Pedagogy, Curriculum, Policy* (pp 11-32). Dordrecht: Springer.

Fensham, P., & Rennie, L. (2013). Towards an Authentically Assessed Science Curriculum. In D. Corrigan, R. Gunstone, & A. Jones (Eds.), *Valuing Assessment in Science Education: Pedagogy, Curriculum, Policy* (pp 69-100). Dordrecht: Springer.

Flyvbjerg, B. (2011). Case Study. In N. K. Denzin & Y. S. Lincoln (Eds.), *The SAGE Handbook of Qualitative Research (4e)* (pp. 301-316). Thousand Oaks, CA: Sage.

Fraser, B. L. (1978). Development of a test of science-related attitudes. *Science Education, 62*(509-515).

Frey, B. B., & Schmitt, V. L. (2007). Coming to Terms With Classroom Assessment. *Journal of Advanced Academics, 18*(3), 402-423.

Ginsberg, H., & Opper, S. (1979). *Piaget's Theory of Intellectual Development* (2nd ed.). New Jersey: Prentice-Hall Inc.

Goodrum, D., Rennie, L. J., & Hackling, M. (2001). *The Status and Quality of Teaching and Learning of Science in Australian Schools*. Canberra: Department of Education, Training and Youth Affairs

Gipps, C. (1999). Chapter 10: Socio-Cultural Aspects of Assessment. *Reveiw of Research in Education*. doi:10.3102/0091732X024001355

Goodrum, D., & Rennie, L. J. (2007). *Australian School Science Education National Plan 2008-2012 Volume 1 The National Action Plan*. Canberra, Department of Education, Science and Training.

Griffin, P. (2009). Teachers' Use of Assessment Data. In C. Wyatt-Smith & J. J. Cumming (Eds.), *Educational Assessment in the 21st Century* (pp. 25). Dordrecht: Springer. doi:10.1007/978-1-4020-9964-9

Guba, E. G. (1981). Criteria for assessing the trustworthiness of naturalistic inquiries. *Educational Communitcation and Technology Journal, 29*(2).

Hackling, M. (2004). Chapter eight: Assessment in science. In G. J. Venville & V. M. Dawson (Eds.), *The Art of Teaching Science for middle and secondary school* (pp. 126-144). Crows Nest, NSW: Allen & Unwin.

Hackling, M., Peers, S., & Prain, V. (2007). Primary Connections: Reforming science teaching in Australian primary schools. *Teaching Science, 53*(3), 12-16.

Hammersley, M. (2008). *Questioning Qualitative Inquiry: Critical Essays*. Los Angeles, CA: SAGE.

Hand, B., Yore, L. D., Jagger, S., & Prain, V. (2010). Connecting research in science literacy and classroom practice: a review of science teaching journals in Australia, the UK and the United States, 1998-2008. *Studies in Science Education, 46*(1), 45-68. doi:10.1080/03057260903562342

Hargreaves, E. (2005). Assessment for Learning? Thinking outside the (black) box. *Cambridge Journal of Education, 35*(2), 213-224.

Harlen, W. (2004*).* A systematic review of the evidence of reliability and validity of assessment by teachers used for summative purposes*.* Retrieved from http://eppi.ioe.ac.uk/cms/Default.aspx?tabid=116

Harlen, W. (2005). Trusting teachers' judgement: research evidence of the reliability and validity of teachers' assessment used for summative purposes. *Research Papers in Education, 20*(3), 245-270. doi:10.1080/02671520500193744

Harlen, W., & Deakin-Crick, R. (2002). A systematic review of the impact of summative assessment and tests on students' motivation for learning. Retrieved from http://eppi.ioe.ac.uk/

Hattie, J. (2003a). *Formative and Summative Interpretations of Assessment Information*. Auckland, NZ: University of Auckland.

Hattie, J. (2003b). *Teachers Make a Difference, What is the research evidence?* ACER research conference paper. Melbourne. ACER. https://research.acer.edu.au/research_conference_2003/4

Hattie, J. (2005). *What is the nature of evidence that makes a difference to learning?* Paper presented at the Using Data to Support Learning, Grand Hyatt Hotel, Melbourne 7-9 August 2005.

Hattie, J. (2009). *Visible Learning, Tomorrow's Schools, The Mindsets that make the difference in Education*. Paper presented at the Guest Lectures by Visiting Academics, The Treasury, Wellington, NZ.

Hattie, J. (2012). *Visible Learning for Teachers: Maximising the impact on learning*. London: Routlege.

Hattie, J. (2018). Visible Learning^{plus} 252+ Influences on Student Achievement. Retrieved from https://visible-learning.org/wp-content/uploads/2018/03/VLPLUS-252-Influences-Hattie-ranking-DEC-2017.pdf

Hattie, J., Jaeger, R. M., & Bond, L. (1999). Persistent Methodological Questions in Educational Testing. In P. Asghar Iran-Nejad & D. Pearson (Eds.), *Review of Research in Education* (Vol. 24 (1)). Washington DC: AERA.

Hattie, J., & Brown, G. (2004, September). *Cognitive Processes in asTTle: The SOLO Taxonomy (asTTLe Technical Report # 43)*. Auckland: University of Auckland / NZ Ministry of Education

Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research, 77*(1), 81-112. doi:10.3102/003465430298487

Heritage, M. (2010). Formative Assessment and Next-Generation Assessment Systems: Are We Losing an Opportunity*?* Retrieved from http://www.ccsso.org/

Hickey, D. T., Taasoobshirazi, G., & Cross, D. (2012). Assessment as Learning: Enhancing Discourse, Understanding, and Achievement in Innovative Science Curricula. *Journal of Research in Science Education, 49*(10), 1240-1270.

Huber, P., Tytler, R., & Haslam, F. (2010). Teaching and Learning about Force with a Representational Focus: Pedagogy and Teacher Change. *Research in Science Educatiion, 40*, 5-28. doi:10.1007IsIII65-009-9154-9

IEA, International Association for the Evaluation of Educational Achievement. (2013). TIMSS and PIRLS Home. Retrieved from http://timss.bc.edu/

James, M. (2006). Learning how to learn, in classrooms, schools and networks. *Research Papers in Education, 21*(2), 101-234.

James, M. (2009). Assessment in Schools: Fit for purpose? Cambridge: University of Cambridge Faculty of Education.

James, M., McCormick, R., Black, P., Drummond, M.-J., Fox, A., MacBeath, J., . . . Wiliam, D. (2007). *Improving Learning How to Learn: Classrooms, schools and networks*. London: Routledge.

JFF, Jobs for the Future. (2007). The STEM workforce challenge: the role of the public workforce system in a national solution for a competitive, science, technology, engineering, and mathematics (STEM) workforce. Retrieved from https://www.doleta.gov/youth_services/pdf/STEM_Report_4%2007.pdf

Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Towards a Definition of Mixed Methods Research. *Journal of Mixed Methods Research, 1*(2), 112-133. doi:10.1177/1558689806298224

Jones, A., & Buntting, C. (2013). International, National and Classroom Assessment: Potent Factors in Shaping What Counts in School Science. In D. Corrigan, R.

Gunstone, & A. Jones (Eds.), *Valuing Assessment in Science Education: Pedagogy, Curriculum, Policy*. (pp 33-550). Dordrecht: Springer.

Klenowski, V., & Wyatt-Smith, C. (2012). The impact of high stakes testing: the Australian story. *Assessment in Education: Principles, Policy & Practice, 19*(1), 65-79. doi:10.1080/0969594X.2011.592972

Laerd Statistics. (2013). Multiple Regression Analysis using SPSS Statistics. Retrieved from https://statistics.laerd.com/spss-tutorials/multiple-regression-using-spss-statistics.php

Laerd Statistics. (2017). Testing for Normality using SPSS Statistics. Retrieved from https://statistics.laerd.com/spss-tutorials/testing-for-normality-using-spss-statistics.php

Laerd Statistics. (2018). One way ANOVA using SPSS Statistics. Retrieved from https://statistics.laerd.com/spss-tutorials/one-way-anova-using-spss-statistics.php#procedure

Lane, D. M. (n.d.). *ONline Statistics Education: A Multimedia Course of Study.* D. M. Lane (Ed.) (pp. 692). Retrieved from http://onlinestatbook.com/2/index.html

Lemke, J. L. (2001). Articulating Communities: Sociocultural Perspectives on Science Education. *Journal of Research in Science Teaching, 38*(3), 296 - 316.

Lim, X.-S., Tan Eng Thye, J., & Kang Lu-Ming, T. (2009). *Avoiding the "prolonged agony" of studying for standardized national exams: At what price?* Paper presented at the AARE 2009 conference, Canberra. http://www.aare.edu.au/09pap/abs09.htm

Lyons, T., & Quinn, F. (2010). Choosing Science: Understanding the declines in senior high school science enrolments. Retrieved from https://simerr.une.edu.au/pages/projects/131choosingscience.pdf

Lyons, T., & Quinn, F. (2012). Rural High School Students' Attitudes Toward School Science. *Australian and International Journal of Rural Education, 22*(2), 8.

Lyons, T., & Quinn, F. (2014). How Relevant Are Australian Science Curricula for Rural and Remote Students? *Australian and International Journal of Rural Education, 24*(2), 8.

Mansell, W., James, M., & The Assessment Reform Group. (2009). *Assessment in schools. Fit for purpose? A Commentary by the Teaching and Learning Research Programme.* London: Economic and Social Research Council, Teaching and Learning Research Progamme.

Marzano, R. J. (2000). What are Grades For? *Transforming Classroom Grading* (pp. 10). Denver, Colorado: Mid-continent Research for Education and Learning (McREL).

Marton, K., & Säljö, R. (1976). *On qualitative differences in learning: 1—outcomes and process.* British Journal of Educational psychology, 46, 8.

Masters, G. N. (2009). *Assessing science learning (PAT science tests).* Melbourne, VIC: ACER.

Masters, G. N. (2013). Reforming Educational Assessment: Imperatives, principles and challenges. In S. Mellor (Series Ed.) *Australian Education Review*. Retrieved from http://research.acer.edu.au/aer/12

Matters, G., & Curtis, D. (2008). *A study into the assessment and reporting of employability skills of senior secondary students*. Retrieved from https://research.acer.edu.au/cgi/viewcontent.cgi?article=1000&context=ar_misc

MCEETYA, Ministerial Committee of Education, Employment. Training and Youth Affairs. (1998). *A Review of the 1989 Common and Agreed Goals for Schooling in Australia (The 'Hobart Declaration').* Carlton South, Victoria: MCEETYA Secretariat.

MCEETYA, Ministerial Council of Education, Employment, Training and Youth Affairs. (2008). *Melbourne Declaration on Educational Goals for Young Australians*. Melbourne, Victoria: Curriculum Corporation.

Merriam, S. B. (1998). *Qualitative research and case study applications in education*. San Francisco: Jossey-Bass.

Messick, S. (1995). Validity of Psychological Assessment: Validation of Inferences From Persons' Responses and Performances as Scientific Inquiry Into Score Meaning. *American Psychologist, 50*(9), 741-749.

Millar, R., & Hames, V. (2003). Towards Evidence-based Practice in Science Education 1- 4. In *Teaching and Learning Research Programme TLRP* (Ed.). York, UK: University of York.

Millar, R. (2013). Improving Science Education: Why Assessment Matters. In D. Corrigan, R. Gunstone, & A. Jones (Eds.), *Valuing Assessment in Science Education: Pedagogy, Curriculum, Policy* (pp 55-68). Dordrecht: Springer.

Mislevy, R. J. (2008). How Cognitive Science Challenges the Educational Measurement Tradition, 18. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.471.4656&rep=rep1&type=pdf

Mitchell, I., Mitchell, J., & Lumb, D. (2009). *Principles of Teaching for Effective Learning: The Voice of the Teacher.* Clayton, VIC: PEEL Publishing.

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff, C. (2009). *TIMSS 2011 Assessment Frameworks*. USA: TIMSS & PIRLS International Study Centre, Lynch School of Education, Boston College.

NAEP, National Assessment Governing Board. (2011). *Science Framework for the 2011 National Assessment of Educational Progress*. Washington, DC: US Government Printing Office.

NESA, NSW Education Standards Authority. (n.d.). *Home*. Retrieved from https://www.esa.edu.au/solutions/our-solutions

NESA, NSW Education Standards Authority. (2017). *Statistics Archive--Stage 5 results (Archived Year 10 Grade reports)*. Retrieved from http://www.boardofstudies.nsw.edu.au/ebos/static/ebos_stats.html

NESA, NSW Education Standards Authority. (2018). Assessment for, as and of Learning. Retrieved from http://syllabus.nesa.nsw.edu.au/support-materials/assessment-for-as-and-of-learning/

Newton, P. (2007). Clarifying the purposes of educational assessment. *Assessment in Education, 14*(2), 149. doi:10.1080/09695940701478321

Newton, P. (2010). The Multiple Purposes of Assessment. In P. McGaw, E. Peterson, & B. Baker (Eds.), *International Encyclopedia of Education (Third Edition)* (pp. 392-396). Oxford: Elsevier.

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. *Studies in Higher Education, 31*(2), 199-218. doi:10.1080/03075070600572090

NRC, National Research Council. (1996). National Science Education Standards (NSES) (pp. 273). Retrieved from https://www.csun.edu/science/ref/curriculum/reforms/nses/nses-complete.pdf

NRC, National Research Council. (2001). Knowing What Students Know: The Science and Design of Educational Assessment. Retrieved from http://www.nap.edu/catalog.php?record_id=10019

NRC, National Research Council. (2008). Research on Future Skill Demands: A Workshop Summary. Retrieved from https://www.nap.edu/catalog/12066/research-on-future-skill-demands-a-workshop-summary

NSW D of E, NSW Department of Education. (2013). Policies: Curriculum planning and programming, assessing and reporting to parents K-12. Retrieved from https://www.det.nsw.edu.au/policies/curriculum/schools/curric_plan/PD20050290.shtml?query=Curriculum+planning+and+programming%2c+assessing+and+reporting+to+parents+K-12

NSW D of E, NSW Department of Education. (2017). Class Size. Retrieved from: https://education.nsw.gov.au/about-us/our-people-and-structure/history-of-government-schools/facts-and-figures/class-size

NSW D of E, NSW Department of Education. (2018). VALID program. Retrieved from https://education.nsw.gov.au/teaching-and-learning/student-assessment/assessment-and-reporting/assessment/valid

Nuffield Foundation. (2018). The Assessment Reform Group (ARG). Retrieved from http://www.nuffieldfoundation.org/assessment-reform-group

Nusche, D., Radinger, T., Santiago, P. (co-ordinator), & Shewbridge, C. (2013). *Synergies for Better Learning: AN INTERNATIONAL PERSPECTIVE ON EVALUATION AND ASSESSMENT*. Paris: OECD. Retrieved from http://www.oecd.org/education/school/synergies-for-better-learning.htm

OCS, Office of the Chief Scientist. (2014). Science, Technology, Engineering and Mathematics: Australia's Future. Retrieved from http://www.chiefscientist.gov.au/2014/09/professor-chubb-releases-science-technology-engineering-and-mathematics-australias-future/

OCS, Office of the Chief Scientist. (2017). Science and maths in Australian secondary schools datasheet. Retrieved from http://www.chiefscientist.gov.au/2016/07/science-and-maths-in-australian-secondary-schools-datasheet/

OECD, Organisation for Economic Co-operation and Development. (1996). The Knowledge-based Economy (pp. 46). Paris: OECD. Retrieved from http://www.oecd.org/sti/sci-tech/1913021.pdf

OECD, Organisation for Economic Co-operation and Development. (1997). Thematic Review of the Transition from Initial Education to Working Life: Australia, Country Note. Paris: OECD. Retrieved from http://www.oecd.org/edu/innovation-education/1908315.pdf

OECD, Organisation for Economic Co-operation and Development. (2003). Key Competencies for a Successful Life and a Well-Functioning Society. Paris: OECD. Retrieved from http://deseco.ch/bfs/deseco/en/index/02.html

OECD, Organisation for Economic Co-operation and Development. (2011). OECD Reviews of Evaluation and Assessment in Education: Australia. Paris: OECD. Retrieved from https://www.oecd.org/australia/48519807.pdf

OECD, Organisation for Economic Co-operation and Development. (2014). About PISA. Retrieved from http://www.oecd.org/pisa/aboutpisa.htm

OECD, Organisation for Economic Co-operation and Development. (2017). PISA 2015: Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving, revised edition, PISA. Paris: OECD. Retrieved from https://www.mecd.gob.es/dctm/inee/internacional/pisa-2015-frameworks.pdf?documentId=0901e72b820fee48 doi:10.1787/9789264281820-en

OECD, Organisation for Economic Co-operation and Development. (2018). How's Life? Measuring Well-being. Retrieved from http://www.oecdbetterlifeindex.org/

Osborne, J., & Dillon, J. (2008). Critical Reflections: A Report to the Nuffield Foundation. Retrieved from http://efepereth.wdfiles.com/local--files/science-education/Sci_Ed_in_Europe_Report_Final.pdf

Palmer, T.-A. (2015). *Fresh Minds for Science: Using marketing science to help school science.* (Doctor of Philosophy PhD), University of Technology Sydney, Sydney. Retrieved from http://hdl.handle.net/10453/37019

Panizzon, D. (2003). Using a cognitive structural model to provide new insights into students' understandings of diffusion. *International Journal of Science Education, 25*(12), 1427-1450. doi:10.1080/09500690320000052108

Panizzon, D., Arthur, D., & Pegg, J. (2006). Essential Secondary Science Assessment: Development and scope of a test to explore scientific literacy and achievement in NSW. *Teaching Science, 52*(4), 6.

Panizzon, D., & Bond, T. (2006). Exploring conceptual understandings of diffusion and osmosis by senior high school and undergraduate university science students. In X. Liu & W. J. Boone (Eds.), *Applications of rasch measurement in science education* (pp. 137-164). Maple Grove, MN: Jam Press.

Panizzon, D., & Bond, T. (2007). *Measuring scientific understanding: A pedagogical problem and its potential solution?* Paper presented at the AARE Conference, Fremantle, WA.

Panizzon, D., Callingham, R., Wright, T., & Pegg, J. (2007). *Shifting Sands: Using SOLO to promote assessment for learning with secondary mathematics and science teachers*. Paper presented at the AARE Conference, Fremantle, WA.

PEEL, Project for Enhancing Effective Learning. (2009). About PEEL. Retrieved from http://www.peelweb.org/index.cfm?resource=about

Pegg, J., Panizzon, D., Arthur, D., Scott, J., & Aylmer, W. (2011). *Assessing student responses in secondary science: Examples from ESSA*. Unpublished manuscript, SiMMER Centre, University of New England. Armidale, NSW.

Pellegrino, J. W. (2009). *The Design of an Assessment System for the Race to the Top: A Learning Science Perspective on Issues of Growth and Measurement*. Paper presented at the Exploratory Seminar: Measurement Challenges Within the Race to the Top Agenda. Retrieved from: http://www.k12center.org/publications.html.

Pierce, R., & Chick, H. (2011). Teachers' intentions to use national literacy and numeracy assessment data: a pilot study. *Australian Educational Researcher, 38*, 433-447.

Polesel, J., Dulfer, N., & Turnbull, M. (2012). *The Experience of Education: The impacts of high stakes testing on school students and their families. Literature Review*. Rydalmere: The Whitlam Institute.

QERC. (1985). *Quality of Edcuation in Australia: Report of the Review Committee*. Canberra, ACT: Australian Government Publishing Service.

QSA, Queensland Studies Authority*. (2012). Memo: Future of QCATS.* Retrieved from https://www.qsa.qld.edu.au/qsa_secure/memos.act?year=2012&type=MEMO&docType=MEMO&orderBy=audience

Rowe, K. J., & Hill, P. W. (1996). Assessing, Recording and Reporting Students' Educational Progress: the case for 'subject profiles'. *Assessment in Education: Principles, Policy & Practice, 3*(3), 309-352. doi:10.1080/0969594960030304

Rowe, K. J. (2006). School Performance: Australian State/Territory Comparisons of Student Achievements in National and International Studies. Carlton, Victoria: ACER. Retrieved from http:research.acer.edu.au/learning_processes/5

Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L., & Klein, S. (2002). On the Evaluation of Systemic Science Education Reform: Searching for Instructional Sensitivity. *Journal of Research in Science Teaching, 39*(5), 369-393.

Ruiz-Primo, M. A. (2009). *Towards a Framework for Assessing 21st Century Science Skills: Commissioned paper for the National Academies*. Denver: University of Colorado Denver. Retrieved from https://sites.nationalacademies.org/cs/groups/dbassesite/documents/webpage/dbasse_072612.pdf

Ruiz-Primo, M. A., & Li, M. (2012). Examining Formative Feedback in the Classroom Context: New Research Perspectives. SAGE Handbook of Research on Classroom Assessment. SAGE Publications, Inc. In J. H. McMillan (Ed.). Thousand Oaks, CA: SAGE Publications, Inc.

Ryan, C. (1997). *NSW Key Competencies Pilot Project Report.* Sydney: NSW Department of Education Co-ordination.

Sadler, R. (2007). Perils in the meticulous specification of goals and assessment criteria. *Assessment in Education, 14*(3), 387-392.

Sadler, R. (1998). Formative Assessment: revisiting the territory. *Assessment in Education: Principles, Policy & Practice, 5*(1), 77-84. doi:10.1080/0969595980050104

Sandelowski, M. (2000). Whatever Happened to Qualitative Description? *Research in Nursing & Health*, 23, 334-340.

Schraw, G., Crippen, K. J., & Hartley, K. (2006). Promoting Self-Regulation in Science Education: Metacognition as Part of a Broader Perspective on Learning. *Research in Science Education, 36*, 111-139. doi:10.1007/s11165-005-3917-8

Schroder, H., Driver, M., & Streufert, S. (1967). *Human information processing*. New York: Holt, Rinehart, & Winston.

SCSA, WA Schools Curriculum and Standards Authority. (2010). Western Australian Monitoring Standards in Education (WAMSE): Science tests. Retrieved from http://www.scsa.wa.edu.au/internet/Years_K10/WAMSE/Tests/Science

Sfard, A. (1998). On Two Metaphors for Learning and the Dangers of Choosing Just One. *Educational Researcher, 27*(4), 10. doi:10.3102/0013189X027002004

Shayer, M. (1976). Development in thinking of middle school and early secondary school pupils. *School Science Review, 57*.

Shayer, M. (2003). Not just Piaget; not just Vygotsky, and certainly not Vygotsky as alternative to Piaget. *Learning and Instruction, 13*, 465-485. doi:10.1016/S0959-4752(03)00092-6

Shenton, A. K. (2004). Strategies for ensuring trustworthiness in qualitative research projects. *Education for Information*(22), 63-75.

Shepard, L. A. (1993). Evaluating Test Validity. In L. Darling-Hammond (Ed.), *Review of Research in Education* (Vol. 19 (Issue 1), pp. 46). Washington DC: AERA.

Shepard, L. A. (2001). The Role of Classroom Assessment in Teaching and Learning. In V. Richardson (Ed.), *Handbook of Research on Teaching* (pp. 1278). Washington DC: American Educational Research Association.

Shute, V. J. (2007). *Focus on Formative Feedback*. Princeton, NJ: Educational Testing Service.

Sjøberg, S., & Schreiner, C. (2010). The ROSE project: An overview and key findings. Retrieved from http://roseproject.no./publications/english-pub.html

Smith, M. (2005). *Data for schools in NSW: What is provided and can it help?* Paper presented at the Using Data to Support Learning, Grand Hyatt Hotel, Melbourne 7-9 August 2005.

SCCS&F, Standing Committee of Communities, Schools and Families. (2008). The Purposes of Testing and Fitness for Purpose (Third Report). Retrieved from www.educationengland.org.uk/documents/pdfs/2008-testing-and-assessment.pdf

Stake, R. E. (1995). *The Art of Case Study Research*. Thousand Oaks: Sage.

Stiggins, R. J. (2002). Assessment Crisis: the Absence of Assessment FOR Learning. *Phi Delta Kappan, 83*(10), 758-765.

Stiggins, R. J. (2004). Student-Centred Classroom Assessment. Retrieved from 169.204.228.86

Stiggins, R. J. (2007). Assessment Through the Student's Eyes. *Educational Leadership, 64*(8), 22-26.

Stiggins, R. J., & Chappius, J. (2005). Using Student-Involved Classroom Assessment to Close Achievement Gaps. *Theory into Practice, 44*(1), 11-18.

Stiggins, R. J., & DuFour, R. (2009). Maximizing the Power of Formative Assessments. *Phi Delta Kappan, 90*(9), 640-644. doi:10.1177/003172170909000907

Tebbutt, Hon Carmel. (2005). Science Students Assessment (page 14956). Retrieved from http://www.parliament.nsw.gov.au/prod/PARLMENT/hansArt.nsf/0/31FF954D026D2F48CA256FE5007AE34C

Thomson, S., De Bortoli, L., & Underwood, C. (2017). PISA 2015: Reporting Australia's results. Retrieved from http://www.acer.org/ozpisa/reports/

Thomson, S., Hillman, K., & De Bortoli, L. (2013). A teacher's guide to PISA scientific literacy (pp. 58). Retrieved from http://www.acer.edu.au/ozpisa/reports/

Thomson, S., Wernert, N., O'Grady, E., & Rodrigues, S. (2017). TIMSS 2015: Reporting Australia's results. Retrieved from https://research.acer.edu.au/timss_2015/2/

Tobin, K. (2012). Sociocultural Perspectives on Science Education. In B. J. Fraser, K. Tobin, & C. J. McRobbie (Eds.), *Second International Handbook of Science Education* (pp. 3-17). Dordrecht: Springer.

Torrance, H. (2007). Assessment *as* learning? How the use of explicit learning objectives, assessment criteria and feedback in post-secondary education and training can come to dominate learning. *Assessment in Education, 14*(3), 281-294.

Treagust, D. F. (2006). *Diagnostic assessment in science as a means of improving teaching, learning and retention*. Paper presented at the Assessment in Science Teaching and Learning Symposium, The University of Sydney. http://science.uniserve.edu.au/pubs/procs/2006/index.html

Tytler, R. (2007). *Re-imagining science education: engaging students in science for Australia's future*. Camberwell, Vic.: ACER Press.

Tytler, R., & Hubber, P. (2010). *A representation-intensive signature pedagogy for school science?* Paper presented at the AARE Annual Conference, Melbourne.

Tytler, R., & Prain, V. (2010). A Framework for Re-thinking Learning in Science from Recent Cognitive Science Perspectives. *International Journal of Science Education, 32*(15), 2055-2078.

Tytler, R., Prain, V., Huber, P., & Waldrip, B. (Eds.). (2013). *Constructing Representations to Learn in Science*: SensePublishers.

Tytler, R., & Symington, D. (2015). Science Learning in Rural Australia. Not necessarily the poor cousin. *Teaching Science, 61*(3), 7.

UNDP, United Nations Development Program. (2018). Human Development Index. Retrieved from http://hdr.undp.org/en/content/human-development-index-hdi-table

UNESCO, United Nations Educational, Scientific and Cultural Organisation. (2005). *UNESCO World Report: Towards Knowledge Societies*. Paris: UNESCO Publishing.

Vygotsky, L. S. (1978). *Mind in society*. London: Harvard University Press.

Waldrip, B., Prain, V., & Carolan, J. (2010). Using Multi-Modal Representations to Improve Learning in Junior Secondary Science. *Research in Science Education, 40*, 60-80. doi:10.1007/sI1165-009-9157-6

Wasson, D. (2009). *Large cohort testing--How can we use assessment data to effect school and system improvement*. Paper presented at the 2009 ACER research conference, Perth. Conference theme—Assessment and Student Learning: Collecting, Interpreting and Using Data to Inform Teaching.

Webb, N. L. (1997). Research Monograph No. 6: Criteria for Alignment of Expectations and Assessments in Mathematics and Science Education ED 414 305*.* Retrieved from http://eric.ed.gov/?id=ED414305

White, R., & Gunstone, R. (1992). *Probing Understanding*. London: The Falmer Press.

Wiggins, G. P. (1998). *Educative assessment: designing assessments to inform and improve student performance*: Jossey-Bass.

Wiliam, D. (2011a). *Embedded formative assessment*. Bloomington, IN 47404: Solution Tree Press.

Wiliam, D. (2011b). What is assessment for learning? *Studies in Educational Evaluation, 37*, 3-14.

Wilson, M., & Sloane, K. (2000). From Principles to Practice: An Embedded Assessment System. *Applied Measurement in Education, 13*(2), 181-208

Yin, R. K. (2003). *Case Study Research 3rd edition* (Vol. 5). Thousand Oaks: SAGE.