An investigation into the challenges faced by international school teachers when interpreting the results from standardised tests

PATRICIA JANE POMROY

Abstract

An investigation into the challenges faced by international school teachers when interpreting the results from standardised tests.

Patricia Pomroy

This study aims to identify factors that could affect the validity of the inferences made when large-scale standardised tests are used in international schools. These tests are exported in that they are administered to a population which is different to the one for which they were designed.

Document analysis identified some areas where construct-irrelevant variance might occur. The transitive nature of international school populations can lead to disruptions in student learning trajectories. There are potential differences in the curriculum that is used to develop the test and those used in international schools. International school populations are also linguistically and culturally diverse.

Interviews conducted with a small number of international school teachers found that these teachers can lack the skills necessary to interpret the score reports that provide the feedback from testing. It is the test user who bears the responsibility for establishing the validity of any inferences made from exported tests. The lack of skills of such educators raises concerns about their ability to assess the validity of inferences that can be made from these tests.

The study contributes to research on the use of standardised on linguistically and culturally diverse populations as well as the ability of teachers to interpret information from testing data.

An investigation into the challenges faced by international school teachers when interpreting the results from standardised tests

Patricia Pomroy

A thesis submitted in fulfilment of the requirement of the degree of Doctor of Education

School of Education

Durham University

2020

Table of Contents

List of tables

Abbreviations

ACER	Australian Council for Educational Research
ACT	American College Test
AERA	American Education Research Association
AP	Advance Placement
APA	American Psychological Association
CAT	Computer Adaptive Test
CIS	Council of International Schools
DDDM	Data-driven decision making
DIF	Differential Item Functioning
EBRW	Evidence-Based Reading and Writing
ELL	English Language Learner
GCSE	General Certificate of Education
IBDP	International Baccalaureate Diploma Program
ISA	International Schools Assessment
IUA	Interpretation/Use Argument
MAP	Measures of Academic Progress
MYP	Middle Years Program
NAEP	National Assessment of Progress
NCLB	No Child Left Behind
NCME	National Council of Measurement in Education
OECD	Organisation for Economic Co-operation
PISA	Programme for International Student Assessment
PYP	Primary Years Program
TIMMS	Trends in International Mathematics and Science Study
US	United States of America

Declaration

No material contained in the thesis has previously been submitted for a degree in this or any other institution.

Statement of Copyright

Acknowledgements

I would like to express my gratitude to my current supervisors, Professor Jens Beckmann and Dr. Dimitra Kokotsaki. I am grateful for the help and support that you have given me during the writing-up phase of my thesis. The constructive feedback has really helped me to improve the standard of my writing. I would also like to thank my original supervisor, Professor Robert Coe for his input in helping me to make decisions during the research phase of this work.

I also want to thank all the friends who have supported me and consoled me during this journey. I particularly want to thank Professor Bernard Wilson for being willing to proofread my early work. I also want to thank Professor Brendan Wilson and Mary Saso. There have been times when both of you have offered words of encouragement that have made a real difference to me, and I appreciated your comments. There are so many people who have wished me well in my endeavours and kept me going on the days when I have felt discouraged that I am too scared to mention names for fear of offending those I miss. Thank you to everybody who has offered words of encouragement as I have worked through this project.

Finally, I would like to thank my family. Mum and Dad, you have always encouraged me in everything I have tried to do in my life. You raised me to accept challenges and not to give up even when the going gets tough. Thank you for always supporting my decisions, no matter how crazy they are. To Patrick and Angela, thank you for all the encouragement you have given me along the way.

Chapter 1

Introduction

This research aims to describe the use of large-scale standardised testing in international schools. For that purpose, the technical manuals of assessment tools used in those schools will be reviewed to establish the intended purpose(s) of those tests and to reflect on the kind of inferences that can validly be drawn from the obtained test scores. The insights gained will then be projected onto the specific characteristics of international schools and their needs in terms of utilising test information. Interviews were also conducted to ask a small number of teachers from international schools to describe the challenges that they experience when using the feedback from large-scale standardised testing.

One of the fundamental aims of education is to facilitate and improve student learning. In order to do this, it is necessary to assess where students are in their learning journey and to adapt teaching strategies to help them make the next steps. As explained by Earl (2005), this has been traditionally left to the experience and professional judgement of teachers. However, there has been a recent shift towards the use of more evidence-based teaching in which data from assessments are analysed and used to inform practice. According to Firestone & González (2007), data from large-scale standardised testing can provide an important source of information and this is expected to facilitate schools' educational and pedagogical decision making. For instance, a report by the European Comission (2009) acknowledges that results from large-scale standardised tests are used by schools to identify the learning needs of their students with a view to adapting their teaching.

1.1 Standardised Tests

The use of large-scale standardised testing has become more prevalent in recent years. It was noted by Phelps (2000) that in a sample of developed countries including Russia, China and those in the Organisation for Economic Co-operation (OECD), there was evidence that during the 25 years in the run up to the millennium, twenty-seven countries and provinces had increased the amount of testing they carried out while only three had shown evidence of a reduction in testing. The report by the European Comission (2009) states that 21 European countries had introduced or had been scheduled to introduce standardised testing between 1990 and 2012. Further there had been changes or additions to the tests in five countries that were seen as early adopters of large-scale standardised testing. They comment that nearly all European countries now have some form of national standardised testing. Additionally, in an analysis of the data collected from 15 year-old students who participated in the Programme for International Student Assessment (PISA) in 2009, Schleicher (2015) reported that 76% of students indicated that they take at least one standardised test a year, while 34% of Dutch students stated that they take a standardised test at least once a month.

Standardised tests have been developed to fulfil many purposes in education. One purpose is to enable comparison across countries. For instance, tests such as the PISA and the Trends in International Mathematics and Science Study (TIMSS) are promoted as providing a way for evaluating and comparing the quality and efficiency of school systems. This results in the identification of the characteristics of high-performing countries which allows for other countries to adapt their systems with the aim of increasing their performance. However, at the school level,

comparisons are made between individuals and groups of students within different subject areas and across different schools. The European Comission (2009) reports that the increase in national tests came about due to the decentralisation of education policy in many countries in Europe which resulted in more autonomy for schools. They describe national tests as those that are administered nationally to the public education sector by the central organisations that are responsible for the country's education policy. These decentralisation policies also saw the introduction of new evaluation requirements which have a variety of purposes including the selection or streaming of students and for certification at the end of schooling. Tests are also used to inform decisions regarding the student's readiness for progression into the next level education, whether that be moving to the next grade within the school or moving onto the next strata of the education system. It is suggested by Smith (2014) that much of the increase in standardised testing is due to the introduction of accountability systems in education. Accountability in an education context is defined by Figlio & Loeb (2011) as "the process of evaluating school performance on the basis of student performance measures." (p. 384). As will be discussed further in chapter 2, the underlying assumption and subsequent expectation is that by implementing programs in which students, teachers and others involved in educational provision are made responsible for the results from standardised tests, attainment will improve.

Standardised tests are designed in such a way that all test takers should be subject to the same experiences in that they complete the same tasks which are administered under the same conditions and then marked using the same scoring rules. By standardising the conditions under which the test is conducted sources of potentially

irrelevant variability in measured performance is minimised so that results are comparable across test takers (Kane, 2016).

Tests are designed for different purposes. Wiliam (2010) points out that test classifications are not based on the tests themselves but rather the types of inferences that are to be made from the results. For example, according to Reynolds, Livingston, & Willson (2010) standardised tests can be classified into aptitude or achievement tests. Aptitude tests tend to be associated with norm-referenced inferences while achievement test mainly use criterion-referencing. The roots of norm-referencing are in psychometric theory where inferences are based on determining how the performance of an individual student compares to others who are similar in some way, say in terms of age or grade level. According to Willis, Dumont, & Kaufman (2012), psychometric theory rests on the assumption that people possess certain differing abilities in areas such as reading comprehension and that the level of ability can be inferred from observable, and subsequently measurable behaviours. A norm-referenced standardised test is constructed using a defined set of items that allows the registration and scoring of behaviours (responses to those items) that is considered indictive of the ability, skill, or knowledge level targeted. To obtain a valid norm reference the test is administered to a representative sample of individuals from the population of interest, for instance, all the students of a given age in a country. Their results are used as a yardstick to interpret the relative performance of other individuals from the same population when they take the test. With norm-referenced testing, percentile ranks are a prominent format of score reporting. These tests are designed to measure the test taker's current level of

proficiency by assessing the cognitive skills that have been acquired up until the point of testing.

While norm-refencing may be useful if the requirement is to relate student skill levels to the population of interest, sometimes the requirement is to classify students' performance in reference to a defined list of externally determined criteria. This requires a different form of testing. Achievement tests are often criterion-referenced in that they are linked to specific learning objectives and are designed to test an individual on content in which they have received instruction. Criterion-referencing starts with a group of subject matter experts producing a description of the knowledge and skills which students are expected to have mastered at a specific point in their education. The aim of testing is to evaluate how well the student has learnt that required knowledge. Criterion-referencing may be used to identify if a student had reached a pre-determined level of understanding or mastery, in which case the results may be reported in terms of absolute grades such as pass or fail or categories such as proficient, and not proficient. While these labels are easily understood, the score ranges within the label can be wide. Therefore, as Gulek, (2003) tells us, students who appear in the same performance level category can vary widely in their attainment depending on whether they have gained scores nearer the lower or the upper boundary of the level. It is also difficult to classify students with certainty as testing only gives an estimate for performance because testing can only include a sample of the concepts from the content domain. To acknowledge this measurement error, test scores are sometimes shown with confidence intervals that acknowledges this measurement error. While this band may include the score

needed to pass or to reach an appropriate level, the actual score that is attributed to the student may mean that they are placed in a higher or lower category.

Problems in using test information occur when, for instance, norm-referenced tests are used for criterion-referenced purposes. For example, Gipps (2012) quotes examples of norm-referencing being used in tests to determine what level of mastery the average 7 year-old will have so that this can inform decisions about where grade boundaries should be applied. She concludes that using the results of criterion-referenced testing for norm-referenced purposes causes "enormous problems" (p. 7). To avoid this sort of misuse of testing, it is important that test users understand the types of testing that they are using, the inferences that can be justified from those tests and the limitations that given test formats have.

1.1.1 Feedback from large-scale standardised testing

After students have participated in large-scale standardised tests, results are generally returned to test users such as schools, students and parents in the form of score reports (See appendix 1). Such reports are considered by Roduta, Roberts & Gotch, (2019) to be the primary source of information given to the test user. According to Zenisky & Hambleton (2012) these reports have historically been used to relate details about scaled scores and performance bands while Hattie (2014) tells us that trained professionals such as psychologists were considered to be the traditional audiences that these score reports were designed for. Plake & Wise (2014) and Hattie (2014) comment that as a result of the introduction of accountability measures, there has been a change to the audience who use these reports. It is now more likely that teachers and educational administrators will be

expected to interpret the information that is given in these reports. However, it is unlikely that teachers will have received training to understand such reporting. Popham (2018) tells us that these reports are designed to give guidance to test users on how to make score-based interpretations from the test-taker's performance. However, without the appropriate education, teachers may not be able to understand the information they are given.

Accountability measures have also led to an expansion in the information included in score reports to include diagnostic information, according to Zenisky & Hambleton (2012). There are increasing expectations that this diagnostic information will be used as a means for teachers and school administrators to inform curriculum practices by identifying areas in which individuals or groups of students have gained mastery and pinpointing areas where there are weaknesses. For instance, Mandinach & Honey (2008) argue that by using this information in their lesson planning, teachers should be able to increase the attainment of all students. They also state that by looking at assessment data over time, school leaders should be able to identify where gaps occur in the curriculum that is taught. This information can be used to identify where there is a need to invest resources, for example, to implement new learning programs or provide targeted professional development on new learning strategies, so that performance can be improved. It is acknowledged by Plake & Wise (2014) that many of the test users who are now expected to use the information provided in score reports have not received training in educational measurement. The presumed lack of ability of teachers and administrators to interpret the information that is provided in score reports is a major cause for

concern as it may lead to teachers and school leaders making invalid interpretations with detrimental consequences.

As will be discussed in greater detail in chapter 2, there is extensive evidence of barriers that prevent teachers from being able to gain insights from the provided information. For instance, Roduta Roberts, Gotch, & Lester, (2018), comment that score reports can use text, graphs and numbers to describe student performance in the test but research by Brown & Hattie (2011) shows that teachers cannot understand the statistical diagrams that are used in score reports. The way the data is presented can also have an impact on its use. Firestone & González (2007) tell us that if presentation methods are confusing or if reports are too densely packed, this may be intimidating for teachers and result in the data being ignored. However, Mandinach & Schildkamp (2020) warn that information may be rendered useless if it is oversimplified. Even where teachers are able to understand what is presented, they may struggle to translate this information into appropriate instructional interventions. As Hattie (2014) acknowledges, no matter how carefully tests are designed, this cannot prevent the harmful effects of interventions based on poor interpretations. He states that it is essential for those interpreting testing data to have received appropriate training. However, Popham (2009) acknowledges that many serving teachers would not have undertaken appropriate courses during their teacher training. Moreover, he comments that teachers report that they have not undertaken any professional training in how to interpret such data. Further, Mandinach & Jimerson (2016) report a lack of funding for professional development to help teachers to develop the skills that are needed to interpret the data correctly.

The use of standardised testing is not without controversy and there is extensive discussion surrounding the appropriateness of standardised testing for specific uses and the validity of the inferences that are made from the data that is produced. Standardised tests are developed for specific purposes and there are elaborate and extensive procedures that the test has to undergo during its development to show its fitness for those purposes. Possible inferences (i.e. valid) are identified as part of the test development procedure and the evidence that is collected during the development process that warrants these claims. However, standardised tests are frequently used for purposes such as accountability in ways that go beyond those for which validity has been established and therefore there is no evidence to justify the given use of the test.

Particular concerns are raised when these inferences result in unwarranted and unacceptable consequences. Fear of potential negative consequences can lead to the use of unsound educational practices which also have implications for the validity of the inferences that are being made. For instance, pressures to show that the school is achieving acceptable levels of attainment can result in teaching to the test and test score inflation which will affect the validity of inferences that can be made from the data obtained because improvements in the test score may not be attributable to improvements in understanding of the domain that is being tested (Means, Chen, DeBarger, & Padilla, 2011). Because of a lack of understanding of the purposes that tests are validated for, users such as the bodies responsible for educational provision, educational professionals and parents see the results of large-scale testing as a measure of quality of educational provision but the tests lack the instructional

sensitivity that would allow them to determine how well students are taught and are therefore not suitable for this purpose.

1.2 Validity

When establishing the validity of a test, it is the proposed use of the test score that needs to be validated rather than the test itself (Nichols & Williams, 2009). Validity is not an absolute concept and tests cannot be declared to be either valid or invalid per se. Establishing the validity of a test involves evaluating the degree to which the evidence from the test warrants the interpretations that are being made. There must be adequate evidence for each inference made and it is therefore essential that teachers are able to understand applications of testing and read score reports correctly so that the likelihood of incorrect interpretation is minimised.

When developing tests, measurements experts focus on ensuring the quality of the data that is produced and they provide documentation detailing the procedures carried out in developing the test. Kane (2001) states that when tests are developed, some of their uses can be anticipated and that the test developer should include information that evaluates the suitability of the test for those particular types of inference while Brennan (2013) asserts that their responsibility should extend to giving warnings about the limitations of their test and advice about anticipated incorrect uses. However, test use is a complex process and test developers may not be able to anticipate all the ways in which an educator may use a test score within their own particular context. This places the onus on users of score reports to be aware of the fundamental issues that can affect assessment results. A standardised assessment is only an evaluation of how a student performed at a particular moment

in time and users need to be aware that all evaluation processes are subject to error. The users of test scores have the responsibility to gain understanding of such issues and cannot just be passive users of score reports. If teachers are unable to use testing appropriately or to read and interpret the score reports correctly then the validity of any inferences made is threatened.

1.2.1 Potential threats to validity

Kane (2006) discusses two potential threats to validity. These are, construct underrepresentation and construct irrelevant variance (Messick, 1989). Construct underrepresentation occurs when the range of observations fail to represent the range of processes associated with a trait. He defines a trait as "a disposition to behave or perform in some way in response to some kinds of stimuli or tasks under some range of circumstances" (Kane, 2006, p. 30). When sampling from a domain, underrepresentation will always occur and should be considered in the extrapolation inference, which should guarantee that underrepresentation does not have a significant effect on the score obtained. However, serious underrepresentation occurring due to "restrictions in the universe of generalisation relative to the target domain", (Kane, 2006, p. 38) is more problematic when the target score is substantially impacted. Construct-irrelevant variance occurs when factors that are not associated with the trait influence the measures of that trait. Kane (2006) comments that construct-irrelevant variance is not just a threat to validity as a whole but can compromise the validity of results for individuals or groups if there are skills that are variable within the population that are not being tested but which bring about differences in the measurement of the trait.

1.3 International Schools

Hayden (2006) describes international schools as schools that were set up to provide education for expatriate children who are living away from their home country often because their parents were working as diplomats or for international corporations. While there are international schools that teach in languages such as French or German, the most common language used in these schools is English (Clark, 2014). Because the schools were originally set up to provide education for the children of expatriates in situations where local education was not considered suitable, their populations were initially made up predominantly of native English-speaking children. However, in several countries, there are an increasing number of parents who perceive that it would be advantageous for their children to attend an English-medium school for their pre-university education as this will allow them access to a wider range of university options and more advantageous employment opportunities later in life (Bates, 2011; Sears, 2015). According to Clark (2014), the current demand for placement in international school is dominated in many countries by non-English speaking local families and as a result, Hayden & Thompson (2011) tell us that the composition of the student body has become increasingly multilingual and multicultural.

International schools need to consider both construct irrelevant and construct underrepresentation when using the information from the results of standardised testing. Because most standardised tests are designed with national populations in mind, they will reflect the curriculum of that nation. Differences in the linguistic background of the international school population may result in construct irrelevant variance due to the inability to demonstrate understanding of concepts while cultural

differences may affect test-taking strategies. Further, where international schools do not follow the same curriculum, construct underrepresentation may occur for instance if they teach concepts outside of the defined curriculum for the test.

1.3.1 Curricula design in international schools

In many cities, there will be international schools with national affiliations and their curricula will be determined by the curriculum offered in the national context (Hayden, 2006). So, for instance, a British School would offer the English national curriculum. However, Hayden (2006) acknowledges that in many other international schools, a combination of adaption, integration and creation to form their curriculum. Some schools will include adaptations of recognised programs such as the Advanced Placement (AP) program or the International GCSE. Alternatively, they may integrate programs such as the International Baccalaureate Diploma Program (IBDP). These "off-the-shelf" curricular programs include a terminal examination. However, many international schools will have created their own programs with designs based on the context of the school. For instance, Clark (2014) identified 14 different curricula used across Dubai's 227 international schools. As Sears, (2015) tells us, it is common for such international schools to draw from multiple sources when developing their curriculum. According to Cambridge (2011), international schools will seek to build a curriculum that incorporates the best practices from the successful curricula of different countries or systems. So it would not be uncommon for teachers to draw their resources for humanities classes from the US while their mathematics resources come from Australia, or that they use science resources from England for one phase of schooling and resources from the US for another.

We are told by McClelland, (2017) that international schools will use published standardised tests to demonstrate that a high quality education is being delivered. She goes on to say that schools may find standardised tests useful to monitor attainment and to determine if standards have improved. However, McClelland, (2017) acknowledges that these tests are standardised for populations that are very different to those of the international schools and that such programs ignore variations within the contexts in which they are used. As Goldstein & Thomas (2008) comment, the curricula which are used in national contexts reflect the particular ideas, aims and expectations of that nation regarding the purpose of the curriculum and the content and standards that they identify as important. It is noted by Catling (2017) that while there is a similarity in the set of subjects that are included in the curricula of different nations, there is only a superficial similarity in the content that is actually delivered within schools. Large-scale tests do not have sensitivity to the order in which curriculum elements are taught in different contexts according to Gipps (2012), and the resulting small changes in the performance of students can lead to appreciable differences in ranking scores. The differences in the curricula used in international schools leads to questions about how data from standardised testing can be used in a way that ensures that inferences have validity.

One method of ensuring the quality of the education provided by international schools is through the accreditation process (Bartlett, 1998). Schools will view accreditation by an internationally respected organisation as an indicator of their success (Hayden, 2006), while parents are likely to see accreditation as a minimum requirement to guarantee the quality of education in the school that they choose as

'placing their children in an unexamined school is a risk they do not wish to take.' (Murphy, 1998; p. 242). Acceptance of students into American universities can also be dependent on schools having the appropriate accreditation to guarantee the standards of education that is being provided. The administrations of many schools aim to have accreditation from a combination of internationally recognised organisations as this is seen to be beneficial to their students, and thus may combine accreditation from the Council of International Schools (CIS) with accreditation from one of the American-based organisations (Council of International Schools, n.d.-a). As such, CIS now offers a joint accreditation process with several US agencies; (Council of International Schools, n.d.; Murphy, 1998). The Council of International Schools includes a standard that requires that

> External examination and/or testing results are used to measure students' learning of the taught curriculum, benchmarks with other, similar schools and to support on-going students' achievement (Council of International Schools, n.d.-b, p. 2).

Partner agencies such as the Western Association of Schools and Colleges (2017) also require international schools to have an understanding of the use of standardised testing data and they state that the accreditation process should include a review of group test data.

1.3.2 Use of testing data in international schools

The expectation that teachers in international schools will use assessment data to inform their practice does not differ from conventional "national" schools. Teachers in international schools, therefore, are likely to experience the same difficulties in converting data from standardised testing into usable information. However, there are additional factors which affect their ability to use such data to improve student learning. Students are frequently learning in a second language. The differing

language profiles of students makes it challenging for teachers to separate errors made due to the student's language background from those that are as a result of academic deficits. For instance, students' ability to read and interpret questions may be inhibited because they are still learning English and this may give the impression that student attainment is less than it actually is. As Drennen (2002) points out, there is more potential for discontinuity for student learning in international schools. According to Skelton (2005), international schools have a higher level of student turnover than other schools as there is a higher likelihood that students will move after two or three years due to their parents relocating. Sears (2015) explains that there is a pattern of students entering or leaving the school in the middle of the academic year for instance because the students move from the local education system or the parents move from one hemisphere to another. The discontinuity in the student's education may mean that there are gaps in their learning while frequent movement is likely to make tracking students' progress more difficult in international schools than for those in a national context.

Teacher turnover may also mean that even where students do stay in the same school, there is no opportunity for consultation between current and previous teachers. Because international school teachers are not able to witness student growth over a number of years, they lack the opportunities to know their students sufficiently well. Consequently, it is important to have reliable and valid information that can inform pedagogical decision making, and standardised testing is one potential source of that information (McClelland, 2017). However, teachers seem to have problems with being able to interpret the information that they are given from testing correctly.

International school teachers could also lack the opportunities that their nationally based counterparts have when it comes to receiving professional development and consequently, they may find it difficult to improve their skills in using the data from testing. Holderness (2002) comments that if such teachers are not living in an English-speaking country then it may be that there are no appropriate courses locally while Black & Armstrong (1995) note that there may be significant budgetary implications either in allowing teachers to attend appropriate courses or bringing expertise to the school because both are likely to incur the costs associated with international travel. Holderness (2002) goes on to suggests that teachers who are on short-term or fixed-term contracts may be reluctant to ask for additional funding to attend overseas training events and schools may be reluctant to finance professional development for teachers if it is likely that the teacher will leave after a short while and the school will not benefit from the investment in professional development.

1.4 Purpose of the study

This thesis will consider the impact that the curriculum, the student population and the context of international schools has on the validity of inferences that can be made from the testing data. The technical manuals from the large-scale standardised testing will be reviewed to describe the standardisation procedures and determine how they reflect the population and curriculum of international schools. A picture of how standardised testing is currently used in by a sample of teachers working in four international schools will then be presented. Teachers who are currently working in these schools will be asked about the expectations that are placed on them to use the

data from large-scale standardised testing, how they actually use this data in practice and what challenges they face when trying to use this data.

1.5 Research questions

The study was designed to answer the following research questions:

1. What are the potential threats to validity that teachers need to be aware of when using large-scale standardised tests in international schools?

This question will be addressed by looking at the following sub questions.

   a. Which standardised tests are used in international schools?
   b. How do the standardised tests that are used reflect the curricula used in international schools?
   c. How do the standardised tests that are used reflect the population in international schools in terms of language?
   d. How do the standardised tests that are used reflect the population in international schools in terms of culture?

2. How is the information from standardised testing currently being used as demonstrated by a sample of teachers drawn from four international schools based in the Kanto Plains region of Japan?

In order to answer this question, the following sub questions will be addressed.

   a. What requirements do these schools place on teachers to use data from tests?
   b. What challenges do these teachers identify when using test data?
   c. What training and support have these teachers received to enable them to understand the score reports from the standardised testing that they receive?

The first research question will be answered through an analysis of the technical manuals that accompany the testing that is used in international schools. The second question will be addressed through teacher interviews and will seek to provide descriptions of the work that is done with testing data and the reasons these teachers give for the way that the data is used.

1.6 Rationale for the study

There has been a lot of research carried out regarding the challenges of using standardised testing and the information released in the score reports in national settings such as that in the US (Gallagher, Means, & Padilla, 2008; Means, Padilla, & Gallagher, 2010), Australia (Pierce & Chick, 2010, 2011a, 2011b) and across Europe (Bolhuis, Schildkamp, & Voogt, 2016; Poortman & Schildkamp, 2016; Vanlommel, Vanhoof, & Van Petegem, 2016). In spite of the importance that is attached to using data to identify areas for improvements within schools and the requirement that this places on teachers to be able to understand and use data effectively (Matters, 2006), there is little, if any, research into the use of standardised testing in international schools. According to ISC Research, (2018) there are currently over five million students attending international schools. Their research shows that the numbers of students attending international schools are increasing rapidly. While international schools are not subject to the same accountability requirements as schools in national settings, they are accountable to their stakeholders. As such, they need to show that they are providing students with an appropriate standard of education that will enable them to progress to desired higher education provision. External measures such as large-scale standardised tests are one source of evidence of their academic standards. The transience of students and teachers in international schools leads to some common challenges brought about by the lack of continuity in learning. Standardised testing could provide a source of information to help teachers in planning or to support placement decisions when students move schools.

However, many of the standardised tests that are used by international schools are not validated for their student populations. As Oliveri & Lawless (2018) explain, construct-irrelevant variance could occur because of differences in exposure to the curriculum that is being tested or a failure to understand the cultural references used in the assessment. This could mean that score-based inferences are rendered invalid. As the International Test Commission (2019) warn, when a test is used outside of the country in which it was developed, there are

> limitations related to the appropriateness and relevance of the use of the original normed scores; maintenance of the construct definition or curricular relevance across groups; or the comparability of scores for the multiple test-taker populations. (p. 4).

Schools that are labelled as international are diverse but share certain characteristics such as populations which include students with different levels of language proficiency and students who experience interrupted development trajectories as they move schools often. Using testing where inferences have not been validated for the population present a number of challenges including the ability to identify appropriate uses of the information that is provided after students have taken standardised tests.

The study aims to provide a snapshot of current practice in a small number of international schools in a specific region to gain a more in-depth description of these challenges. However, because the sample is small and cannot be considered representative, no attempts at generalisation can be made. As highlighted by Oliveri & Lawless (2018), increased globalisation and student mobility have resulted in a rise in the use of exported assessments, where a test that is designed for one country is used in a different situation. Ensuring that assessments are valid for students from a range of different cultural and language backgrounds is seen a complex process

and one of the "most pressing and challenging issues confronting test developers and test users." (Schwabe et al., 2016, p. 300). This study will contribute to the knowledge regarding how factors such as language, culture and differences in curricular exposure might jeopardise the validity of the inferences from large-scale standardised testing when they are used in different populations and it could be useful for those who are carrying out studies with a similar research focus.

1.7 Overview of the study

The thesis consists of six chapters. A review of relevant literature is given in chapter 2. As the aim of the thesis is to investigate the validity of inferences from standardised tests in the setting of international schools, it starts with a discussion of current theories on validity. It goes on to consider the literature relating to the perceived threats to the validity of inferences from large-scale standardised testing. The challenges that teachers experience when they try to understand and make valid inferences from the score reports will then be discussed, and finally, the use of standardised testing in international schools will be considered. In chapter 3, the research framework is described. This will explain the rationale for the interpretative qualitative approach that was used and describe the methods used for data collection and data analysis. It will also reflect on some ethical considerations. The research findings are described in chapters 4 and 5. Chapter 6 will relate the research findings and the relevant literature to the two research questions. Here, the implications of the research will be discussed and the limitations of the study and suggestions for future research will be given.

Chapter 2

Literature Review

2.1 Introduction

The call for greater use of data in education has partly been driven by legislation such as No Child Left Behind (NCLB) in the US which called for an increased use of data-based decision making (Bernhardt, 2004). There has also been a push for data to be used to inform practice at the school and classroom level by educational researchers such as Hattie (2005), Mandinach (2012) and Boudett, City, & Murnane, (2005). This process is referred to as data-driven decision-making (DDDM), which Gallagher, Means and Padilla (2007) define in an educational context as being

> a set of expectations and practices around the on-going examination of student data to ascertain the effectiveness of educational activities and subsequently to refine programs and practices to improve outcomes for students. (p. 1)

According to Mandinach & Jackson (2012), DDDM involves the collection, examination and interpretation of data to inform decisions on instruction, administration and policy. They assert that when teachers are trained to select appropriate data and to translate it into information, they can identify ways in which their practice can be changed and this will have a positive impact on education. Mandinach & Honey (2008) state that when school improvement is based on valid inferences from multiple sources of data, then instruction will improve.

Data from standardised testing, as noted by Lai & Schildkamp (2013), is a source that can provide potentially relevant information in this context. According to Masters (2001), when we use measurement instruments such as tests, we want to learn about the knowledge and skills that have been acquired by our students. This

may be to find out how well students are doing compared to their peers (norm-referenced perspective), the progress they have made in relation to a continuum of knowledge or skills (criterion-referenced perspective), or the improvements they have made compared to an earlier time at which they were tested (ipsative norm perspective). Lai & Schildkamp (2013) point out that standardised testing data should not be used alone as it will not serve to identify what teaching and management practices should be implemented to improve learning. However, they go on to comment that ignoring or using information from standardised testing in a limited way can have negative effects. For instance, teachers may underestimate the abilities of their students and this could result in targeting instruction at a level below that of the students' current abilities. However, Booher-Jennings (2005) also warns that when standardised testing is the only source of data that is used, teachers may resort to practices that are designed to increase scores on the test without bringing about a comparable improvement in the student's educational attainment. These practices will affect the validity of the inferences that can be made from the data, and will ultimately stifle the educational development of students.

The following review will consider research relating to the uses of standardised testing, the impact that those uses have on the related validity arguments and how this relates to the use of data-driven decision making in education. The first section will look at the current prominent definition of validity and the rationales for establishing validity in educational testing will be discussed. Validity can be compromised when large-scale standardised testing is used for accountability purposes and so the second section, will discuss concerns relating to the use of tests for accountability. As policies have been implemented both in the US and

worldwide, there has been an expanding body of research analysing the appropriateness of using standardised testing data to measure school effectiveness. Research also examines schools' responses to the introduction of accountability measures and the effects this can have on the validity of inferences that can be made from standardised testing. The second section will discuss this research with the aim of establishing the conditions that are needed to maintain the validity of inferences. The third section will discuss research relating to the use of score reports in schools. Validity requires that the information in reports is understood and correctly interpreted. However, the research highlights a number of obstacles that prevent the information in score report from being used appropriately. Teachers' opinions about the worth of using data, their knowledge about making appropriate interpretations and systematic hurdles including timely access to data may prevent appropriate use of the information. The final section will highlight the particular challenges that international schools need to consider to ensure they are using the information from standardised testing in a valid way. It will look at the characteristics of schools and their populations that could impact the validity arguments that are put forward for the established uses of testing.

2.2 Validity

Validity has been described by Crooks, Kane, & Cohen (1996) as the 'most important consideration in the use of assessment procedures' (p. 266), while Popham, (2018) states that 'without validity, educational testing would have no point, no purpose and no legitimate application' (p. 17). It is not a new concept and theories about how validity should be established have been in existence since the turn of the twentieth century. According to Newton, Shaw, Lagrange, & Robinson

(2014) conceptualisations of validity have evolved over the years with the ideas contained in seminal writings, such as those included in the National Council on Measurement in Education's handbook, Educational Measurement (e.g. Kane, 2006; Messick, 1989), gradually being reflected in practitioners' and test users' considerations. This is shown by changes in successive versions of the Standards for Educational and Psychological Testing (e.g. AERA, APA, & NCME, 2014), often referred to as the Standards, published by the American Education Research Association through the years and seen as the professional standards which measurement specialists need to adhere to.

The definition of validity that is cited most frequently is that of Messick from the third edition of Educational Measurement (Linn, 1989) which states that

> Validity is an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment. (Messick, 1989, p. 13)

According to Kane (2013), Messick's (1989) work had led to the acceptance of a unified model for validation but this model did not give details of the methodology that should be used to establish validity. In the most recent edition of Educational Measurement (Brennan, 2006), the chapter on validity was written by Kane (2006) and this has been described by Brennan (2013) as the most extensive treatment of the subject of validity.

2.2.1 Kane and the interpretative argument

In this work, Kane makes some changes to the conceptual framework of Messick's (Linn, 1989) unified notion of validity. According to Kane, validity theory had seen changes such as a reduction in the prominence of nomological networks and the

more elevated status of the consideration of social consequences. He considers validity to involve the evaluation of the rationale for the claims that are being made for the proposed interpretations or uses of the test score (Kane, 2006).

Kane's main contribution to the discussion of validity was the development of the argument-based approach to validation, which started to gain prominence in the new millennium. He stated that

> The evidence needed for validation necessarily depends on the claims being made. Therefore, validation requires a clear statement of the proposed interpretations and uses. (Kane, 2006, p. 23).

He initially described validity as being composed of two arguments; the interpretive argument and the validity argument. The proposed interpretations and uses form the interpretive argument. Here, a network of assumptions and inferences is laid out that lead from the observed performance to the conclusions and decisions that are to be made based on those performances. The validity argument seeks to show that the interpretive argument is coherent and based on reasonable inferences and plausible assumptions. For example, if there is a statistical generalisation in the interpretative argument then the dependability of the generalisation should be tested in the validity argument. He does warn that there are a large number of potential assumptions for each interpretive argument but states that some may be accepted unless there is evidence to the contrary. For example, he says that it is acceptable to assume that students understand the instructions and have a sufficient level of English to complete the test unless there is evidence that suggests otherwise. With content though, he states that we should always question how well the test covers the content domain as it draws a sample from the domain-related task universe. While he does state that there is no need to develop the interpretative or the validity argument anew

for each person's test performance he warns that there may be special cases where one or more of the assumptions used in the interpretative argument may not hold and this will result in the failure of the interpretation in that special case.

Kane became concerned that his original classifications resulted in too much emphasis on test interpretation and not enough on the test use. So later in his work he used the expression "interpretation/use argument" (e.g. Kane, 2013, p. 2), which was generally referred to as the IUA. He explains that the IUA specifies proposed uses and interpretations of scores from a testing program and applied to a specific population over a range of contexts and will generally include at least three main inferences; scoring, generalisation and extrapolation. In a testing program, we combine task scores from observed performances to give an observed score. The scoring inference is the process of turning the observed performance into an observed score. However, this observed score is only useful as an estimate of performance in a larger domain and so the generalisation inference takes us from the observed performance and observed score to make a claim about performance across the universe of the domain. Where scores are used to predict future performance in another context then an extrapolation inference is required.

2.2.2 Other perspectives on validity

However, Kane's (2006) definition of validity is not without controversy. For instance, Borsboom, Mellenbergh, & Van Heerden (2004) comment that the focus on test score interpretations rather than about the test score itself in current validity theory results is a disconnect between the questions that researchers pose when they are establishing validity in their work and the areas that are focused on by the

validity theorists. They comment that validity theory has evolved into an all-encompassing term which aims to incorporate all the important test related issues but which fails to help the professionals who are trying to apply the concepts because they lack a sense of direction. They deem that many aspects included in the validity argument are irrelevant and that validity theory should be focused on considering whether a test measures what it purports to measure (e.g., Messick, 1989).

In their work, Borsboom et al., (2004) draw a distinction between validity, which they comment is a property of a test, and validation which is an activity that is undertaken to show that a test has validity. They claim that most of the literature which discusses validity is actually talking about validation. Indeed, Kane (2001, 2013b, 2016), frequently refers to validation in his work. His chapter in Educational Measurement (Kane, 2006) is entitled Validation and in that work he does draw the distinction between validity and validation. He uses the term to describe the process of establishing validity. However, Kane (2016) argues that an interpretation that is not justified is not valid and that the core notion of validity is the evaluation of those interpretative claims. Borsboom et al., (2004) state that Messick's commonly cited definition (see above) is incorrect because they state that validity does not involve judgement but is rather a property that is being judged. Their view is that establishing a test's validity should mainly be carried out during the construction of the test and not in the analysis phase. They suggest that current validity theory has the process the wrong way around as it is focused on establishing what has been measured after it has been measured. They state the process should start with establishing what is to be measured by the test.

In their definition, Borsboom et al., (2004) state that validity is a property of the test itself and not of the interpretations that are to be made. They say that there are two requirements to establish validity. These requirements are that the attribute that is being measured does exist and that variations in the scores derived from the measurement tool occur as a direct reflection of variations in the attribute targeted by the testing. They state that

> The crucial ingredient of validity involves the causal effect of an attribute on the test scores implies that the locus of evidence for validity lies in the processes that convey this effect. (Borsboom et al., 2004, p. 1062).

For a causal relationship to be established, they require that variations in the level of the attribute must produce variations in the outcome of the measurement tool and if there are no differences in the attribute then the outcome of the measurement tool would be expected to remain unchanged. Borsboom, Mellenbergh, & Van Heerden (2003; 2004) consider validity within a latent variable framework in which the attribute that is being measured is generally unobservable but has a causal role in bringing about a particular behaviour, i.e. what can be observed, registered, and evaluated in testing situations.

Referring to the correlation between the attribute being tested and the variation in the measurement instrument is not enough according to Borsboom, Mellenbergh, & Van Heerden (2003; 2004); it is necessary to develop an understanding of the potential causality between variations in the attribute being tested and the response variations captured in the measurement tool. Correlations are necessary, but not sufficient to establish causality. They refer to the example of height and weight being related but neither can be used as *causally* predict the other.

One further objection that Borsboom et al., (2004) have to current validity theory is the idea that validity is a matter of degree. Using their definition, validity is established if, and only if the two above-mentioned requirements are met. Borsboom et al., (2004) concede that a valid test may not be the optimal measurement tool for a given situation. They acknowledge that while multiple tests may be valid for measuring a particular attribute, the ability of the tests to measure that attribute may not be equal in terms of reliability or bias. A test exhibits bias when scores have different meanings for different subgroups as a result of test characteristics which are not related to the construct being measured (AERA et al., 2014). Reliability "refers to the degree to which test scores are consistent across time, conditions and test-takers" (Phelps, 2008, p. 110). While reliability is a necessary condition, it is not a sufficient to show validity (Kane, 2013a). As Kane (2013a) explains, establishing validity requires support to be established for all inferences. However, some inferences such as extrapolation are not addressed by reliability evidence.

2.2.3 The current position on validity

In response to Borsboom et al., (2004), Kane (2016) argues that it is not possible to validate a test, or indeed to begin developing a test without knowing the purpose that the scores from that test will be used for. Hence, the validity is relative to the intended interpretation. According to Kane (2013, 2016), Borsboom et al.'s (2004) use of a causal attribute has the potential to provide a rich interpretation but only if the causal theory can be justified. In discussing Borsboom et al's., (2004), Kane (2013) states that if traits exist and are shown to be the main or only reason for an observed performance then causal explanations are extremely powerful. However,

he states that a causal relationship is hard to establish and therefore the theory is seen to be limited in its applicability. He sees Borsboom et al., (2004) position as restricted to one type of implicitly-assumed interpretation which he says limits the ability to respond to the various applications of test scores.

In his comments, Kane (2013) reminds us that we are not interested in the test scores for their own sake but rather as an estimate of an underlying, latent attribute. Assumed knowledge of the level of that attribute may then be used to make decisions. He states that tests are developed with a purpose or use in mind and test score interpretations are generally claims regarding attributes of individual test takers, groups of test takers, or even teachers and schools. Scores are used to make decisions about these groups. Because such interpretations are connected to practice, then the appropriateness must rely heavily on how relevant the score interpretations are. According to Kane (2013), there are multiple possible uses and interpretations and he restates his position that validity is the property of the proposed uses and interpretations of the test score and not of the test itself. Further, validity depends on the degree to which the evidence supports the proposed interpretation or use. The most recent version of the Standards was released in 2014 to provide a framework for professionals involved in educational measurement to evaluate testing practices and the use of score interpretations and to promote ethical testing practices (AERA, APA, & NCME, 2014). The Standards describe validity as 'the most fundamental consideration' (p. 11) and it is the first of the core chapters in the Standards. They reflect Kane's position of validation as the construction of arguments for interpretations and planned uses.

2.2.4 Validity and educational practitioners

As was stated in the introduction, Hattie (2014) argues that teachers and other educational professionals are considered to be test users. As a result, they will be expected to select the appropriate tests to fulfil requirements in their particular situation and to interpret the results correctly. We are reminded by Wiliam (2014), the responsibility of ensuring that a test is valid for a specific use rests with the test user. The Standards (AERA et al., 2014) instruct that

> those who are participants in the testing process should have appropriate knowledge of tests and assessments to allow them to make good decisions about which tests to use and how to interpret the results. (p. 3).

Consequently, teachers and educational professionals are required to have an understanding of validity and the associated evidence that is provided by the test developer.

However, much of the debate about validity is conducted by measurement professionals and researchers, and is far removed from teachers and administrators working in schools. For instance, in their explanation of the importance of the Standards, Plake & Wise (2014) state that the intended audience are professionals in the fields of psychological and graduate students in programs relating to educational measurement. Others, such as Pitts & Naumenko, (2016) comment that teachers are not part of the intended primary audience of the Standards even though it could be perceived that they are an essential group of stakeholders. However, Plake & Wise (2014) inform us that one of the goals of the Standards was to promote assessment literacy amongst teachers and therefore it would be beneficial for classroom teachers to read the Standards. They note that it was decided that classroom assessment was beyond the scope of the Standards. However, Ferrara (2014) disagrees with Plake &

Wise's (2014) view that reading the Standards would be beneficial to teachers as he notes that they are written in the technical language of measurement professionals which is not suitably attuned for practical use by classroom teachers. It is acknowledged by Pitts & Naumenko (2016) that even the guidelines that are most applicable to teachers do not facilitate teacher engagement. Furthermore, Ferrara (2014) comments that the Standards do not include the right concepts to develop teachers' assessment literacy. He states that teachers should develop assessment literacy by using texts written for such a purpose before they read the Standards.

It is not just the Standards whose design excludes educational professionals. Hattie (2014) comments that the documentation that test developers are expected to produce to inform test users about how scores are intended to be used, are also prepared with the assumption that users will be measurement professionals. He reminds us that users are expected to understand the validity evidence that supports intended interpretations. However, there is a questionable expectation that users will be professionals who have training or credentials which will help them to determine if they are making appropriate interpretations from the testing data. Test developers fail to take into account that many of today's users do not have the benefit of training in educational measurement and lack understanding of the concepts used.

So how should teachers gain understanding of concepts such as validity when they are trying to understand testing results? Ferrara (2014) tells us that there are many good textbooks for teachers produced by measurement professionals, however, those that are mentioned tend to focus on classroom assessment. We are advised by Popham (2011) that many of the textbooks on educational measurement that are

designed for teachers may be out-of-date. There are textbooks that give what Wiliam (2014) refers to as the consensus definition but not all offerings include the consensus definition. For instance, Bernhardt (2013) tells us "The validity of a test or assessment refers to whether it provides the type of information desired." (p. 77) before giving definitions for seven different types of validity. It is also noted by Wiliam (2014) that outdated definitions of validity are promoted in other resources that are designed with teachers in mind. He cites standards produced by the International Reading Association & National Council of Teachers of English, (2010, p. 52), while other sources such as The Graide Network, (2019) and Te Kete Ipurangi, (n.d.) also use the outdated definition that a valid test measures what it "purports to measure".

2.2.5 Summary of section 2.2

Validity is deemed to be one of the most important considerations relating to testing. The onus of ensuring that inferences are valid rest with test users and consequently it is essential that test users understand the concept of validity and are able to relate it to the circumstances and context in which testing takes place. Frequently nowadays those test users will be teachers and educational administrators who have not been trained in the principles of educational measurement and so lack the understanding of concepts such as validity. Further, teachers lack the opportunities to be able to develop an appropriate level understanding of this topic. In spite of changes that mean teachers are now a major group of stakeholders in the interpretation and use of test results, they are frequently not considered as stakeholders when work is published that relates to the major considerations in testing. Furthermore, the outcomes from the work of measurement professions can be slow to trickle down

into resources designed for teaching professionals, meaning that teachers may be left working from resources with outdated concepts. The disagreements of educational measurement professional may also be reflected in the materials that are provided resulting in confused understanding of these essential concepts.

Included in the materials that are produced by test developers are technical documents which are designed for test users and explain the validation processes that tests have undergone. However, this information is only useful to test users such as teachers if it is accessible, and if it provides the appropriate details to support them in making decisions regarding the tests that they plan to use. So what information do these documents give regarding the validation procedures that are undertaken during test development? How accessible are these documents to the test users who are making decisions about appropriate test use? It is also important for teachers to understand that tests are validated for specific purposes and that validity is compromised if the conditions surrounding test use are violated.

2.3 Validity issues relating to the uses of tests and testing data

As was stated in the introduction, the growth in the use of standardised testing has been attributed to the increased use of accountability measures across national educational systems (Smith, 2014). Accountability is seen by the National Research Council (1999) as a process in which schools give account of their performance so that they can be monitored by policymakers, tax payers and parents. A UNESCO (2017) report comments that governments struggled to manage educational provision because of a rapid expansion during the latter half of the 20th Century. This resulted in a shift towards emphasising efficiency and equality (Smith, 2014). According to

UNESCO (2017), governments devolved control over educational provision to local authorities. However, they maintained responsibility for monitoring and funding with monitoring taking the form of large-scale standardised testing. UNESCO (2017) report that the in spite of the increased numbers of students attending school there was no proportional increase in educational funding. Rather countries expected better value from their investment. Results from standardised testing led many countries to conclude that their education systems were not providing a good quality education (Kim, 2018; Linn, 2005). As will be discussed below, poor test performance was attributed to poor quality teaching (Alhamdan et al., 2014; Ball, 2008; Smith & Benavot, 2019), differences in opportunities for students based on race or social class, and the lack of a uniform curriculum (Kim, 2018; Linn, 2000, 2005).

2.3.1 Accountability and teachers

Accountability is based on the premise that high achieving students are the product of a good quality education and any differences in achievement are solely attributable to differences in the standard of education that students receive (Wiliam, 2010). According to Smith & Benavot, (2019) public trust in schools and teachers has been undermined by crisis narratives. These narratives seek to apportion blame for poor educational standards on teachers who are incompetent, (Ball, 2008) lazy or unmotivated (Alhamdan et al., 2014). Thus, by making teachers work harder or more efficiently, an improvement in standards in the form of increased test scores will be seen. However, the introduction of accountability measures can have mixed effects on teaching quality.

Research by Williamson, Bondy, Langley, & Mayne (2005) showed that some teachers respond to accountability measures by introducing more child-centred policies. Their study was based at a school that was deemed to be failing by their state authority. Two teachers acknowledged that a student's backgrounds could influence their ability to relate to different representations of a concept. These teachers gained an understanding of their students' backgrounds before deciding on the methods of presentation and the explanations and analogies that would be used to introduce concepts. The students performed well in their high-stakes tests. The school had previously been categorised as failing but the higher percentages of students achieving the required standards lead to it being re-categorised as a passing school. The researchers suggest that the introduction of a high-stakes policy does have not to mean the end of child-centred pedagogy. However, the introduction of accountability measures does not always have a positive impact on teaching quality.

Case studies were conducted in two schools in Chicago by Anagnostopoulos (2007) to establish the effects of the introduction of a state accountability program based on high-stakes testing on reading attainment. This accountability program included the introduction of state-wide curricular standards with associated standardised testing as part of an assessment program and a promotion policy for students at four stages in their compulsory education which required, amongst other things, for them to achieve grade level scores in the standardised testing. A designation of 'probationary' was applied to schools that failed to achieve a prescribed proportion of students attaining prescribed levels in the standardised tests. According to Anagnostopoulos (2007), provision was put in place to support students and teachers in schools given this category. Monitoring of scores and the course failure rate was

used to judge the performance to teachers. This policy served to make teachers more responsible for student attainment in the standardised tests. Principals were also monitored and evidence of unacceptable attainment levels could be used to fire the principal. The accountability measures in this study were seen to lead to a number of reforms within the schools. These included the establishment of before school recovery programs, lunchtime tutoring programs and extra classes for students with below standard levels. A Reading Task Force was formed which led to the introduction of a reading calendar which designated days for departments to include reading strategies in their classes. However, principals also met with teachers whose students were not seen to be making appropriate progress and it was implied that teachers were responsible for the success and failure of their students with messages that "good teachers do not fail students" (p. 303). In response to the introduction of high-stakes accountability in the Anagnostopoulos (2007) case studies, teachers instituted practices that were focused solely on increasing test scores. This resulted in devoting significant amounts of class time to test practice and the development of test taking skills. Teacher-centred methods were employed and the curriculum was narrowed to focus on skills that teachers thought would appear in the test. While some teachers were seen to introduce practices to help individual students achieve, these were considered to be ad hoc methods that were responses to students' individual circumstances rather than resulting from their educational needs. There was also evidence of a lowering of students' requirements and goals by eliminating assignments, altering grading structures and reducing the number of set texts to be read. However, attributing poor student performance solely on underperforming teachers treats the issue in a very simplistic manner.

Wiliam, (2010) argues that differences in the quality of education only have a very small impact on student outcomes. According to Blazer (2011), the differences in test scores may be a reflection of differences between schools and that differences in funding, facilities and neighbourhood environment also impact student achievement. For example, Wiliam, (2010) uses within-school variance to measure the impact that schools have on their students. The effects of an individual school is included in the between-school variance which Wiliam, (2010) tells us is composed of the systematic differences in attainment between schools and the school-effect. Tests do not just incorporate knowledge and skills that are taught in school but also skills that students accumulated throughout life. Systematic differences arise because the school does not have any impact on those accumulated skills. Socioeconomic factors are seen to be one such factor. Those students who come from higher socio-economic background are likely to experience benefits through lifestyle factors. These include having parents who have a higher level of education who are more able to help students with homework or who are willing to provide experiences that are specifically aimed at benefitting their child's education, perhaps including additional classes or programs outside of school to support their learning. According to Baker et al. (2010), because students are not randomly assigned to schools, the effects of factors relating to the socio-economic situation of students will vary depending on the school demographics. It is estimated by Wiliam, (2010) that between-school variance represents about 25% of total variance and that 69% of this can be accounted for by difference in the population of students that are attending the school (within-school variance). This would mean that schools are only responsible for around 8% of the variance in student outcomes.

This argument attributes a large proportion of the variability in school performance to differences in school populations rather than to differences in the quality of education provision. However, Hochberg & Desimone (2010) comment that of all school-based factors, being taught by high-quality teachers has the biggest effect on student achievement. Indeed, Wiliam, (2014a) also argues that replacing the weakest teachers with average teachers would still have a significant impact on attainment in schools. In defence of accountability measures, Good, Wiley, & Sabers (2010) argue that even if accountability measures were only to result in a small amount of additional higher quality teaching, the cumulative effect over several years would be significant. However, Wiliam, (2014a) goes on to say that identifying the weakest teachers can take time as learning is not an instantaneous process. Consequently, practice that may seem good at the time may not lead to long-term retention of knowledge. Arguing for the replacement of the weakest teachers also assumes that there are better teachers to replace them with. However, schools do not have equal access to appropriately qualified staff (Aragon, 2016; Garcia & Weiss, 2019). Countries such as England and the US report teacher shortages (Adams, 2019; Sutcher, Darling-Hammond, & Carver-Thomas, 2016), and frequently such shortages occur in mathematics (Ingersoll & Perda, 2009; Noyes, 2007; Sutcher et al., 2016), which is one of the subjects frequently included in the testing. It is also the case that the schools in those areas where student attainment is the lowest are the ones that have the most problems in recruiting and retaining teachers (Aragon, 2016; Diamond, 2012; Garcia & Weiss, 2019). Equity is one of the frequent reasons cited for the implementation of accountability measures but systematic failures such as the supply of high calibre teaching staff can undermine the attempts to address this issue. It is argued by Harris (2012a, 2012b) that

differences in teacher expectations may be one source of inequity in the educational provision for students from different socioeconomic, ethnic groups and language backgrounds.

2.3.2 Equity of educational opportunities

According to Schraw (2010), accountability measures were justified on the grounds of leading to equal opportunities for all students. As stated by Spillane (2012), the NCLB mandated proficiency standards were based on the ideal that everybody should have the chance to excel regardless of identifiers such as race or social class. However, accountability measures do not necessarily result in equitable opportunities for students. As Harris (2012b) tells us, teacher's expectations can be skewed by the student populations that they teach. Rather than teachers using testing as information to help move students forward in their learning, she tells us that teachers blame students for their results and then use the student's background to justify the poor results they attain. Teachers may have the perception that students' capacity to learn is dependent on their socioeconomic or ethnic backgrounds. They adopt a deficit mindset in which poor performance for such students is attributed to a lack of motivation and low ability levels. Consequently, teachers may have lower expectations of African American and Latina/o students or those from lower socioeconomic backgrounds. Harris (2012b) tells us that these lower expectations lead to differences in opportunities to learn including a focus on teaching basic skills. Teachers perceive a need to improve understanding of concepts that should have been taught to these students in previous grades. According to Noyes (2007), this bias can affect the academic paths that students are assigned to. This includes both the groupings they are assigned to within schools and their opportunities to

move on to schools that might allow them the chance of a better-quality education. He also finds that the bias affects the students' own perceptions of themselves as learners and acts to reinforce students' expectations of success or failure. It is argued by Diamond (2012) that these lower expectations also affect pedagogy. He comments that teachers in lower-performing schools use more didactic methods involving lecture and recitation while higher-performing students experience more authentic instruction involving meta-cognition and problem solving. Nevertheless, there was some evidence that the introduction of accountability measures did lead to improved attainment for students particularly in groups where attainment was seen to be the lowest.

According to Carnoy & Loeb, (2002) implemented accountability measures saw significantly higher gains in the low-stakes National Assessment of Progress (NAEP) mathematics tests for grade 8 students compared to those that did not have such accountability measures. A review of the NLCB carried out by Dee & Jacob (2010) also detected positive effects in the long term trend data of mathematics scores provided by the NAEP test. Improved scores were achieved by younger students, particularly those from disadvantaged populations and larger improvements were experienced by Hispanic students according to Dee & Jacob (2010). However, they go on to comment that this improvement was not mirrored by black students and that results for reading did not show similar improvement. Dee & Jacob, (2010) note that the results at the extremes of the distribution were not affected equally which could mean that improvements were not attained by making provision more equitable but rather by focusing on the groups of students who are more likely to impact accountability measures. Given that NCLB measures are based on

proportions of students achieving a specific standard of proficiency, this could imply that the changes mentioned by Dee & Jacob, (2010) were as a result of schools focusing on the students at the middle of the distribution who were closer to attaining the required level of proficiency, while students who were farther away from achieving proficiency received no benefits.

However, the introduction of accountability measures can serve to exacerbate the problem of inequity. According to Diamond (2012), schools serving children from ethnic minorities or those from families with lower socio-economic status are more likely to experience sanctions as a result of accountability measures. He tells us that these low performing schools frequently respond to the threat of sanctions by reallocating instructional resources to focus on those students who are close to attaining the cut scores needed to be classified as proficient as they are the ones most likely to help schools meet prescribed pass rates. Research by Booher-Jennings (2005) provides an example of a school in which students' prior testing results were used to decide their academic path. This research was conducted at a school which had a minority population that was economically disadvantaged. She explains that the response to the introduction of testing was for students to be separated into three strands according to their ability. The first group were considered to be safe in that teachers were certain they would reach the required standard in the test. The middle strand comprised of so-called bubble kids who are those students considered to be closest to achieving the standard for passing but were not quite there. It was considered the most likely way for the school to reach its target passing rate was by getting these students to the required standard so the school focused its teaching and financial resources on these students. Consequently, resources were allotted to them

with the aim of turning them into passing students. This was at the expense of the third group who were considered least likely to pass the test. The school tried to have the students in this final group excluded from their accountability subset through such practices as assigning them to special needs programs. The policy of focusing on those students who were most likely to pass was justified by claiming data-driven policies were being employed to target resources based on the perceived needs of students as identified through deficits in their assessments. In this case, accountability measures did not result in students receiving equal opportunities. Rather the opportunities of the lowest attaining students were restricted as attempts were made to assign them to special needs programs that may not have been suitable for them. Accountability measures aim to promote educational equity by linking academic standards and curriculum content to standardised testing according to Diamond (2012). In this way, all students should have access to the same curricular content.

2.3.3 Alignment between content standards and assessment

It is argued by Kim (2018) that poor standards were blamed on differences between the curricula that were being taught. He states that content decisions were based on the whims of teachers or influenced by the choice of textbooks. Students were being tested on content that they had not been taught because schools did not focus on core standards, and this led to poor performance in testing. To address these inequities in the academic experiences of student, accountability proponents argued that there was a need to standardise the curriculum (Kim 2018). According to Forte (2010), accountability starts by defining performance standards which detail what a student should know and be able to do in each content area and for each grade level.

Darling-Hammond & Rustique-Forrester, (2005) argue that aligning standardised tests to these performance standards will result in control over what is taught in school leading to more curricular coherence. By measuring the content, the tests act as a signal to educators of what is important and influence what is taught and how it is taught. This should lead to a rise in achievement in the domain that the standards represent according to Polikoff (2012). We are told by Figlio & Loeb (2011) that standardised test scores are then used to provide assurances that educators are focusing on the right content. According to Darling-Hammond & Rustique-Forrester, (2005), these curriculum changes, along with more focused professional development and better allocation of financial resources will lead to an improvement in scores attained in the tests that are being used as accountability measures. However, the validity of interpretations based on test results is dependent on alignment between the content standards and what is being tested.

As Koretz (2005) acknowledges, even when tests are aligned to appropriate curricular standards, inappropriate teaching methodology can lead to score inflation which can undermine the validity of the inferences that can be made. We are told by Hannaway & Hamilton (2008) that tests can only sample a small number of the skills that are included in the content standards and some skills are not amenable to testing. According to Jennings & Bearak (2014), when teachers become familiar with test formats, they can focus their teaching on skills that are tested rather than on the curriculum as a whole. Instances where teachers are seen to adopt methods of teaching to the test rather than to the content standards are identified by researchers such as Anagnostopoulos, (2007); Clarke et al., (2003); and Firestone, Mayrowetz, & Fairman, (1998). As stated by Jennings & Bearak (2014), aligning instruction to

the test leads to an incomplete alignment with the standards. We are told by Koretz (2005) that incomplete alignment means that the sample of items tested is no longer representative of the domain. This renders inferences about attainment gains or proficiency levels invalid. It is therefore important that changes reflect the content of the curriculum and not the content of the test.

A qualitative meta-synthesis of 49 studies carried out by Au (2007) found a strong link between the introduction of accountability testing and changes in curriculum content and pedagogy. The dominant changes were in content alignment, with 69.4% of studies reporting that the introduction of testing resulted in contractions in the curriculum that was delivered while 28.6% of studies reported an expansion of the subject matter that was taught in schools. In selecting articles for the meta-synthesis, Au (2007) only used studies that employed qualitative research methods. Such research does not employ comparison but will provide analysis of observational data and so causal inferences cannot be made. Given that much of the research that is published on accountability tends to focus on the negative aspects (Wiliam, 2010), there is a danger that this bias is reflected in Au's (2007) analysis.

An example of curriculum expansion is seen in a case study by Brimijoin, (2005). A teacher's response to the implementation of standard-based learning and high-stakes testing details the procedures she goes through to structure lessons that take into account the abilities of all her students is studied. The teacher starts with a process of backwards design which starts by interrogating the content standards that her class are expected to learn. The results are used to design lessons and a class format that enables student learning and formative assessment is used to identify areas for

differentiated instruction. By aligning the curriculum to the standards that were being tested, the teacher in the study was able to raise the level of attainment for all her students.

Evidence suggests that changes to teaching practices are not restricted to what happens within the lessons of the subjects which are involved in high-stakes testing. For instance, in surveys of 500 school districts reported by Berliner (2011), 80% of districts reported increasing the curriculum time allocated to Language/arts by 75 minutes or more per week while 63% reported similar increases to the time allocated to mathematics. More than 50% stated that this increase in time allocated to Language arts was over 150 minutes and 19% of districts reported increasing mathematics time by more that 150 minutes. Provision was made for these changes by reducing the time allocation to other subjects such as social studies, science and music or by reducing the time students were given for breaks during the day. However, Hannaway & Hamilton (2008) argue that this may not be evidence of negative changes. By reviewing teacher's logs and survey information, they found that prior to the introduction of accountability measures, there was significant variation in the amount of time schools devoted to teaching mathematics in fourth and fifth grades. They found that the difference across the two grade levels between those schools who taught the most and those that taught the least could be equivalent to 23 weeks of instruction. As was mentioned above, studies have highlighted grade 5 mathematics as an area where improved attainment can be seen after accountability measures have been introduced. Hence, the change in curriculum timings could be a redressing of the balance rather than an elimination of the subjects that are not tested. They acknowledge that the situation will be different across schools. Therefore, it is

difficult to know the extent to which the changes are due to necessary increases in time allocation for the tested subjects and the proportion of cases where the changes lead to inadequate time allocations for non-tested subjects.

A school's unique response to suggestions that curriculum time should be reallocated is described by is described by Vogler (2003). Rather than reducing the time allocated to social studies to allow for more time to be devoted to mathematics and language/arts, a curriculum review was instigated. The school saw the aim of accountability as giving opportunities to students to develop their creative and intellectual potential and so they explained the curriculum standards to the students and then tasked the students to design an integrated curriculum. The expectation was that by involving the students in curriculum planning, they would find the curriculum more interesting and would therefore be more motivated to work. This would lead students to gain a better understanding of what they were supposed to be learning, substantial improvements to the curriculum and improved experiences for the students. However, the ability of the students to understand the content defined in the curriculum standards is potentially problematic. Furthermore, the students who were meant to benefit most from the implementation of standards-based education are those who are most in need of more intensive educational support and therefore least likely to be able to unpack what is meant by the curriculum standards.

As was mentioned above, much of the published research into accountability highlights negative practices in which changes introduced may lead to increases in the results of the high-stakes testing which are not evident in other lower-stakes measures of the same constructs. Research that identifies positive effects resulting

from the implementation of high-stakes testing is more limited. Although there are case studies showing evidence that high-stakes testing can result in appropriate modifications to teaching practice that have a positive impact on student attainment, these tend to be more difficult to find. Phelps (2015) argues that the use of standardised testing to monitor productivity is perceived as a threat to the status quo of education and so evidence that supports its use is suppressed and does not make its way into the prominent research journals.

However, it is not just schools' responses that can undermine the validity of inferences from large-scale standardised tests. There can be problems with the tests themselves. Popham (1999) quotes research that found that less then 50% of the items that were included in tests of mathematics received more than a cursory treatment in the textbooks that were chosen to deliver content. While educational objectives are generally similar, there can be differences in the knowledge and skills that are identified as important in different localities, but test companies may try to use or adapt similar tests across these localities meaning that the tests are not truly aligned to the content in a particular area. Meanwhile, Buckendahl, Plake, Impara, & Irwin, (2000) note a mismatch between the standards that teachers view are being included in standardised tests and those that test publishers think they have included. For instance, in one test, teachers identified that six of sixteen standards had been included, while two publishers stated that eleven and sixteen matches were found respectively. These were tests for which publishers had released validity alignment analyses (Buckendahl et al., 2000). Research by Polikoff, Porter, & Smithson (2011) also raises questions about the alignment of testing to curriculum standards. They found that even when using the most relaxed definition of alignment, the highest

alignment they could obtain was 50%. Misalignment was identified in content and cognitive demands. This lack of alignment means that it is important that teachers are able to access and understand test documentation so that they are aware of differences as such mismatches can lead to incorrect conclusions being made.

2.3.4 Appropriate uses of test results

There is opposition to the use of the results of standardised tests to judge the performance of schools. As Crooks et al. (1996) argue, there is no test that provides interpretations that are valid in all situations or for all purposes. Morris (2011) states that the purpose of the test must be stated clearly as this will drive the test design. Any intended purposes of the test need to be well validated. We are told by Gipps (2012) that a different type of testing is needed for accountability purposes. Tests that are designed and validated for measuring student performance are not appropriate means for this purpose. Standardised tests are generally normed to measure the performance of individual students, not teachers or schools and any attempt at accumulating the results across students to evaluate the efficacy of either teachers or schools will be affected by sources of variance that will render the inferences invalid. To use this testing to judge schools would require all schools to have the same intake but schools have very diverse intakes. As Leighton (2020) argues, schools are complex social environments and children do not walk into schools as blank slates but rather have experienced different combinations of various different cultural, linguistic and socioeconomic variables throughout their life and this will have an impact on their level of attainment. According to Morris (2011), to maximise validity and reliability, the test must be developed using established

content standards. Failure to align testing to curriculum standards will give misleading information about the extent of student success in meeting the standards.

There are some disputes about the best way to use test results. For Wiliam (2010), tests results should be used to report student mastery in curricular aims with results being given to teachers in a timely manner so that they can use the information to inform their teaching. Teacher's capacity to use results effectively and appropriately is described by Morris (2011) as a critical pillar of an assessment system. However, as will be discussed later, this requires teachers to have a sufficient level of assessment literacy (Morris, 2011).

As Valli & Buese (2007) acknowledge, the introduction of high-stakes testing does not have to result in negative teaching practices. Rather, it is both the situation and the people involved that will affect the responses to the introduction of high-stakes testing. In their comparison of two states where high-states testing had been implemented, Firestone, Mayrowetz, & Fairman (1998) found that teachers may be encouraged to think about changes they could make to their practice but this does not necessarily bring about a motivation to change and even where is does, resources such as professional training may not be available to help the teacher to facilitate changes in their teaching practices. It is argued by Bowman (2018) that bringing about change in educational establishments is a complex issue that is dependent upon the different personalities involved. Novice and veteran teachers and school leaders all have different mindsets and goals. According to Bowman (2018), curricular change affects all levels of the school and changes that might be positive for one level may not be received well by another level. However good the change

is, some people will have resistance as they don't like the chaos change might bring. Bowman (2018) states that change requires the conviction of stakeholders and advocates for more communication and collaboration which leads to negotiated changes.

Wiliam (2010) concludes that there is evidence both from comparisons of different states in the US and from different education systems around the world that suggests that the implementation of accountability measures can bring about positive impact on student achievement, but there is also evidence that suggests that the impact can be negative. Much of the evidence above comes from what Wiliam (2010) describes as naturalistic studies where researchers use case studies to investigate the impact of the implementation of policies on individual teachers, schools or districts which means that the results are not generalisable. For generalisable results, an experimental design which employed comparison would need to be used. While it is more difficult to carry out such research as in educational settings as students and teachers are not randomised into class settings, quasi-experimental designs or clustered randomised trials could be used to give a better understanding of factors that would improve educational outcomes.

To encourage positive change in student achievement, Wiliam (2010) suggest that curriculum aims should be defined broadly around a small number of key concepts which are written in language that teachers can understand. In addition to this, Anagnostopoulos (2007) states that there is a need to balance the pressures that the introduction of standards-based reforms places on teachers. She advocates for intensive training to support teachers in using assessment to identify the needs of

students. Teachers also need to support in altering practice to meet those needs including the supply of appropriate teaching resources.

2.3.5 Summary of Sections 2.3

Accountability assumes that test scores are a direct indicator of the educational quality that a school provides. However, it fails to take account of the many other factors that contribute to the scores that students attain in large-scale standardised testing. For test-based interpretations to be valid, it is necessary that students have the opportunity to demonstrate what they have learnt. This means that the test design should incorporate what is defined in the curriculum so that all learners have the opportunity to display what they have learnt (Gipps & Stobart, 2009; Polikoff et al., 2011). It also requires that every effort is made to reduce sources of construct-irrelevant variance by making sure the test only includes what is defined in the curriculum. Factors relating to a student's out-of-school life experiences contribute to the scores and tests are not able to separate the impact of these experiences from the contribution that is made by the school. When tests are validated, they are validated with a particular purpose and population in mind. Attempts to use the tests in ways that they are not designed for is likely to result in inferences that are invalid and incorrect. To understand the ways in which tests should be used and the inferences that can be made from them, it is important for teachers to have an adequate level of assessment literacy.

Education is not an automated process and the implementation of new policies within specific educational contexts is dependent on the personalities involve. Thought needs to be given to the changes that need to be made and to how to

introduce them in a way that results in all stakeholders implementing the changes in a positive manner (Bowman, 2018). This includes initiatives involving the use of data from standardised testing to improve attainment.

2.4 Using data to inform practice

Schools have been described by researchers such as Mandinach (2012) as being data rich but information poor. By this she means that schools have always been in possession of a lot of data about their students but have previously made little use of that data. As was stated at the beginning of the literature review, accountability measures place high demands on educators to use evidence-based processes in their planning. According to Earl & Katz (2002), this has led to a rise in the profile of student achievement data. There is now an expectation that data from testing will be analysed and the interpretations will be used to understand the current position of attainment within the school and to formulate plans for future improvement. As Matthews, Trimble, & Gay (2007) tell us, for teachers to use data, they need to understand what the data means, be willing to accept it and be prepared to change their instruction in light of what they find. However, as a result of a number of perceived barriers, the information from testing is frequently not used. This section will consider the effect of teacher attitudes on data use. It will then discuss how teachers' lack of knowledge and skills impact their ability to use the information that is produced after students have undertaken standardised testing. It will discuss the provision of professional development to improve teachers' skills. Finally, it will consider how teachers' ability to use data is affected by practical considerations such as the timeliness of receiving data and the accessibility of that data in schools.

2.4.1 Teacher attitudes

Teachers choose not to use data because they do not see the need to reflect critically on their own practice (Schildkamp & Ehren, 2013). They express the opinion that students are too complex for data to offer solutions (Murray, 2013). They view their experience as being more relevant than data (Schildkamp & Ehren, 2013). Teachers decisions are seen to have a profound impact on the students' they teach and it is therefore important that those decisions are objective. In research on grade retention, Vanlommel, Van Gasse, Vanhoof, & Van Petegem (2017) found that teachers made decisions about students relatively early in the academic year. These decisions were informed by student behaviours including puzzled looks when asked questions or frequency of crying. When teachers chose to use data, they looked for data that would confirm their opinions. If data contradicted their opinions, they don't see this as an opportunity to learn something they didn't know but rather focus their explanations around why the data was wrong. We are told by Vanlommel & Schildkamp (2019) that data is filtered through teachers' own lenses. As was mentioned in section 3 above, teachers' expectations can be influenced by the socioeconomic and ethnic background of their students (Harris, 2012b). This can lead to data being used to reaffirm stereotyping leading, for instance, to students from lower socio-economic groups being incorrectly designated to special education programmes (Booher-Jennings, 2005) or to bias in choices about which students should not progress to the next grade (Vanlommel et al., 2017). Thus initiatives that were meant to resolve inequalities are being used to reinforce the biased decisions that are made. Mandinach & Schildkamp (2020) tell us that by using a range of data that emphasises the strengths that students bring to the classroom, it is possible to change teachers' mindsets regarding their expectations of students from different

backgrounds. The research of Vanlommel et al., (2017) highlights the importance of the systematic collection and use of data to prevent such confirmation bias. Vanlommel & Schildkamp (2019) advocate for triangulation using of multiple sources of data to increase the likelihood of making appropriate interpretations.

We are told by Schildkamp & Ehren (2013) that for teachers to be successful in using data, they need to have a positive attitude to its use. During a two year study of an intervention designed to improve data use Keuning, Van Geel, & Visscher (2017) found that there was more hindrance to the introduction of the measures in schools where teachers displayed negative attitudes to data use. Their intervention was implemented in 101 schools across two phases. There were 53 schools in the first phase and 48 in the second. Questionnaires were used to compare the attitudes of teachers in the 5 schools within each phase that had the highest effects to those of teachers in the 5 schools with the lowest effects. Reasons given for negative attitudes included a lack of leadership, differences in perceptions of DDDM and because the intervention was imposed on the teachers. Positive attitudes formed as teachers became aware of the benefits of DDDM. Wayman & Stringfield (2006a, 2006b) report that even teachers with negative attitudes will engage in using data when teachers see that their students will benefit. As Hattie's (2010) work shows, of the variables that are controllable, it is teacher behaviours which are likely to have the biggest impact in education and one resource that can influence their behaviour is data-based information about student performance. However, Wayman & Stringfield (2006a, 2006b) say that thought must be given into how those measures are introduced. This will help teachers to overcome negative emotions such as fear.

A survey by Moore & Shaw (2017) found that teachers were more positive about data use if they were confident in their ability to use data.

According to survey information quoted by Pierce & Chick, (2011b), about 40% of teachers indicated that they did not perceive that the reports they received from testing were important, while Mandinach, Rivas, Light, Heinze, & Honey (2006) stated that educators are reluctant to use standardised testing data because they have doubts about the accuracy of the information given. Many teachers question the validity of the data they received because they associate it with accountability. Furthermore, Bernhardt (2000) says that teachers express a sense of fear of embracing data-driven methods because they are scared that data may reveal something that they do not want to see and this will lead them to question their own competence. There is fear that the data to be used as evidence against teachers and schools as a result of the processes of accountability. Because many of the tests are misused for accountability purposes, teachers see accountability as the sole purpose of testing. As has been expressed in previous sections, these tests are not validated to assess teacher efficacy but rather student learning. As Popham, (2009, 2018) advocates, teachers need to have an understanding of the purposes for which tests are designed. He states that this would enable teachers to campaign against inappropriate uses of testing such as when tests designed to measure students are used to evaluate schools. However, this does not acknowledge the power dynamic between teachers and those who legislate for the use testing for accountability purposes. Testing is also popular amongst the general public, many of whom do not have the needed level of assessment literacy to distinguish between appropriate and inappropriate used of testing results.

According to Means et al. (2011) teachers did show more confidence and were more likely to use data if they were allowed to analyse data with colleagues. Their research drew participants from a number of schools who were considered to be active in their use of data to inform instruction. Individual and small group interviews were conducted. Participants were presented with data scenarios and asked to think out loud when responding to questions about interpreting information in the scenarios. It was found that groups were able to answer more questions correctly and could extract more useful information than teachers who were working alone. Analysis of the interviews showed that one teacher's misconceptions could be corrected by other colleagues in the group. Working in groups was also seen to mitigate biases. According to Dunlap & Piro (2016) teachers prefer working with colleagues as this allows them to socially construct knowledge, particularly around the use of effective teaching strategies. However, this identification of effective strategies presupposes their ability to identify the areas that they need to improve which requires that they can analyse data first. To be successful in group analysis would require that at least one of the teachers could successfully interpret the given information otherwise at best, this would result in wasted time and at worst could lead to bad decision being made based on invalid interpretations. In their study, Means et al. (2011) only used schools that were already considered to be using data successfully. Therefore, teachers had already received training in data use. When they carried out an intervention to improve teachers understanding of data, Pierce & Chick (2012a) found that the teachers who expressed more confidence in their ability were not always better able to provide correct interpretations of the data. For example, Pierce & Chick (2012b) report instances of teachers who stated that they

could correctly interpret box plots but who were found to have misconceptions when questioned. They held the common misconception that the regions of the box plot related to frequency rather that density. Rather than teachers learning to interpret data properly, this could result in having a teacher who is confident but does not understand the data leading other teachers to make incorrect inferences and to learn incorrect methods for interpreting the data they have.

According to Popham (2009), assessment literate teachers are able to make better informed decisions because they know how to use testing and assessment to gain insight into the progress their students are making. Popham (2018) tells us

> Assessment literacy consists of an individual's understanding of the fundamental assessment concepts and procedures deemed likely to influence educational decisions (p. 2).

He goes on to explain that the focus of assessment literacy is on the use of educational measurement to influence decisions. There are two clusters of learning required for assessment literacy as defined by Popham (2009). The first involves the ability to develop classroom assessment. However, he states that it is also essential that teachers have an understanding of standardised testing so that they can evaluate whether such tests are suitable for a given purpose.

2.4.2 Teachers' lack of knowledge regarding how to make appropriate interpretations from the information provided in tests.

One of the skills that teachers need to develop to be considered assessment literate, according to Popham, (2009) is the ability to understand the components of score reports and to interpret the information contained in them. Teachers' inability to interpret data reports is probably the major factor that inhibits the constructive use of data from the reports that they receive according to Pierce & Chick (2011b). Many

teachers indicated that they do not use the data to inform their teaching even though they perceived that the data could be useful. Teachers commented that data was of little importance to their teaching because they were unable to interpret the summary reports that were sent to the school (Pierce & Chick 2011a). Teachers also reported that they find the reports are too complex and Pierce & Chick (2011b) found that teachers choose not to engage with data because they were overwhelmed. As Stiggins (1991) notes, while interpretive guides often accompany test score reports, test users such as teachers have very little background training. Many teachers lack the skills and understanding that is necessary to turn educational data into information that can aid their teaching.

To investigate teachers' use of data received from standardised testing, Pierce & Chick (2011b) surveyed 84 teachers. Their sample consisted of mathematics and English teachers as the tests were focused on numeracy and literacy. They found less than half the mathematics teachers said they found the reports easy to understand while a third of them suggested they were neutral or not confident in their ability to understand the information. Their sample was voluntary and represents only a relatively small number of teachers but their findings mirror the results of other researchers. However, choosing not to engage with something because it is difficult is contradictory to the expectations that teachers have of their students. Teachers are often questioned about the relevance of the subject content that they teach but it would not be acceptable for students to decide what is appropriate for them to learn and what they can ignore. Students who find difficulty in understanding particular concepts are encouraged to keep trying. If there is evidence that data interpretation can lead to improvements in student learning then

teachers have the same requirements to overcome the difficulties that they place on their students.

2.4.3 Professional development to improve teacher competence

The ability of teachers to understand the implications of data depends on them having a sufficient level of assessment literacy. However, as was mentioned in the introduction, teacher education programs have not included the requirement for teachers to learn about educational assessment and consequently many serving teachers have had little, if any, understanding of the fundamentals of educational measurement. Research from the US Council on Teacher Quality found that less than 2% of a sample of 180 undergraduate and graduate courses in teacher education included any work relating to the analysis of test results (Bruniges, 2011). The situation does not improve when teachers gain employment with many reporting that they do not have access to professional development courses in the analysis of data (ACER, 2009; Mandinach, Honey, & Light, 2006). Of the few courses that are available, most are designed for administrators rather than teachers (Mandinach, 2012). Consequently, it is entirely possible that many teachers may be expected to interpret educational data to make decisions about their teaching practices and the future learning opportunities of their students, even though they have had little or no training in understanding the forms of presentation that are used and the possible conclusions that may be drawn. So, as Popham (2009) tells us, due to the number of teachers who are now in the profession who have not had the opportunities to develop the essential knowledge to interpret testing and assessment data, it is important that appropriate professional development programs are developed and delivered to fill the gaps. Teachers themselves have expressed a desire to have more

professional development on the use of data (Gallagher et al., 2008). They suggest that the provision of appropriate professional development is a major factor that affects their decisions to use data (Angelico, 2005) and a survey reported by Gallagher et al., (2008) found that 48% of teachers expressed a need of professional development to help them to interpret testing data.

Before professional development is provided, there should be an understanding of the current level of proficiency and identification of the difficulties that prevent better analysis of the data (Means et al., 2011). While targeted professional development has been shown to improve teachers' capacity to use data, as with all professional development, it is important that the content that is delivered is chosen carefully. If not, there is a danger that the professional development will fail to provide teachers with adequate skills to enable them to use data after the professional development has been delivered (Means et al., 2010). Professional development is designed to improve teachers' knowledge about how to improve student learning with the expectation that they will use that knowledge to transform their practice (Avalos, 2011). However, courses are usually of short duration, frequently lasting only one or two days. It is expected that teachers will learn specific concepts related to their own position within the school and then make changes to their work practices. As such, this type of professional development results in teachers receiving training which is designed to bring immediate change to a specific part of their practice. This is not an effective way to ensure that practices change because teachers may not retain the skills that are taught (Pierce & Chick, 2012a). Developing assessment literacy and understanding the complex issues surrounding the use of data to inform practice is something that will require longer term

professional development. As Popham (2018) acknowledges, it is an area in which only having a little knowledge can cause more problems than it solves. Care must be taken to make sure that professional development is not a short term training process that aims to give teachers a set of procedures which are lacking in depth or applied without understanding as the resulting analysis may not prove beneficial to student learning (Chick, Pierce, & Wander, 2014). Rather teachers need to develop an understanding of the important concepts so that they can apply them across their practice.

Professional development which is aimed at improving teachers' understanding of assessment and their knowledge of pedagogy in curricula-specific content is most likely to benefit student learning. According to Bruniges (2011) and Ingvarson (2005), there also needs to be information about how teachers can help students to narrow the gap between their current level of attainment and the level that they are required to reach. Teachers wanted to undergo professional development in developing diagnostic assessments and in using diagnostic data to adjust instruction (Gallagher et al., 2008). Teachers are likely to gain more if professional development provides them with a chance to work with data from their own students and identify areas of weakness within their own environment (Ingvarson, 2005). Hess, (2008) suggests that any professional development opportunities must allow teachers to familiarise themselves with the types of presentations that are used in score reports and the types of valid inferences that can be made (Hattie, 2009). Courses involving multiple training sessions that immerse teachers in the analysis of data have been shown to improve teachers' understanding of statistical concepts while remedying some of their misconceptions (Confrey, Makar, & Kazak, 2004;

Doerr & Jacob, 2011; Makar & Confrey, 2002). Such courses also improve teacher attitudes and perceptions towards the use of data. Such professional development would be both time-consuming and costly, and with schools having other priorities, it is unlikely that such opportunities would be offered to all teachers who could benefit from such training. However, failure to provide training could mean that teachers either do not use data or they use it badly. This could lead to detrimental consequences for students as they are not provided with the most appropriate educational opportunities based on their needs.

The provision of professional development on its own is not enough to guarantee that teachers will use data analysis to inform their practice. The school culture must encourage the use of the skills that are developed as otherwise there is a danger that teacher practices will remain unchanged. Support must be on-going as Means et al. (2010) suggest that teachers may find that they are not sufficiently confident in the taught skills when they try to implement the practices on returning to the teaching environment. It can also be the case that teachers do not use data analysis often in their practice and so do not remember the concepts in the longer term. Consequently, it is important that procedures are in place to constantly renew and refresh teachers' skills. One suggestion from Pierce & Chick (2012a) is to include the provision of online courses that review the skills taught in the professional development programme which can be used by the teacher when they are most in need of support.

A report from the United States suggests that some schools choose not to provide professional development but prefer to encourage teachers to use data by employing

data coaches who can assist teachers in examining their own data and collaborate with them to develop instructional plans (Means et al., 2010). The introduction of data teams, where a group of teachers is given training and then provide support to their colleagues, was seen as a successful initiative that could improve data use. Collaboration with colleagues was encouraged and teachers were found to be more engaged in data analysis and to make better decisions when they worked in small groups. Furthermore, as mentioned above, teachers reported that they found the data analysis process more enjoyable when they worked with colleagues (Means et al., 2011). Yet, Means et al. (2007) reported that teachers are more likely to be working alone when using data to make instructional decisions relating to their class or individual students that they are teaching. Many researchers such as Wayman & Stringfield (2006b) comment that DDDM works best when all teachers are fully involved in all stages of the process.

2.4.4 Practical considerations that affect data use

It is not just the lack of training that can affect teachers' decisions to use data to inform their practice. There are a number of factors identified by teachers relating to organisational arrangements in school which prevent them from analysing testing data. Teachers state that they do not have an adequate amount of time to spend on reviewing the score reports. Furthermore, getting access to the reports in a timely manner can also prevent teachers making use of the information in these reports.

2.4.4.1 Time Issues

Teachers need to be given adequate time to study data reports if they are to gain maximum information from them. Yet many teachers state that they do not have the

necessary time to devote to the process so that they are able to gain useful knowledge from data (Coburn & Talbert, 2006; Means et al., 2007; Schildkamp, Karbautzki, & Vanhoof, 2014). The teachers commented that data use is something that they are expected to do in their own time (Means et al., 2010). Teachers who state that they already lack the time to complete their current duties will be unwilling to commit time to interpreting data because this would reduce the time they give to teaching and other preparation (Kerr, Marsh, Schuyler Ikemoto, Darilek, & Barney, 2006). Furthermore, they will see this as wasted time because they perceive that they cannot interpret the data (Pierce & Chick, 2011a) or because they do not see the relevance of the information gained. However, in spite of teachers' comments that they lack the time to analyse data, there is little evidence to suggest that providing teachers with time and resources has improved the use of data in ways that will benefit student outcomes (Bruniges, 2011). As discussed above, having additional time to review data will not necessarily be beneficial until appropriate pre-service and in-service training is used to address the deficit in teachers' skills (Pierce & Chick, 2013).

2.4.4.2 Procedural issues

Data is only useful if the diagnostic information is returned to schools in a timely manner and issues relating to the timeliness of data can also result in teachers failing to make appropriate use of data (Smith, 2005). It can take many months for testing data to be returned to schools and so by the time it is received, the usefulness and relevance of the diagnostic information given is questionable (Wasson, 2009). Teachers want data that is current to their students and have expressed frustration

when they find that the only data that is made available to them is from assessments that were taken more than six months previously (Means et al., 2010).

It can also be a problem for teachers to gain access to the data that is available within the school according to Lachat & Smith (2005). While the data may be available in schools, according to Wayman & Stringfield (2006b) it is stored in a way that means that teachers have difficulty in gaining access to it and when they do it is not in a usable format. The data is generally held in large electronic files and if teachers are trying to combine testing data with demographic or attendance data, they may have to access different systems to get the information they need. Initiatives to improve the use of data work best according to Kerr et al., (2006) when data is received in a timely fashion, efforts were made to improve the ease of teacher access to data and schools provided facilities such as computer applications that enabled teachers to work with the data.

As with appropriate use of standardised tests, Wayman & Stringfield (2006) comment that research into DDDM has centred around case studies involving descriptions of best practice in a few schools. However, there is little evaluation of these practices in the research and so teachers cannot identify which practices would work in their own setting. They state that the research community need to move beyond research methods that provide description to establish a better base which seeks to understand and explain the practices that work so that appropriate interventions can be identified.

2.4.5 Summary of 2.4

In spite of the importance that is attached to the use of data-driven methods, teachers still lack the necessary understanding of how to interpret and use assessment data. Furthermore, teachers lack the opportunities to develop the necessary skills, with little opportunity to attend appropriate training. Other challenges that are identified by teachers including a lack of time, score reports that are out-of-date before they are received in schools and difficulties in gaining access to data within their school.

Much of the research used here is from the US and Australia. However, this leads to questions regarding the challenges faced by teachers in other situations. How do schools such as international schools that are not subject to accountability legislation use standardised testing? Are their teachers able to make appropriate use of the data? If not, how do they improve their skills? As mentioned in the introduction, the provision of professional development is very different in these situations. Long-term professional development is recommended in the literature and this would be very costly. International schools may not be willing to pay such costs for teachers if they don't think it will benefit the school in the longer term. There may also be difficulties in getting consistency in professional development for teachers who move frequently between schools.

2.5 Standardised tests and international schools

Testing programs evolve in response to an identified need and their development will be instigated by an organisation who will take the lead in determining the features of the test. Because of this, tests are designed and validated with a particular population in mind. As we are reminded by McClelland, (2017), the tests

that are being used in international schools have generally not been standardised for use in such schools but rather for schools in national contexts. As Stoelting (2019) tells us, the characteristics of the national population that the testing is designed for will differ greatly to the populations in international schools.

Tests that are developed for use in one country but then are used with a population that is different to the one for which the assessment is developed are described by Oliveri, Lawless, & Young (2015) as exported assessments. It is estimated by Oakland (2004) that up to 50% of tests are developed in one country and exported for use in another. When testing is used with a population that is different to the population that was used for the validation studies, Kane (2006) states that 'Aspects of the procedure, the context, or the population being tested may interfere with the effectiveness of the procedure in a particular context.' (p. 55). He explains that when the interpretations are validated for a specific defined population then the assumptions that are used to build the interpretative argument may not hold up when the test is given to a population which is different in terms of characteristics such as educational background, age or language proficiency. With norm referencing, the student's performance is compared to that of a preselected group but it is important that the comparison is made with a group that is relevant to the test taker (Reynolds et al., 2010). For instance, in the US, the sampling plan

> stratifies the samples by gender, age, education, ethnicity, socioeconomic background, region of residence, and community size based on population statistics provided by the US Census Bureau (Reynolds et al., 2010, p. 57).

When scores are used with different populations to those that they are developed for, results may be impacted by irrelevant sources of variance resulting from differences in curricular opportunities, linguistic backgrounds or cultural understanding and so

Oliveri & Lawless (2018) warn that special attention needs to be given to the results gained from such uses of these assessment. As explained by Oliveri & Lawless (2018), this means that differences in attainment by the new population may not be as a result of differences in skill level but rather because of differences in the ability to understand cultural references or idiomatic expressions or because of differences in test-taking strategies and behaviour. These differences could result in construct-irrelevant variance and this would mean that inferences based on the results of testing may lack validity for the new population being tested.

2.5.1.1 Establishing the validity of inferences for imported tests

A number of threats to validity are identified by Oliveri & Lawless (2018). Firstly, test takers may not have had the opportunity to learn the curriculum that is being assessed. Secondly, score-based inferences may not be the same for the different population as sources of construct-irrelevant variance that are influencing the ability of the new population to answer questions. Finally, it is possible that the test does not measure the intended construct.

According to Wendler & Powers, (2009), when a test is used in a new situation, procedures that should be carried out before the test is used, including explicitly identifying differences between the original use the test was developed for and its proposed new use, the development of a plausible argument that explains why the function of the test should not differ with the new population, and creating a plan to obtain evidence to determine the validity of the intended uses and interpretations. Oliveri et al., (2015) put forward a framework that involve using diagrams based on Kane's (2013) IUAs. They suggest that a panel of experts should be involved in each

stage of the process. However, when it comes to using large-scale testing across international schools, it is the teachers and administrators at the school who will have the responsibility for establishing validity and questions remain about their assessment literacy and their abilities to carry out the necessary processes.

2.5.2 Current research into the use of standardised testing in international schools

One analysis of the implications of standardised testing is carried out by Walker (2017) who uses a case study to describe how the International Schools' Assessment (ISA) is used in an international school in France. The ISA is a standardised test that is designed to provide benchmarking for international schools (ACER, 2019a). Walker (2017) comments that international schools have particular challenges when using standardised testing. As mentioned in the introduction, these challenges include the transience of students which makes it more difficult to use data to measure student growth. Furthermore, the educational patterns that these students follow may add to the challenges. For instance, students may not always attend English-speaking schools as they may return to local schools if they spend time in their home country between moves or if their parents perceive the language of a new country useful and want their child to learn it.

Challenges also relate to the diverse student populations that international schools have and, as Sears, (2015) comments, it would not be unusual for schools to have students from over 60 nationalities. This diversity includes students with varied linguistic backgrounds, including students who are English language learners (ELL) as well as those who are bilingual or multilingual. Sears (2015) comments that it is acknowledged that bilingualism is accepted as a spectrum of language use. Those at

one end of the spectrum will have a balanced command of all aspects of the languages they use. However, these balanced bilinguals are considered to be rare. At the other end of the spectrum are those who may only use limited aspects of one of their languages, for instance they may be able to read or understand to a limited degree. It is acknowledged that different levels of proficiency in language use is shown by bilinguals in different aspects of their life. This diversity will have added complications for teachers when they try to interpret data relating to their students' progression.

2.5.3 Language and standardised testing

Kieffer, Lesaux, Rivera, & Francis, (2009) argue that there is a unique set of challenges created for policy makers and educators due to inclusion of ELL students in large-scale assessments and these raise questions about the validity of the inferences that are made about ELL students' abilities. Young, (2009) states that where ELL students take the same content tests as native English speakers, it is necessary to show that the interpretations that have been validated for other learners also apply to the ELL students. However Abedi (2006) indicates that where the performance of ELL students is lower on tests in mathematics, science and social studies, this may because all tests require a certain amount of language proficiency and therefore it may be that the level of English that is required prevents ELL students from demonstrating their content knowledge. When lack of proficiency in English results in difficulty in understanding instructions or items on a content test then test scores may suffer from construct-irrelevant variance. Young (2009) states that this can be a concern even in English language arts tests where it is particularly

difficult to distinguish the effects of English proficiency from that of content knowledge.

Abedi (2006) comments on a series of studies based on controlled experiments which found that linguistically complex items prove difficult to ELL students but that when the complexity was reduced, the performance gap between ELL and non-ELL students was reduced. The linguistic complexity of the test may result in measurement error leading to construct-irrelevant variance and this will lead to a reduction in the reliability and validity of outcomes in assessments for ELL students. It has also been noted by Oliveri & von Davier, (2016) that the performance of ELL students on test items may be affected by their ability to understand linguistic features such as idioms and to some of the contextual settings used in questions. They state that this may threaten score comparability as ELL students may not be able to demonstrate their proficiency in a construct because they do not understand the language. However, they also highlight the impact that lack of familiarity with the texts used in test situations or differences in test taking strategies may have on test scores.

There is evidence cited by Durán (2008) that the results ELL students attain in standardised assessments include measures from characteristics outside of the knowledge and skills that they are intended to measure and this leads to concerns about the validity of any inferences that could be made. In his paper, Durán (2008) comments that ELL students form a very heterogenous group in terms of the languages other than English that they speak, the amount and type of instruction that they have had in their non-English language or languages, the curriculum that was

used in that non-English language, and the age at which they have started to learn English. All of these variables will impact on their readiness to learn in English and will mean that even where the same scores are attained in a test, it could have very different meanings in terms of the instructional needs for the student. In considering language minority students from around the world, Schwabe et al., (2016) comment that composition differences occur both within countries as well across countries. Schwabe et al., (2016) define language minority students to be those whose primary home language is different to the community language of their country. They note that ELLs in the US are likely to be living in low-income families which means they will be raised with different cultural norms to others in the US. Parents of students in international schools are more likely to be from the middle classes within their own countries.

The Standards (AERA et al., 2014) note that there are special challenges when testing students who know more than one language because there may be differences in the way they use their languages and therefore an understanding of the degree of bilingualism is necessary when testing is used. For instance, students may use their native language in social situations while English is the language they use in school. Teachers need to be aware that students' conversational English may give the impression of fluency but they are not as competent in reading and writing which may impact their ability to show their skill level in testing. The Standards (AERA et al., 2014) go on to acknowledge that subgroups within the testing population may be heterogenous which may in turn affect the appropriateness of test content and the relation of test scores to other variables.

2.5.3.1 Use of accommodations

One method that is employed to mitigate the effects of language skills on the validity of interpretations that can be made from the results of testing is to provide ELL students with accommodations. Accommodations are adjustments to the standard assessment procedures which are not relevant to the test construct being measured but which minimise the impact of irrelevant differences in student characteristics. While the Standards (AERA et al., 2014) promote the principle that tests should be designed in a way to minimize the potential for construct-irrelevant variance to impact test scores, they acknowledge that it is not possible to make all tests accessible to all members of the target population, and therefore it may be necessary to provide accommodations for individuals to prevent characteristics impeding their access to the test. However, they warn that appropriate accommodations should not change the construct that is being measured or the score meaning.

In the case of ELL students, there are a range of accommodations that can be made available including allowing extra time to complete the test, providing bilingual dictionaries and having the test read aloud either with a teacher reading the test to a group or individuals having access to on-demand recordings. The types of provision that are allowed are different depending on which test is taken. Staehr Fenner, (2016) advocates for the use of testing accommodations for ELL students as she says that the provision of appropriate testing accommodations can give ELL students a greater chance of demonstrating their skills on content tests. However, she warns that different students will benefit from different accommodations and incorrect provision can actually prove detrimental to students. For instance, students who are at the beginning level are more likely to need accommodations but may be

overwhelmed if given a word-to-word dictionary. However, there was no evidence to suggest many of the different accommodations that were provided for ELL students were effective in improving their ability to demonstrate their attainment in different subject areas according to a meta-analysis carried out by Kieffer et al., (2009). For instance, providing a translation of a test which uses the student's dominant or home language may not prove helpful when their academic language is English as they may not have learnt the academic vocabulary in the dominant or home language. The only accommodation that was found to show a statistically significant improvement was the provision of a glossary or English dictionary. Many of the documents used in their analysis were based on small samples and it is likely that they were either related to a specific language group or that conclusions could be affected by the variability of the ELL learners in the study. Given that the number of ELL students in national education is rapidly increasing and the inclusion of this group in large-scale standardised testing, there is a need for more detailed research to be carried out on how effective the different accommodations are, the point of the language development spectrum in which they are most appropriate and whether there are differences based on the language background of the student.

As Murphy, (2017) points out, even with accommodations, ELL students may be subject to different conditions to others taking the test. For instance, if the test is being read aloud, the student will have to work at the pace of the interpreter and will not have the benefit of a silent environment. It is also possible that adequate accommodations cannot be provided. For instance, with all the different language backgrounds that are present in international schools it may not be possible to find an appropriate interpreter who speaks the required language.

2.5.4 Testing and culture

The Standards (AERA et al., 2014) warn that tests should not include content that confounds measurement of the target construct and include examples such as use of words or expressions associated with particular socioeconomic groups, cultural backgrounds or geographical locations. However, Valdes & Figueroa, (1996) comment that tests are culturally biased and favour students whose background give them the appropriate cultural capital. Meanwhile, it is noted by Schwabe et al., (2016) that ELL students are likely to have different cultural backgrounds and this can lead to additional challenges for test developers. International school students may not have experienced life in the country in which the test was written so there is a higher danger that they will encounter questions with settings that they have no experience of. However, it is not just the questions that can cause differences in the ability of students to respond to test questions. Differences is cultural norms can also affect the way students answer questions. For instance, it is acknowledged by Park et al., (2013) that there are cultural differences in tendencies to express anger between western and Asian culture. They tell us that the US, adopts an independent stance for the pursuit of individual goals. Meanwhile culture is Eastern Asia is focused on interdependence which views the person as being a member of a larger cultural groups according to Cheung (2004). In such cultures, Park et al., (2013) tell us that expressions of disagreement and anger are seen to threaten that interdependence. Oliveri & Lawless (2018) comment that these differences may affect the way in which students from Eastern Asia respond to questions in testing.

Guidelines from the International Test Commission, (2019) state that fairness reviews should include representatives from all cultural groups, but when a test is exported it is likely that there will be no representation from the new countries that are taking the test. Therefore it becomes the responsibility of the test user to review the test to establish the impact of the cultural setting of a question on the test-taker's results.

2.5.5 Differential item functioning

One method that is used by testing developers to ensure that items test fairly across all subgroups of the testing population is Differential Item Functioning (DIF). DIF procedures are designed to calculate the probability of answering questions correctly for different groups who have been matched by ability within the testing population. Differences in the probabilities could indicate that performance is affected by items other than the construct that is being measured. Young, (2009) states that when DIF is used to analyse the responses of ELL students, the items that are found to have exhibited DIF are those which have high levels of language complexity.

There are different procedures used for DIF and according to Young (2009) they can produce markedly different results. Oliveri & Lawless (2018) point out that the challenges of using DIF include the classification of the groups, especially if the test is being used on a different population to that for which validity has been established. For instance, possible classifications used to compare ELL groups home country, language group or first language proficiency. Each will lead to different results. Even here, the heterogeneity of the groups can make conclusions problematic, and small sample sizes can lead to difficulty in interpreting results. It is

important that the procedures used to identify testing items that prevent ELL students from showing their understanding of testing constructs are reliable in carrying out that function. Oliveri & von Davier, (2016) promote that idea that ELL students should be catered for in all stages of test development rather than considering the effects solely towards the end of the development procedures. It would be assumed that test writers do not deliberately write questions that show bias towards a particular subsection of the testing population. However, test developers are generally picked because of their subject knowledge. It may be that there is a need for ongoing training to keep writers up-to-date on theories related to ELL.

2.5.6 Teacher turnover and initiatives in international schools

We are told by Chandler (2010) that issues relating to recruitment and retention teachers in international schools are complex. A survey of over 22000 international school teachers quoted by Odland & Ruzicka (2009) found a turnover rate of 14.9%. According to Odland & Ruzicka (2009), this is comparable to rates that would be described as troublesome in national contexts. To find out about teacher retention, Hardman (2017) surveyed 30 teachers working in international schools. Teachers in the sample stated that they believed that the optimal length for staying at a school was five to six years. However, he found that most initial contracts were for two years and that less than half the teachers had renewed their contract more than once, meaning that a majority stayed for less than four years. This survey involved a small number of teachers in four locations. Teachers' desire to move may be affected by location according to Chandler (2010). He tells us that working and living conditions vary between countries. For instance, he found teachers who described working in Africa as a hardship post. Hardman's (2017) survey included teachers

from Tanzania and Egypt, so it may be that their location influenced their decisions to stay. Hardman (2017) comments that the transient nature of the international school population can be disruptive to the learning environment. He includes comments from follow-up interviews. One head of department states that they have experienced discontinuity in a series of postings. This discontinuity led to haphazard curriculum delivery. A deputy headteacher also commented that the discontinuity meant it was difficult to measure academic learning, particularly in the short-term. This may be a justification for implementing analysis of standardised testing measures amongst teachers. However, this discontinuity can also mean that initiatives to measure academic performance can fall by the wayside as senior teachers move school.

As a former international school teacher now working for one of the testing agencies, Stoelting (2019) relates how he went from volunteering to go on a training conference for the Measures of Academic Progress (MAP) testing to being nominated as new testing coordinator in the international school he worked for. His assumption was that MAP had been introduced with a great vision in mind. However, reflecting the transient nature of teachers in international schools, the person who had introduced MAP testing into the school had moved on and nobody else had sought to continue the work. Consequently the vision had not been realised. He explains that when he began working at the school, testing was viewed as something the school just did for accreditation and that the data was used to make sure that the school was keeping up with US norms in education. As stated in the introduction, accreditation is seen as a measure that a school is maintaining an appropriate level of educational standards. Parents of international school students

look for accreditation as a sign that an international school is providing their child with an education that at least matches the quality of the education that they would receive in their own country. Universities will look for the school to have accreditation before offering places to the students who are graduating from the school.

The data from testing was seen as unimportant at Stoelting's (2019) school because it was something that was not used and so he attempted to reboot data use by working with stakeholders to show them how data could be utilised and this led to the information from the testing being embedded into processes including those for making decisions such as who should be placed in enriched programs and whether students were ready to exit English as a Second Language (ESL) program. The data was then used by classroom teachers for differentiation and to form work groups in classes.

2.5.7 Summary of 2.5

It is important to identify reliable sources of information that can help international school teachers to progress the educational attainment of their students because of the transience of students, teachers and administrators in the community. International school teachers face challenges when trying to interpret the information provided in the score reports of large-scale standardised testing because of the combination of students with different language background and different cultures. While accommodations are given with the intention of ameliorating the language issues, it is suggested that such accommodations may make little if any

improvement. Differences in the curriculum taught in international schools also lead to threats to the validity of inferences that can be made.

Before they can establish that the test is suitable for the purposes they wish to use it for, teachers and administrators need to have an understanding of the processes that have gone into the development of tests, while the challenges of interpreting test results from different subgroups such as ELL students supports the need for teachers and educational professionals to have undertaken appropriate professional development.

2.6 Summary of the literature review

The literature review started by considering validity, which is an essential component to legitimise the inferences that can be made from testing. The current definition and the processes required for establishing the validity of uses and inferences of tests were described. It was also highlighted that essential literature, such as the Standards (AERA et al., 2014) and the documentation that is required to explain the development of tests, are produced with educational measurement experts in mind. However, the test users who will need to understand that documentation nowadays are frequently teachers who have little or no training in measurement concepts.

It is important to consider the uses and associated interpretations for which testing has been validated when considering the use of a test. So the review continued by considering some of the uses of testing to highlight that even where testing has been validated for one use, it does not mean all uses of testing are valid. This highlights

the importance that test users understand and can determine appropriate uses for testing and more importantly that they are able to establish when uses and interpretations are unwarranted.

The review then looked into the different factors that prevent teachers from being able to interpret and apply the information they are given from large-scale standardised testing. It highlighted that most teachers have not received training in the use of such information. However, recommendations indicated one-off training was not enough and that there was a need for long-term professional development to ensure that teachers develop and use the necessary skills to make maximum and appropriate use of the information they receive.

Finally, the review considered research regarding the current use of standardised testing data in international schools. It noted that there was need for reliable information because many students and teachers move frequently in the international system but it highlighted that much of the testing that is used is not designed with international school populations in mind. It was also acknowledged that the composition of the student population in terms of language and cultural profile could be different to the populations that tests were validated for.

The difference in composition of the populations of international schools compared to those for whom testing is validated led to the development of the first research question which aims to use the documentation produced during test development to describe the population and uses for which validity has been established by the test developers and the procedures that are carried out to establish that validity.

Concerns that teachers in international schools are expected to choose and interpret tests even though they do not have the knowledge and skills to understand measurement concepts led to the second research question. This question aims to describe the skill level and professional development opportunities that internationals school currently have. It also asks international school teachers to describe the perceived barriers to data use in their situations. The next chapter describe the research framework that was used to investigate the two research questions.

Chapter 3

Methodology

3.1 Introduction

This chapter will outline the methodology that was used to answer the research questions.

1. What are the potential threats to validity that teachers need to be aware of when using large-scale standardised tests in international schools?
2. How is the information from standardised testing currently being used?

The research started by reviewing the websites of international schools to identify the large-scale standardised testing that is used in international schools. Question 1 was then addressed by reviewing the technical documentation that is published to support the use of these standardised tests in schools. The information relating to the test design was analysed. This included, where relevant, the curriculum that was being tested, the standardisation sample that was used and the population the test was designed for. Question 2 was addressed by interviewing a small sample of teachers drawn from four international schools based in the Kanto Plains region of Japan. The teachers were asked to describe the expectations that their school placed on them to use the information from standardised testing. They were also asked about the challenges they encountered when using this information and the training and support they had received to enable them to use this information.

3.2 Research design

Because the aim of the research was to explore potential challenges relating to the use of large-scale standardised testing in international schools, it was decided that an interpretative paradigm was the most appropriate. The research was exploratory

(Cohen, Manion, & Morrison, 2018). The aim of the research was to build a subjective interpretation of use of standardised testing as it is used in a small number of international schools. There was no intention to develop theory but rather the aim was to describe the experiences of using the information that is received from these tests through the eyes of the teachers working in those schools.

According to Kivunja & Kuyini (2017) the interpretative paradigm assumes a subjectivist epistemology in which the researcher makes meaning of data through their own thinking. Knowledge is constructed socially through the personal experiences of real life that the researcher gains within the natural settings that they investigate, while Cohen, Manion, & Morrison, (2011) tell us that researchers have their own lenses through which they view the world. While the documents that are used in the analysis are factual, the analysis will focus on the interpretation of those documents to support the use of standardised testing in an international school setting. I am an international school teacher with experience of using this data in my practice and ultimately the information will be interpreted through my eyes. The interpretative view acknowledges that people's actions are based on their interpretation of the world around them. Teachers' decisions about using the information from score reports will be influenced by factors within their school, and their own conceptualisation of the usefulness of that information.

3.3 Sample/participants

The sample was drawn from international schools in the Kanto Plains region of Japan. I reasoned that there was a sufficiently large number of international schools in this area and this would make it possible to get a large sample of teachers to share

their experiences. There was no reason to see these schools were atypical of international schools in other countries (Merriam, 1998). The schools are all English medium and their websites state that they draw their student populations from a large number of countries and a significant proportion of students are not Japanese (Hayden, 2006). These schools do not base their curricula decisions on the curriculum used in Japanese national schools (Hayden, 2006). The eleven international schools listed on the Kanto Plains Association of Secondary Schools website (KPASS, 2019) were invited to take part. However, several of the international schools explained that they frequently received requests to conduct research and so only permit their own teachers to carry out research. Consequently, only five schools gave permission for their teachers to be involved in the research. In one of these schools, there were no teachers willing to participate. Therefore the final sample consisted of nine educators from four schools (see table 3.1 and 3.2 below).

Schools A, B and D catered for students from Kindergarten through to the age of 18 while school C catered for students from Kindergarten up to the age of 15, which was considered by them to be the end of Middle School. Schools B and C followed the International Baccalaureate (IB) programme. Schools A and D were not following a specific curriculum program but students in both schools followed programs and entered exams for the Advanced Placement (AP) examination during their High School years, mainly in grades 11 and 12. School C followed the Primary (PYP) and Middle Years (MYP) programmes.

Table 3.1: Schools represented in the sample.

School	Curriculum	Nationalities in the student population	Testing Used
A	North American	Over 25	IOWA PSAT AP
B	IB MYP and DP	Over 50	ISA PSAT IB DP
C	IB MYP but described as aligned to the curricula of leading developed nations	Over 60	ISA
D	Described as drawing on the best practices from around the world	Over 40	ISA PSAT AP

The teacher from school B taught in both the MYP and DP. Two of the teachers held management positions. There were four English arts teachers, one mathematics teacher, one science teacher and a teacher of modern foreign languages who had a particular interest in the use of data and was involved in implementing practices into his school. Five of the teachers were from the US, three were Australian and one was trained in England.

Table 3.2: Interview Participants.

	Subject taught	School	Teaching Level
Teacher 1	Mathematics	A	Middle School High School
Teacher 2	English	A	Middle School High School
Teacher 3	Science	B	High School
Teacher 4	Spanish	C	Middle School
Teacher 5	English	D	Middle School High School
Teacher 6	English	D	High School
Teacher 7	English	D	Middle School High School
Administrator 1		A	High School
Administrator 2		C	Middle School

3.4 Instruments

The aim of the first research question was to identify the potential threats to the validity of inferences when standardised tests are used in international schools. To do this, it was necessary to understand the procedures that are carried out to establish the validity of inferences for the population that each test is designed for. According to the current version of the Standards for Educational and Psychological Testing -

> Information relating to tests should be clearly documented so that those who use tests can make informed decisions regarding which tests to use for a specific purpose, how to administer the chosen test, and how to interpret test scores. (AERA et al., 2014, p. 125)

I decided that the most appropriate way to answer the first research question was to carry out a document analysis of the reports that were released by the test developers in response to the requirements in the Standards. This documentation should serve multiple purposes according to Ferrara & Lai (2016), including supplying a thorough and comprehensive description of the test with details of the technical information regarding its psychometric characteristics. It should give sufficient information to support test users in selecting tests and using scores. The documentation should provide explicit guidance to support users in making appropriate interpretations from test scores while cautioning against identified uses that are inappropriate, particularly those that can lead to negative consequences. This documentation should also provide users with the basis of the validity argument for the proposed test interpretations. It should describe the test development procedures that have been employed and the intended test takers should be identified. Details of the quality control procedures used in the development of test items should be given, as well as information about scaling and equating procedures, reliability and measurement error. This documentation was analysed to identify the processes used to validate the

inferences for the selected tests. According to Fitzgerald (2007), document analysis can be used to analyse information which is difficult to obtain by other methods such as interview. In this case, it is unlikely that I would be able to get direct access to the measurement professionals involved in the test development. As stated by Creswell (2012), such documents have the potential to provide a valuable source of information and this can help to gain more understanding of a phenomenon. Document analysis involves systematically reviewing or evaluating documents according to Bowen (2009), who identifies manuals and organisational reports such as those that would be produced by the testing agencies as examples of the type of documentation that may form the basis of such analysis.

To answer the second research question, I decided it was necessary to interact with teachers from international schools. This would allow teachers to describe their practices in using the information from score reports in detail. Interviews or questionnaires could have been used to gain this information. It would have been possible to send questionnaires to a larger number of international schools. This may have resulted a greater representation of the experiences of international school teachers. However, there is no guarantee that people would respond and so it is possible that few or no responses would be obtained. Interviewing was deemed to be the more advantageous technique for this research. Interviews can be more time-consuming than administering questionnaires and there is the danger that the information is not as focussed as it would be by using a questionnaire. However, interviewing gave teachers the opportunity to more detailed description of their experiences and practices (Creswell, 2012). It enabled me to direct questions to

elicit additional information and to clarify my understanding of the particular challenges that these international school teachers faced.

3.5 Procedures

To identify the testing that was used by the international schools in the Kanto Plains area, a review of their websites was carried out. The standardised tests that were identified were the International Schools Assessment (ISA), the Iowa Assessments, the Preliminary SAT (PSAT), and the Measures of Academic Progress Test (MAP). Websites also mentioned that schools carried out examinations for the Advanced Placement (AP) Program and the International Baccalaureate Diploma Program (IBDP). These tests are terminal examinations for the programs that are described by Hayden (2006) as adapted or integrated. The other four tests are those that would be used to monitor learning over time. I decided to restrict analysis to those tests that would be used in created curriculum programs (Hayden, 2006). Consequently, APs and IBDP tests were excluded from the analysis.

I aimed to analyse documentation from the identified standardised tests. As acknowledged by Ferrara & Lai, (2016) the supporting information may be given in a range of documentation. Therefore, I sought included the technical documentation, any manuals that were provided to support test administration, and any information that was released to test users to support the interpretation of test results. In response to the Standards, (AERA et al., 2014), test developers generally produce technical manuals that detail the procedures that have been carried out to ensure the validity of proposed uses and interpretations. These contain the information regarding the domains that will be tested and the standardisation processes that are carried out.

They are generally made available on the developer's website. I carried out internet searches to find copies of the technical documentation relating to each of the identified tests. Meanwhile, test publishers produce instruction manuals which include details of how the test should be conducted and the script that is to be used when administering the test. Test publishers also produce documentation giving details of statistical measures and diagrams will be included in the score reports. These documents may be sent to users after the test has been marked and scores have been calculated, or they may be made available through the testing agency's website. I collected the documentation provided test publishers from schools. I also used internet searches to get hold of documents that I could not obtain through the schools. A list of the documentation that was identified for each test is given in appendix 2.

We are advised by Scott (2014) that the quality of the documents should be established using four criteria: authenticity; credibility; representativeness; and meaning. Establishing he authenticity of public reports, such as those used in this research, is consider unproblematic by McCulloch (2004). The documents used here would also be considered credible as they are official documents resulting in the authors having little choice in the information that is provided. Representativeness is assessed by considering survival and accessibility according to Scott (2014). Neither was considered to be an issue here as the documentation has been produced relatively recently and is generally made available through the test developer's website. As the ISA is not a US test, it does not conform to the same requirements so care had to be taken to access all appropriate documentation so that the necessary information could be obtained. Finally, Scott (2014) tells us to be able to analyse

documents, we must understand the meaning of the information provided in them. According to Fitzgerald, (2007) this requires an understanding of key words and phrases. Previous statistical studies and experience of working with the feedback from standardised testing has enabled me to understand many of the terms used. However, I also used reference texts to clarify meanings when I was uncertain.

I chose semi-structured interviewing to gain understanding of teachers' current positions regarding data use (Rubin & Rubin, 2012). For semi-structured interviews, Kvale (2013) tells us that a sequence of themes is identified and open questions related to these themes are developed. Kvale (2013) states that this preparation allows the interviewer to lead the subject towards these themes without steering them towards specific opinions. Research such as that by Mandinach & Honey, (2008); Means, Padilla, DeBarger, & Bakia, (2009); Pierce & Chick, (2011b); Schildkamp, Lai, & Earl, (2013) highlights the increased use of assessment data in educational practice. It was decided that the first theme would ask about the school's policy for data use. A second theme asked teachers about how they used data (Pierce & Chick, 2011b; Popham, 2011). The need for training and professional development were highlighted the literature (Bruniges, 2011; Gallagher et al., 2008; Means et al., 2011). Therefore, the final theme focused on the training and professional development. As suggested by Creswell (2012), an interview protocol was developed to outline the topics that would be covered so that themes could be carefully ordered during the interview (See Appendix 4). By identifying the themes and preparing the topic outline, I was able to maintain the focus on the research topic, but I also had the opportunity to follow up on answers that were particularly interesting and relevant to the research theme. Cohen, Manion, & Morrison (2011)

suggest starting with less threatening questions to put respondents at their ease. Therefore the first questions asked teachers to give factual information including that related to the third theme. Once it seemed that the teachers were more relaxed about being interviewed, they were asked questions relating to the other identified themes.

The sampling method used was purposive as the teachers that were selected were those most likely to be expected to use the information from standardised testing to inform their practice. The standardised tests identified above focus on skills in mathematics, reading and writing however, some included testing or analysis that was related to skills in science and social studies/humanities subjects. Even in instances in which science and social studies are not part of the testing, skills in reading and writing are also likely to impact on how these subjects are taught. Therefore, principals and teachers of English, mathematics, science and humanities/social studies were identified as those most likely to be required to use information from this testing and to apply the feedback to their practice. In school A, participants were identified using snowball sampling (Cohen et al., 2011). Once permission had been given for the research to be carried out, the headteacher was asked to suggest a member of the faculty that they thought would be able to provide information regarding the of use information from standardised testing data in the school. This was done because I did not have any contacts within the school who would be able to suggest suitable candidates who would agree to be interviewed. The headteacher suggested an administrator. At the end of their interview, the administrator was asked if they could supply the names of other teachers from the identified curriculum areas who they thought would be able to give more information about the practices within the school. Two more participants were

identified. One participant was the departmental chair of English and humanities, while the other was departmental chair of mathematics and science. Again, they were asked during their interviews if they could suggest other teachers who could be contacted to take part in the research. However, no further teachers agreed to be interviewed. The interview participant in schools D was recruited by teaching colleague who had professional contacts within the schools. In school C, I arranged to interview a teacher who I had met on a professional training course regarding standardised testing and who I knew was trying to implement data-driven practices into his school. He suggested that I talked to an administrator at his school. It was my intention that the interviews were one-to-one. However, when the interview started, I realised that the administrator and teacher planned to be interviewed together and I was unable to resolve this issue. In school C, I had the email addresses of the teachers of the subject identified above. The teachers who were interviewed were those who responded to an emailed invitation to take part in the research. All of the methods used to select participants had the potential to introduce bias. With the two departmental chairs being nominated by their administrator in School A, it is possible that the teachers that were chosen were those that would express support for the school policies. It was hoped that the confidential nature of the research would mean that these teachers felt that they could express their opinions honestly in the interviews. Meanwhile, it is likely that those who responded to the email invitations could be teachers who have strong opinions regarding the use of standardised testing. Other methods of recruiting could have resulted in fewer participants. Presenting the views of a smaller number of participants would also run the risk of presenting a one-sided or biased picture of the use of standardised testing in international schools. This research aimed to give a

picture of the use of standardised testing information and was therefore exploratory in nature. Given that the sample is made up of a small sample of teachers based in one country, it is acknowledged that the research would not be representative of the practices of all teachers.

The individual interviews lasted approximately 45 minutes each, while the interview with the two teachers lasted about 75 minutes. Teacher 3 commented that there was no expectation to use testing data within her school. Although there was some discussion about training to use data, the interview was shorter as the participant felt that she could not provide any relevant information. Initially, I had planned to carry out face-to-face interviews. However, because of the distance between the international schools that participated in the study and the busy schedules of the teachers who agreed to be interviewed, the ability to co-ordinate schedules proved difficult. Consequently, four of the interviews were conducted using Skype. The remaining five were face-to-face. It is acknowledged by Lo Iacono, Symonds, & Brown (2016) that the use of video applications can overcome scheduling issues such as those experienced here and Nehls, Smith, & Schneider (2014) comment that online interviewing is a viable option to the face-to-face interview. I took account of Krouwel, Jolly, & Greenfield's (2019) reminders and confirmed that the interviewee was comfortable with using Skype and would have access to a suitable internet connection. Researchers such as Sullivan (2012) and Cohen et al., (2018) comment that the more impersonal setting of the online interview means that the interviewee may be subject to distraction, say if others are in the room. All of the interviews were scheduled well in advance and at the convenience of the teacher. Teachers were asked to be in a comfortable place where disruption could be minimised. It is

noted by Deakin & Wakefield (2014) that the interviewer's ability to sense non-verbal cues is reduced when using video applications. At best, they can only see the interviewee's face when the interviewee is facing the camera. However, it was decided that this would not impact the interview results significantly as the analysis of the interviews was semantic (Maguire & Delahunt, 2017) The focus of the analysis was on the meaning of what was being said and the inability to see gestures was not seen to be a disadvantage in this case. In comparing research using face-to-face interviews with Skype interviews, Krouwel et al. (2019) used comparison of the number of statements within each code to confirm that face-to-face and video interviews gave the same width and depth of information. Similar analysis was carried out here to ensure that the different modes did not impact the level of information that was collected.

Interviewing allows for the inclusion of methods designed to confirm the understanding of information obtained according to Kvale & Brinkmann (2009) Based on their suggestions, the researcher included questions that summarised and paraphrased the information gathered. This gave the interviewee the opportunity to rectify any misinterpretations that I had been gained during the interview process. It is also important that there is an accurate record of the interviews to maintain the depth and detail of the phenomenon studied for later analysis, participants were asked to give their permission for the interview to be recorded (Rubin & Rubin, 2012). Where permission was given recordings were made of the interviews. Audio recordings was made for face-to-face interviews while a video application was used to record the four Skype interviews.

3.6 Data analysis

According to Merriam & Tisdell, (2016) document analysis can give descriptive information as the data can be analysed in the same way as that which is gained from interviews and observations. The aim was to identify the information relating to the validation processes and so content analysis was used (Denscombe, 2017). The analysis focused on the facts and information that are described within the text. It looked for evidence using categories identified in research by Becker & Pomplun (2006). These categories related to curriculum and populations characteristics such as language and culture as these were the challenges to validity that were identified in the literature review. The categories chosen were an overview or the purposes of the test; technical characteristics validity evidence; and score reporting and research services. Becker & Pomplun's (2006) work was used to compile a list of questions that would be answered during the documentary analysis.

Table 3.3: Categories for the document analysis

Category	Information sought relating to the category
Overview of the test	What is the purpose of the test? What is the history of the test? What recent changes have been made?
Description of the test	What areas are being tested? What content standards are used? What types of items will be used? How many items are there? How are items written and reviewed?
Technical characteristics	What test-taker samples were used in developing the test? What was the composition of the sample? What evidence is there relating to performance of diverse subgroups such as English Language Learners?
Information on score interpretation and use	What are the appropriate uses of the scores?

These questions are listed in table 3.3 above. The information relating to each of the above questions will be presented and then used to identify any validity concerns relating to the use of the testing for making inferences about students in the international school population.

For the interviews, a thematic analysis was used. Because I had identified themes in preparation for the interviews, this would be considered a deductive thematic analysis (Nowell, Norris, White, & Moules, 2017). The transcribing of interviews is identified by (Braun & Clarke, 2006) as the first stage of analysis. Each recording was transcribed by the researcher immediately following completion of the interview. Once the interviews were transcribed, I uploaded the transcripts into Nvivo software. I then read each interview transcript to familiarise myself with what had been said (Nowell et al., 2017). During a second reading, open coding was used. As Braun & Clarke (2006) acknowledge, analysis is not a linear process. Therefore, repeated readings of the transcripts were carried out to make sure that all passages related to a coding were identified. Nodes were created in the Nvivo software and used to collate passages which had a similar focus. These nodes were then combined to form themes and subthemes (Braun & Clarke, 2006). The coded themes were read to check that that all extracts were appropriate for the theme and any that were incorrectly placed were recoded. Four themes main were identified. The themes and subthemes are shown below in table 3.4. Once the coding and assigning to themes had been completed, descriptive passages were formed for each subtheme. Extracts were chosen from the interviews and included within the passages to illustrate the points raised in the analysis.

Table 3.4: Themes for interview analysis

Theme	Category	Subcategory
School factors that affect data use	Initiatives related to data use	Initiatives to get teachers looking at data more.
		Initiatives where data is analysed to implement curricula change
		Initiatives centred around classroom data
		Use of data teams
	Expectations schools have for data use	Expectations for teachers to use data
		Marketing the school
Factors relating to teachers' use of data	Teachers' perceptions about how they are expected to use data	
	Training in the analysis of data	
	Teachers' concerns about data use	
	Perceived obstruction to using data	
Issues specific to teachers in international schools	Language Issues	
	Transience of students	
	Other issues that were identified	
Changes that would facilitate data use	Presentation of reports	
	Analysis of skills	
	Technological requirements	

3.7 Ethical considerations

During the research, it was important to be aware of ethical concerns from the very beginning of the research process right through to the completion of the final report (Kvale & Brinkmann, 2009). We are told by McCulloch (2004) that there is less potential for ethical issues to arise with documentary analysis as it does not involve

human participation. Indeed, Wellington (2016) suggests that public documents should be considered suitable for critical analysis. However, I had to be aware of legal dimensions including copyright law when using the documents in the research (McCulloch, 2004). As the interview research was carried out on human participants, it was a mandatory requirement to seek permission from the Durham University School of Education Ethics Committee (See Appendix 3). The application for permission to conduct the research required the interviewer to identify potential ethical risks and give information detailing the measures that would be taken to ensure these were mitigated. The first ethical consideration involved gaining permission to use the schools in the research. It is essential to seek permission from the gatekeeper of the organisations. (King & Horrocks, 2010). Therefore, I contacted the heads of the international schools identified in the sample. I provided them with an information sheet giving details of the research (See Appendix 5) and asked permission to interview teachers in their school.

Once permission had been obtained from the headteachers and participants were selected the next concern related to informed consent. I emailed potential participants to invite them to take part in the research. In the email, I provided them with an information sheet (See Appendix 6) which gave them an overview of the research (King & Horrocks, 2010). This explained the aims of the research and the methods used. They were informed of the requirements of participation including the time commitments. A statement acknowledging that participation in the research was voluntary and that withdrawal from the project would not result in negative consequences was also included. As well as informing participants of their rights, they were asked to give permission for the interview to be recorded. Participants

were asked to sign a permission form (See Appendix 7) before taking part in the interview. This gave permission for the use of the interview in the research and for the interview to be recorded. In face-to-face interviews, the form was signed at the beginning of the interview. Where interviews were conducted using Skype, the participant was asked to email or fax a copy of the permission form back to me before the scheduled date of the interview. The participants who were contacted using Skype were reminded of their rights and that the interview was being recorded. They were asked to confirm their permission verbally at the beginning of the interview.

Participants were also told that measures would be taken to ensure that the recording and the transcript made of the interview would be protected. The information sheet also detailed arrangements that would be made to ensure that participants were not identifiable both during the research process and in writing up the thesis. Names would not be used on the transcription of interviews and recordings would be deleted as soon as the relevant information was obtained. The written work would not use the names of the participants and their schools would not be named and no information would be published that would allow for the identification of participants or schools.

3.8 Reliability and Validity

According to Bryman, (2015) judgement of the reliability and validity of a qualitative study is based on an evaluation of the trustworthiness of the research. He identifies four criteria that need to be considered when establishing the

trustworthiness of a qualitative study: credibility, transferability, dependability and confirmability.

Credibility, according to Merriam & Tisdell (2016), is the equivalent of establishing internal validity in quantitative research and questions whether the research findings capture reality. For document analysis, Scott (2014) tells us that credibility is established by assessing the accuracy of the documents used. As stated above, it was assumed that the documents from the testing agencies represented a truthful account of the procedures that were carried out to assure validity. Merriam & Tisdell, (2016) suggest that triangulation is the best-known method for establishing credibility in interviewing. Bryman (2015) states that triangulation requires the use of more than one data source in the study. Additional materials provided evidence of the responses given by teachers and administrators in schools A, C and D during this study. School A shared a copy of the school action plan relating to data use. In school C, the results of interim assessments relating to a reading program that was being implements were projected on a whiteboard throughout the interview. Further, examples of the data feedback that teachers were providing to the administrator were viewed during the interview. In school D, examples of the resources for the classroom assessment initiative were seen, alongside minutes from the meetings of their data team. Another way credibility can be checked according to Busher & James, (2007) is by cross-referencing of the information given by participants from the same school. Factual information given by staff from the same school was cross-referenced to make sure there was agreement.

Transferability is considered parallel to the external validity criteria for quantitative data and considers if findings would be applicable to other contexts. There are many other international schools who use standardised testing and so researchers investigating this phenomenon may be able to apply aspects of both the documentary analysis and the interviews to their particular situation. To help with this, descriptions are given for the context of the research although care has been taken to protect the identities of schools and teachers from the interview sample. It is acknowledged that all international schools have their own individual context and researchers would need to verify that the schools used here are similar to the context in which they were working.

Dependability is parallel to reliability, which is based on the ability to replicate the research. As qualitative research deals with human behaviour which transforms over time, it is unlikely that replication will lead to the same results. Therefore, a detailed description of the decisions that were made and the data analysis process used in the study is given in this chapter so that the findings can be deemed authentic.

Finally, confirmability which parallels the objectivity requirement of quantitative research, requires that the researcher shows that they have not allowed personal bias to influence the research and its finding. This is associated with the reflexivity of the researcher. We are told by Cohen, Manion, & Morrison, (2011) that reflexivity is a central component of interpretative research. Reflexivity requires the researcher to reflect on their own assumptions and preconceptions regarding the research theme. I am a teacher in an international school who has a role that involves reviewing and sharing information from the score reports received after our students have taken

part in large-scale standardised testing. I have studied statistics at the postgraduate level. Because of this, the theoretical perspective of using the information contained in the score reports from testing seemed logical to me. However, through this role, I have become aware of some of the challenges that teachers face relating to the standardisation processes of the testing that is used and the ability to relate the information that is received to international school students. Rubin & Rubin (2012) encourage the interviewer to examine their attitudes so that they are aware of any bias. They can offset such bias when the questions are being written. The interview schedule (See Appendix 4) was written with the aim of keeping the language used for the questions neutral. It was reviewed by my supervisor before being used. Interviewing requires the researcher to play an active role and as an interviewer I also needed to be aware that my responses to the interviewee did not influence the answers that were given (Rubin & Rubin, 2012). I aimed to keep my facial expressions neutral. I tried to make sure my responses mirrored the statements that were being made and did not lead the interviewee. When transcribing the interviews, I reviewed the comments that I had made to reflect on the interactions that I had with the participants (Roulston, 2014). Kivunja & Kuyini, (2017) state that the balanced axiology of the interpretative paradigm assumes that there will be reflections on the researcher's values as they try to present a report that is balanced. In writing the analysis, I aimed to include comments that reflected both the positive and negative experiences of the participants. I included quotes from all participants with the aim that all voices should be heard (Findlay, 2002). Further, the interviewee's responses are included in the analysis so that any reader can evaluate my interpretations of the interview data.

3.9 Summary

This chapter described the use of the interpretative paradigm to collect and analyse data to answer the two research questions. The interpretive paradigm was the most appropriate for this research as the intention was not to test or prove hypothesis but rather to provide a description of how the information from large-scale standardised tests is used in four international schools, including the expectations that these schools place on teachers to use the information provided in score reports and the challenges relating to the use of that information.

The following two chapters will present the findings of the research relating to the two research questions. In chapter 4, the document analysis will be presented to address the research first question regarding the potential threats to validity that teachers need to be aware of when using large-scale standardised testing in international schools. Chapter 5 will present the analysis of the interviews that were used to investigate how standardised testing is currently being used in four international schools.

Chapter 4

Results – part 1

4.1 Introduction

This chapter will present the results from the document analysis of the technical manuals from a sample of the standardised testing that are used by international schools. The analysis was carried out to investigate the first research question. "What are the potential threats to validity that teachers need to be aware of when using large-scale standardised tests in international schools?". By reviewing school websites, I identified the large-scale standardised tests that were being used by international schools in the Kanto Plains area of Japan. The tests that were identified were the SAT suite of assessments, Iowa tests, ISA testing and MAP tests. As stated in chapter 3, the Standards (AERA et al., 2014) suggest that testing organisations should share information regarding their validation processes and Becker & Pomplun (2006) list documentation including the technical manuals, test manuals and user guides as sources that give the test publisher the opportunity detail the evidence necessary to support and defend the test. The technical manuals that were released by the test developers that are responsible for writing each of the tests will be analysed. Additional information manuals that have been released by the test publishers will also be used. The processes that have been used to validate inferences that can be made from the test will be identified. The implications of the details of these processes for the use of the tests to make inferences in international schools will be discussed.

4.2 Format of the document analysis

As Becker & Pomplun (2006) state, careful and thorough documentation which includes explanations of the processes used in creating the items and the test should occur through all processes of test development as it is the interaction between the items and samples that "provides the basis for validating test score interpretation and use" (p. 713). In their work, Becker & Pomplun (2006) identified categories of elements that should be included in the technical manuals provided by the test publishers. As stated in the methodology, categories identified in Becker & Pomplun's (2006) work will be used to give summary of the procedures used to validate the score use and interpretation as described in the technical manuals. Other documentation that is released by the test publishers, such as he instruction manual given to test administrators will also be considered.

The document analysis will start with a description of the information that should be given in each of the categories that Becker & Pomplun (2006) identify for inclusion in the technical manuals. These are the purpose of the test, description of the test including content standards, technical characteristics, information on score interpretation and use. It will then use the categories to relate the information given in the documentation released by each of the testing agencies. Once the description of the tests has been given, the implications that this information could have on the validity of score interpretation and use in international schools will be discussed.

The technical manuals should start with a description of the purpose of the test which includes information on the history of the testing program with a particular

emphasis on any recent changes made according to Becker & Pomplun (2006). According to University of Iowa, (n.d.-a),

> The purposes of a test define how the test should be used, who should use it, who should take it, and what type of interpretations should be based on the results. (p. 1)

Becker & Pomplun (2006) tell us that validity judgements will be based on the test's purpose which will also be used to define the testing specifications.

Following the purpose, Becker & Pomplun (2006) suggest that tests should include the testing specifications. For large-scale tests, this should include the content as process strands while ability tests may include the cognitive components that will be included in each subtest. According to the Standards (AERA et al., 2014), the first part of the test design is for the test developer to make a framework that thoroughly defines the domain or construct which is to be measured and describes the processes and diagnostic features that will be used. The Standards (AERA et al., 2014) inform us that the test content encompasses wording and format of the items and so this section should describe how items are written, reviewed and selected as well as the performance of different subgroups of test takers.

The technical characteristics should include the test and item-level statistics along with information regarding equating and scaling as well as that regarding reliability and measurement error. The sample of test takers used in the development should be described here as these determine the populations that score interpretations are valid for. A description of the sampling procedures should be included. Information on the characteristics of the sample including the participation of groups such as English-language learners should be given. Any field testing that is undertaken

including that to determine the generalisability of inferences across groups should be described. The technical characteristics should include information regarding the scales that will be used and any linking procedures along with descriptions of scale characteristics. A description of the reported scores should be given as well as relevant information such as performance levels that are needed for making criterion or normative interpretations. There should be information regarding appropriate score uses for individuals and groups along with any necessary warnings regarding the use of subtest scores or the interpretations of extreme scores.

This section has described the categories that will be used to carry out the document analysis. The four identified tests will now be analysed starting with the SAT suite of assessments.

4.3 SAT-related assessments

The SAT has its roots in the nineteenth century when the College Entrance Examination Board (CEEB) was formed by a group of leading American universities who were concerned by the lack of universal standards to assess whether students were ready to embark on college level courses (Valentine, 1961). In 1926, the CEEB went on to develop the precursor to the current SAT, which was known at that time as the Scholastic Aptitude Test (Stickler & Breland, 2007). Its purpose was to expand access to university education in the US beyond its traditional domain of white, upper class, Protestant males (College Board, 2015b; Stickler & Breland, 2007). By the 1940's, the majority of private universities in the north-eastern United States adopted the test as part of their admissions requirements (Eduers, 2009) and it became recognised as the test that was needed to gain entry to college (Baird, 2017).

In 1994, the name of the test was changed and it became known only by the initial letters, SAT.

The SAT has undergone many changes during its history. The most recent was in 2016 in response to concerns that the test had become too detached from what was happening in high school, relied on understanding tricks rather than showing student knowledge and was unfair as wealthier students were enrolling in expensive preparation classes which skewed the results in their favour (Balf, 2014 Gumbrecht, 2014). The president of College Board, David Coleman, wanted to design a test that measured student achievement based on the skills that students were being taught in school and that would be necessary in post-secondary education and their working life (Balf, 2014; Gumbrecht, 2014; Letukas, 2014). Research by College Board also showed that 57% of students taking the SAT in 2013 needed to undertake remediation courses before they were capable of succeeding in college-level entry courses (College Board, 2014).

Students generally choose to enter the SAT themselves as part of their application process for university. However, there are a suite of related assessments that students take as indicators of their progress during high school. Schools are responsible for entering their students for these other assessments. These assessments are the PSAT/MNSQT, which students generally take on a prespecified date in grades 10 and 11, the PSAT 10 which is designed to be taken at a school-appointed time in the spring of grade 10, and the PSAT 8/9 which can be taken at intervals decided by the school during grades 8 and 9. The PSAT assessments are designed with the intention that students, teachers and parents can track the progress

of students as they progress through middle and high school. Students and teachers receive score report information after any of the tests are taken, and it is expected that students will share results with their parents. Schools are provided with resources to help them to explain how to interpret the results to students and parents and video presentations are available on YouTube.

The assessments are 'connected by the same underlying content continuum of knowledge and skills' (College Board, 2015b, p. 8) and are aimed at monitoring students' progress throughout their later years compulsory schooling (College Board, 2015d). They are intended to make sure the student is on the path to attain the skills that are identified as important for college or career readiness and to pinpoint the skills that they need improve upon so that they are better prepared when they take the SAT (College Board, 2015a, 2015b). To aid this process, the tests measure the same content and skills but adjustments are made to allow for the expected attainment of students in different grade levels (College Board, 2019a).

4.3.1 The purpose of SAT-related assessments

According to the technical manual released by College Board,

> the primary purpose of the SAT Suite of Assessments is to determine the degree to which students are prepared to succeed, both in college and in workforce training programs (College Board, 2017b, p. 2).

With that aim in mind, the suite of tests aims to determine if students have developed the skills in reading, writing and mathematics which they have identified from research as being essential for success in postsecondary education (College Board, 2017b). According to College Board (2017b), that research indicates that post-secondary institutions value students gain deep learning of a small set of topics

rather than shallow learning over a wider range of concepts. They hope that the SAT suite of assessments will encourage students in high school to take more challenging courses, particularly those students who are considered to be from backgrounds where taking such courses may not be the usual choice. Among the key features of the testing that College Board identify are that problems are grounded in real-world contexts incorporating the US founding documents and documents from the great global conversation about civic rights and that the mathematical concepts that are included have been identified to be the most important for use in postsecondary education.

4.3.2 Description of the test including any content standards

Each of the assessment in the suite comprises of a test in reading, a language and writing test and a mathematics test (College Board, 2019b). All questions for the reading and the language and writing tests in all the tests are multiple-choice. The mathematics test is split into a no calculator and a calculator section and each part involves mostly multiple-choice questions with a smaller number of questions where the student has to work out the answer and indicate their response by writing the answer and shading circles to indicate the digits in the question space on the answer sheet. All answers are computer marked. The breakdown of questions and timing is given in table 4.1. They state that tests place emphasis on key content features which are identified under the titles Words in Context, Command of Evidence, Expression of Ideas, Standard English Conventions, Heart of Algebra, Problem Solving and Data Analysis, and Passport to Advanced Mathematics. While College Board have stated that they have used evidence to identify the skills that are essential for success in college, and they give examples for progression in some content areas

(College Board, 2019e), they do not give any breakdown of the actual content that makes up the areas that are identified and teachers can only identify skills that are included through analysis of the actual tests.

Table 4.1: Overview of number of questions and timings for PSAT assessments

	PSAT/NMSQT	PSAT 8/9
Reading	47 questions / 60 minutes	42 questions / 55 minutes
Writing and Language	44 questions / 35 minutes	40 questions / 30 minutes
Mathematics (no Calculator)	17 questions / 25 minutes	13 questions / 20 minutes
Mathematics (no Calculator)	31 questions / 45 minutes	25 questions / 40 minutes

(College Board, 2019e)

4.3.3 Technical characteristics

The tests are on a common vertical scale so that a student should get the same score regardless of the version of the test that the take. Tests are built using a common set of statistical specifications. This means that a student's score should have the same meaning regardless of which of the tests they take as the tests should be comparable in terms of content and score reliability. The maximum and minimum scores available on each assessment are different to allow for the increased level of difficulty of the tests because more difficult concepts are tested in the SAT but are not included the PSAT 9.

Students receive five different types of scaled scores for the test and each is designed to serve a different purpose. There are two section scores one for evidence-based reading and writing which is derived from scores in the reading and writing sections, and a one for mathematics. This is combined to make a total score for the test. The

scores from the SAT are used by the student for admissions purposes and in the PSAT tests, the scores are colour-coded to indicate whether the student is on track to attain a score deemed to show college readiness when they take the SAT. They are given the three test scores which are equated scores designed to measure growth if students take multiple tests in the sequence. Selected questions from each test are used to derive cross-test scores which show the student's ability to apply core skills within the academic contexts of Science and History/Social Studies. There are then subtest scores for each of the key content features derived either from the individual tests or across the reading and writing tests which are designed to help student identify areas of weakness so that they can work on strengthening their skills.

The scales for the SAT and the related assessments have been developed using what is described as a nationally representative population of college-bound students in the relevant age range for the individual assessment. In the case of the PSAT/NMSQT, the population is composed of grade 10 students while for the PSAT 8/9, grade 9 students are used. Pre-testing was also carried out on a sample of students who are considered to be representative of the population of interest. Differential item functioning (DIF) analysis is carried out to make sure that being a member of different population subgroups is not related to differences in performance on specific items. The ethnic subgroups that are identified within the technical document are White, Alaskan Native, Asian American, Black or African American (College Board, 2017b). While the technical documentation does comment on DIF analysis based on ethnicity, no mention is found of carrying out DIF analysis with ELL students.

The technical manuals comment that measures designed to ensure that their testing is fair to all test takers include having all questions reviewed by secondary and post-secondary teachers drawn from across the nation. They are chosen because they understand the student population that testing is being designed for. This enables them to make sure that questions are unambiguous, clearly stated and accessible by all students within the testing population. Their responsibilities include ensuring that the language included in test questions is widely accessible by choosing generic terms rather than those that could considered to be region-specific in the context of the question. They are also expected to exclude terms that could be considered foreign, dialectical, idiomatic or slang as students are likely to have a different level of exposure to them according to where they live and their socioeconomic or ethnic background.

4.3.4 Information on score interpretation and use.

Data from all testing is released to teachers via an online portal while students are given access to feedback on their performance through an online portal as well as being given a paper version of this feedback. The score reports that have been developed to accompany the assessments give information about appropriate uses and interpretations of the scores. The standard error of measurement (SEM) is used to give students score ranges which tell them about how they would be expected to perform in repeated sittings of the tests if no further learning had taken place. Two percentile ranks are displayed on the report; one which shows how their scores compare to students who typically take the test while the second compares their score to the projected scores of all US students in their grade, including those who would not typically take the test. Benchmarks for each of sections in the PSAT

assessments are designed to show whether students are on track for attaining the SAT benchmark for college readiness. Scores are colour coded with green indicating they have met or exceeded the benchmark, orange that they are within a year's growth of attaining the benchmark for the test taken and red meaning that they are more than a year's growth away. The SAT benchmark is meant to indicate that a student has a 75% chance of earning at least a C grade in a college-level credit-bearing course in a related course i.e. history, social science, literature or writing for the Evidence-Based Reading and Writing (EBRW) score and algebra, statistics, precalculus or calculus for the mathematics score. The test scores and subtest scores are based on the average scores of test takers who have attained the scores necessary for college readiness.

4.3.5 Summary of 4.3

To summarise, the SAT suite of assessments is designed to determine a student's college readiness. It tests skills in reading, language and mathematics using mainly multiple-choice items. There are no specified content standards. The items are reviewed by US secondary and post-secondary teachers. Scaled scores from a US nationally representative sample are used. DIF analysis is carried out to compare different ethnic groups within the sample. In the next section, the categories will be used to analyse the Iowa tests.

4.4 Iowa Assessments

The Iowa assessments are produced by the Iowa Testing Program based at the University of Iowa. They were pioneered at the University of Iowa in 1935 when they were known as the Iowa Test of Basic Skills. Its developers were also

responsible for the introduction of the first optical scanning machine which led to the expansion of standardised testing in the 1950's and 1960's (Koretz, 2009). The testing was last updated in 2012 to align with the Common Core and at this time it was rebranded as Iowa assessments.

4.4.1 The purpose of Iowa assessments

The Iowa assessments have been developed to provide information to stakeholders that can be used to improve instruction and learning. Educators can use the tests to identify areas of strength and weakness and to monitor the growth of individuals and groups of students (Dunbar et al., 2015). The tests are written for students from kindergarten through to grade 12 and are assigned levels ranging from 5/6 through to 17/18, with the levels being designed to roughly correlate to the student's chronological age. At all levels, the tests measure skills in Language, Reading, Mathematics, Science and Social Studies. Additional subjects are also offered according to grade level. The tests are all composed of multiple-choice questions.

4.4.2 Description of the test including any content standards

The technical information that accompanies the program states that the testing is designed to reflect the goals of instruction in schools across the US (Dunbar et al., 2015) and therefore testing is designed to reflect the curriculum standards which reflect successful performance in that national system. Test design is based on the Common Core State Standards (CCSS) and the standards used by individual states in the US (Dunbar et al., 2015).

4.4.3 Technical characteristics

Test production begins with training educators who are then assigned to write items for the grade and subject area that they have experience with. These items are reviewed by content specialists within the testing agency to ensure accuracy, accessibility and fairness to all subgroups within the population. External panels of educators then review the items to ensure they are appropriate for the assigned grade level in terms of content. Following this, the items are assembled into field tests which are included within operational tests so that statistical analysis can be carried out to make sure the items are appropriate in terms of difficulty and discrimination and that they maintain the test's overall reliability. The final testing forms are reviewed by educators and additional experts. The choice of educators and reviewers is designed to represent various different groups and subgroups in the population, including English language learners. They check items for fairness and to make sure that there are no sensitive items included (Dunbar et al., 2015).

Questions and passages used were screened to ensure the reading complexity is at an appropriate level for the student population taking each test. Measures such as word and sentence length, unusual letter patterns and sentence cohesion were used to make sure that validity was not compromised by the reading level required to access the testing materials. Screening was also used to ensure that the testing did not require students to have particular life experiences or could experience bias by the amount of knowledge they need to bring to the test.

The main scale used in Iowa testing is the National Standard Score (NSS). It ranges from 80 to 400. It is a continuum and is used to indicate the achievement for

students from kindergarten to grade 12 in content domains including mathematics, reading and written expressions (Dunbar et al., 2015). The scale is vertically aligned to facilitate the measure of growth between assessments and allow for the comparison of expected growth with peers using a normative growth interpretation. To maintain the validity of the model, growth is measured on a learning continuum that aligns to content standards at an appropriate level of cognitive complexity for the development stage that the student has reached and the skills and content in each test build on those measured in the previous test. They comment that by using a vertical scale, teachers can know that changes in score represent changes in student achievement rather than a change in the test being used. The scale is also used to predict where a student should be in future instances of testing based in the score they have attained. By looking at the difference between predicted and actual score, value-added growth can also be calculated.

To set their scales and norms, the assessments were administered to large groups of students across the US under standard conditions. This was used to form a national probability sample which was designed to closely reflect the US population by ensuring that important groups of students were represented proportionally. Schools were selected for the sample by considering such criteria as region of the country, size of the district and whether they were public or private schools. Socioeconomic characteristics of the school were also identified using the school's title 1 status, which is a status that allows schools to receive extra funding if it has a large proportion of low-income students, and the number of students who were entitled to free or reduced school meals. The sample was also designed to reflect the ethnic composition of the school population. The results from the sample were used to

calculate means, standard deviations and reliability indices, to establish norms and measurements of growth, and to analyse skills and percentage correct data. The growth models were developed using two state-level cohorts of students but little information is given about that sample apart from saying that it is a large sample. Special norms were also developed for students enrolled in Catholic and private schools in the US

The sample did include students who were not native English speakers but depending on how long the student had been in an English-only classroom, schools could choose whether these students would take part and if they did, whether they would be given accommodations such as extended time or use of a translation dictionary. Less than 5% of students in the sample were identified as English Language Learners, and of these, about 15% received accommodations (Dunbar, Welch & Hoover et al, 2015).

As part of the test design, longitudinal score scales designed for measuring growth in achievement were defined using scaling methods. Comparability of scores on parallel forms of the test was established using equating methods, while the basis for measuring strengths and weaknesses was established by looking at long-term trends in achievement and national performance. Their growth model and NSS were developed by gaining a nationally representative sample through spiralling scaling tests within classrooms. This allowed them to work out the variability with subjects and grades and look at the relative achievement between grades. The scale was set so that median performance in the spring of grade 4 was assigned a score of 200 and

that median performance in the spring of grade 8 was 250. Expected grade-to-grade growth ranges were 18 from grade 1 to 2 and 5 from grade 11 to 12.

A National Grade Equivalent (NGE) scale was established. There is a warning that care must be taken as the score describes relative achievement within a grade-indexed metric. It does not indicate the grade level placement but rather that students are responding in the same way that an average student in the given grade level would perform.

Student results include a National Percentile Rank (NPR) which they describe as comparing student performance with a nationally representative sample of public and private schools across in the US. The most recent norms for the NPR were established in 2017 by identifying US school districts who has taken tests for at least two years and applying weightings so that the distributions matched those released by the National Center for Educational Statistics. The distribution was designed to match the demographic characteristics of the US school population in terms of socio-economic and ethnic composition. Again, there were ELL students included in the sample but where schools have ELL students, the schools could decide whether it was appropriate to include the ELL student based on the level of language acquisition that student has. Again, the school can provide any accommodations for the ELL students who took part.

4.4.4 Information on score interpretation and use

Iowa produces a number of different reports at the individual student level, as well as composite reports for classes, schools and districts. The reports include summaries

of the different attainment measures including the standard score, the grade equivalent, the national percentage rank and predicted college entrance exam scores. Alongside the summary numbers are side-by-side bar charts so that performance can be compared with national performance and a narrative that summarises what information can be derived from each performance indicator. Class reports give individual summaries of students as well as average values for the class. Reports are produced for use within schools and a printable version is designed to hand out to parents.

As with the College Board testing, the Iowa tests seek to determine a student's path to college readiness. They have established a predictive relationship between their tests and those that are used as college entrance examinations in the US such at the SAT and the American College Test (ACT). They note a particularly strong correlation between the results of the Iowa testing and the ACT and state that this is evidence that the tests assess the same achievement domains. They say that this supports the use of Iowa testing to predict the likelihood of students attaining or exceeding the College-Readiness benchmarks in the ACT. By using linking studies, they have established target scores for students in earlier grades so that they can establish whether they were on track to meet the benchmarks in the ACT.

Teachers are warned that for maximum precision, students need to be assigned to an appropriate test level as there is less precision in measurement if they take a paper that is too easy or too difficult for them. If the paper is too difficult then students may get higher levels as the effect of getting answers correct by chance will be more

significant, while if a paper is too easy, it will not allow the student to adequately show their skills.

4.4.5 Summary of 4.4

The purpose of Iowa testing is to provide information that can be used to improve teaching and learning. Tests are given in reading, language, mathematics, science and social studies using multiple choice questions. The test is based on the US Common Core State Standards. Items are developed by content specialists and reviewed by external panels which include reviewers from different subgroups within the population. Scaling is done using a US national probability sample. MAP testing will now be analysed using the given categories.

4.5 MAP Testing

Measures of Academic Progress (MAP) testing has been developed by the Northwest Evaluation Association (NWEA). The development of MAP testing is the result of a collaboration between researchers and educators from schools in Oregon and Washington who sought to design a test that helped to inform instruction.

All the questions used in MAP testing are in a multiple-choice format. Since 1985, MAP had taken the form of a computer adaptive test (CAT), which means that students are not given the same questions but rather that the questions they are given are based on whether they answered the previous question correctly. To start the test, they are given a question which is considered to be of medium difficulty. If they get the question right, the computer selects a question considered to be at a slightly higher difficulty while those who get it wrong are given a slightly easier question.

The testing continues in this manner, with a slightly more difficult question being given when the student's response is correct and an easier question being given in response to an incorrect answer. Because each cycle of testing continues by selecting an item closest to that ability, each repeat gives a slightly more accurate estimate. Testing is continued until a desired level of precision is reached in that it is considered the error in the estimate is a small as it can be for a particular testing session. By doing this, according to Thum & Hauser (2015), standard errors remain small across most of the achievement range within a grade level making the test suitable for use in growth models. With tests that have a fixed form, the standard error of measurement (SEM) is at its lowest in the middle of the distribution and generally increases for scores that are away from the middle which means that there is less precision for scores that are at extremes of the score distribution. Furthermore, when testing is used to make conclusions about the change in students' attainment over time, the error accrues. It is argued by NWEA, (2011) that the greater score precision resulting from CAT testing means that not only are better estimates of student's current attainment levels attained, but that this increased precision will mean that conclusions about changes in attainment over time will prove to be more reliable.

According to NWEA, (2011), more than 40 questions will typically be used in most tests and it is expected that each student will answer about half of the questions correctly in each testing session. It is highly unlikely for two students to complete the exact same form of the test and the test is designed so that once a student has seen a question they will not see it again in future testing sessions.

4.5.1 The purpose of MAP testing

The test is diagnostic in nature. It is designed to determine what students know at the time of testing and to track the growth of the student over time. According to NWEA (2011), the score reports can be used to describe performance or for planning for students or classes, to give normative comparisons of performance for individuals and groups of students, and to identify specific skills that students within a score range may find challenging.

There are a number of guiding principles that NWEA (2011) state they adhere to. They state that testing should consist of content that the student has had an opportunity to learn and should provide as much information as possible while being challenging and not wasting students' time. Reliable results should be returned to stakeholders in the shortest possible time and should provide information about the student's current level of achievement and their change in achievement between testing occasions. MAP testing can be taken up to four times a year and the results are available through an online portal within 24 hours of testing being completed.

4.5.2 Description of the test including any content standards

MAP tests are aligned to the content standards of each state or school district and the test blueprint includes the content standards that are used by the educational entity. They do offer an international version of their tests but this is not individualised to the curriculum of each international school. To make sure that testing does match content standards, software is used to look for matches between the key words and phrases used in the content standards for the particular entity to those in the item content descriptors for questions in the databank. A content expert will also validate

the matches and identify content standard areas where no items or insufficient items were identified in the automated search.

The writing of new test items for the test bank is done on a needs basis and new items are written by freelance content specialists who are usually either current or retired teachers or educational consultants. The content specialist is provided with a content guide specific to their area. Once items are written they are reviewed by item reviewers and content editors to ensure that they are fair to all test takers in that the questions do not use idiomatic English, are not based on sensitive topics and are not biased against particular groups, including those of different cultures or language backgrounds. The reviewer also makes sure that the question presents the concept in a way that is consistent to that used in the classroom, and that terminology and information will not become dated. This includes a check to see if the question is appropriate for use in question pools that will be used outside of the US school system. As will be described below, the final check on the quality of items is made by introducing them into operational testing. Items are reviewed at least once every five years, with the oldest 20% of items being rechecked for relevance and validity and items that fail to meet the appropriate standards being retired from use. Before the test is released for use, it undergoes a series of simulations to check that it is working as intended in terms of the accuracy of estimates of ability, particularly for those at the extremes of the scales, identification of weaknesses in the item pool, and the success of targeting from the pool for a given ability.

4.5.3 Technical characteristics

According to Thum & Hauser (2015), the questions from common content area used in MAP testing are calibrated onto the same scale using a Rasch model. This scale is based on calculations which estimate the probability that a student with a specific achievement level will answer a given question correctly and values range from about 140 to 300. It is stated by NWEA (2011) that the scale was deliberately chosen so that it could not be confused with the scales used in other forms of standardised testing.

When testing items are developed, they will be assigned a provisional difficulty level. New test items are introduced into live testing during the final stage of development and the level of difficulty is ascertained using students' responses. Calibrated items are included in operational tests in a way that is transparent to the students but these items are not used in calculating the student's score. Each item will be administered to 1000 students. It is stated by NWEA (2011) that by including items in the way they will be used when they are live, the effect of other factors on student responses to the question such as its interaction with other items or its position in testing, will be minimised. Because of the way in which scores are developed, the test characteristics are defined by the item difficulty values and NWEA (2011) state that both the calibration of item difficulty and achievement level estimates are sample free.

To ensure fairness amongst different demographic groups in the test-taking population, Differential Item Functioning was used to carry out a calibration analysis using data from six states in the US. This sought to carry out pairwise analysis based

on item difficulty between members of a reference group against focal groups formed from different demographic groups based on race and gender. While NWEA (2011) acknowledge that it would be ideal to include DIF analysis in their calibration process, they comment that the number of respondents from some demographic groups would be insufficient to make analysis reliable.

4.5.4 Score interpretation and use for MAP testing

As has been stated, results to MAP testing are generally available the day after testing has taken place. Results include the student's score in the test taken and a percentile which gives the student's position relative to the US population values. However, NWEA also provide means and standard deviations which are designed for use by populations outside of the US. For instance, they give summative data relating to regional school organisations such as the Eastern Regional Council of Schools (EARCOS), and where sample size permits, country averages are also provided. The RIT score is described as a continuous and an equal interval scale according to NWEA (2011), so differences between values are the same no matter where on the scale the student is and this is considered to make the scale reliable for the measurement of growth between testing occasions.

4.5.5 Summary of 4.5

MAP is a computer adaptive test which is designed to measure attainment and growth. There are tests for reading, language and mathematics. Tests are available for a number of different curriculums so schools can select the one that is the closest match to the school's curriculum. The questions are multiple-choice. Students will generally complete about 40 questions. The questions are written and reviewed by

specialists who are generally teachers or educational consultants. These questions are then tested in live test sessions and calibrated onto a Rasch scale which ranges from 140 to 300. DIF analysis is carried out on demographic groups based on race and gender. Comparisons of student attainment and growth are given against a US student sample but means and standard deviations are given for other populations which may be considered more relevant. The next test to be analysed will be the ISA.

4.6 International Schools' Assessment

The ISA was developed by the Australian Council for Educational Research (ACER) specifically for use in International Schools. According to their website, the ISA is taken by 90,000 students in 400 schools across 80 countries and 65% of the students taking the ISA were from non-English speaking backgrounds (ACER, 2019a). It is promoted as reflecting the frameworks used by the Programme for International Assessment (PISA) (ACER, 2019b). While ACER have been involved in the development of PISA testing, the ISA is not endorsed by PISA.

The test is taken annually, with schools choosing to run the test either in October, February and May. Testing takes place across two mornings and comprises of a mixture of multiple choice and open-ended questions for the mathematics and reading sections, and two essays prompts are used to assess the writing component.

4.6.1 The purpose of the ISA assessment

ISA was developed following discussion between ACER and the international school community in the East Asia region who identified the need for the

development of a test that was designed for schools whose student population was from culturally and linguistically diverse backgrounds. They required a test instrument that would provide quantitative and qualitative feedback. This would allow schools to identify areas for improvement and provide them with comparisons with populations that were considered to be more relevant to such schools (ACER, 2015). The test is designed to allow schools to monitor the reliability of their internal assessments to ensure that they are aligned to international expectations of performance. It is designed to provide diagnostic information which includes data on the level of proficiency of individual students. This allows schools to compare themselves against international benchmarks with the aim of stimulating improved achievement

4.6.2 Description of the test including any content standards

The ISA is a norm-referenced test according to Walker (2017). It is not developed for a specific curriculum but rather is designed to measure core skills (ACER, 2015). There are sections in mathematical literacy, reading, and writing, with writing being separated into argument and narrative. Mathematical literacy and reading are further broken down to show different content areas and different process aspects. For instance, mathematics content areas are quantity, shape and space, uncertainty and data and change and relationships, while the process aspects are formulating, employing and interpreting. There is also a science literacy test offered within the program for grades 7 through to 10.

4.6.3 Technical characteristics

The testing is designed with the knowledge that a significant number of students in international schools do not have English as their first language and according to Walker (2017) in 2015, 65% of the 73,000 students who took the ISA came from backgrounds where English was not the first language. ACER also acknowledge that the cultural experiences of students in international schools will be different to those of the majority of English speakers. They acknowledge that even though the context of a question is extrinsic to the knowledge and skills that are assessed by the questions, familiarity with the context can have a significant effect on performance. However, they comment that it is impossible to develop meaningful tests that are culturally neutral. Therefore, the aim is for cultural eclecticism in which tests include questions that cater for a wide range of cultural experiences. In this way, it is hoped that all test takers will find some contexts which are familiar to them while they also realise that they may find some the at are unfamiliar.

Papers are available for all grades between 3 and 10, and students must take the test that corresponds to the grade level they are currently working through. Schools can choose where to use the ISA within their testing program. For instance, in the sample of interviewed schools, one used testing for grades 3 to 8 and switched to a different form of testing in grade 9. Another had recently switched from doing the ISA with odd numbered grades to doing it with all applicable grades.

The tests are made up of both multiple choice and open-ended questions with at least 50% of the questions in each assessment area require the student to construct a response. There are paper and online versions of the tests, although the questions

asked within the versions are essentially the same and the only adaptations being made to accommodate the use of the computer.

ISA reports students' results to schools using a scale of 0 to 800 points which ACER (2015) say is based on the scaled scores that were developed for use in the PISA although they warn that the scores are not PISA scores. A score of 500 was considered to show the average proficiency in mathematical literacy and reading of 15-year-old students in OECD countries. Student achievement in each testing domain is also reported using a proficiency scales which uses ten described levels on the different aspects which are also related to the aspects that were used in the PISA testing. The scaling is vertical and ACER (2015) state that the scaled score that a student receives is not dependent on the grade level of the paper taken. This means that scales can be used to compare the performance of all students across all grades within a domain. Scores can be compared across years to track the performance of individual students as well as providing evidence of the impact of curricular change on student performance.

Questions are pretested by the students in the international schools. It is the expectation that all schools who use the computer version of the testing will run trial tests with their students in the two-week period before the actual tests take part. The teachers and administrators who run the trial tests are asked to provide feedback about any questions which have proved problematic and the progress that students have made in the allotted time (ACER, 2015). Teachers and administrators are also asked to give feedback after they have conducted the testing within their schools. They also state that they have ongoing consultations with teachers and administrators

through attendance at the regional conferences that are arranged for international schools, as well as by making visits to the international schools that take their tests. In-house reviews are carried out by ACER to supplement the feedback that is received from teachers during trial testing.

4.6.4 Information on score interpretation and use.

The results of testing are returned to the school about three months after the test has been taken. Scores are given for individual students. Averages are given for classes and grade levels. Analysis is also given of student performance on individual questions. Between-school comparison is provided in relation to four groups of "like schools" which are based on the proportion of students who are declared by the school to be from an English-speaking background.

4.6.5 Summary of 4.6

The ISA is designed to provide quantitative and qualitative feedback for schools whose populations are culturally and linguistically diverse. There are tests for mathematical literacy, reading, writing and science literacy. Questions are mostly student constructed responses. The testing is not designed for a specific curriculum. Comparisons are based on the whole population taking the test and to schools that are considered to have a similar proportion of students for whom English is a second language. Scores are reported on a scale from 0 to 800 with a score of 500 considered to show average proficiency for a 15-year-old student in the PISA tests. This section will now go on to consider how the different aspects of the tests analysed above relate to the international school population.

4.7 Testing and the international school population

When selecting a test for use, it is essential that education professionals determine that the test that is chosen has validity and reliability in the context in which it is being used (AERA et al., 2014). Tests may be used for a multitude of different purposes in educational settings and validity should be established for each of those uses (AERA et al., 2014). The majority of the standardised testing included in this document review are designed for use in national settings and so international schools need to verify the validity evidence to make sure that the test is suitable for its required purposes in the international school context. However, there are potentially significant differences between the populations that the standardised tests are designed for and those who will take the tests in the international school context. Even where testing is designed for the international context, the identification of subgroups within the testing population may impact the appropriateness of inferences that could be made. This analysis will now go on to highlight some of the potential concerns that arise when the tests mentioned above are used in international schools. The categories derived from Becker & Pomplun, (2006) will be used discuss these concerns.

4.7.1 Description of the test

This section will look at concerns relating to the test description. It will discuss the potential effects that differences in content standards can have on the validity of inferences in the international school context. The Standards (AERA et al., 2014) inform us that the inferences that can be made from a given test are dependent on the appropriateness of the content domain. They draw attention to making sure that the content domain reflects the content that is delivered and that students have had the

opportunity to learn. When determining if a test is fit for a purpose other than that for which it was originally developed, it is essential to establish that the content is appropriate. Neither the SAT suite of testing or ISA testing have a syllabus that gives specific details of the content domains used for testing. Rather they give general information about the concepts that may be included. For instance, College Board, (2019a) give content features which give an overview of the skills that may be included in the test. They also give content alignment within the different testing domains (College Board, 2019c, 2019d, 2019f). However, they do not list the actual skills that will be tested. For instance, they state that the PSAT 8/9 may require the use of 'common geometric equations' (College Board, 2019c) and could include 'introductory probability and statistics', but it does not give detail of the specific skills that could be included under these categories. So when they talk about 'introductory probability or statistics', do they mean finding measures of average and giving the probability based on outcomes for a single event, or are measures of variability expected to be understood and are combined events needed for probability? The vocabulary that could be used also needs to be specified. At what stage should students understand terms such as "mutually exclusive", "independent" or "correlation"? The level at which these skills are included also needs to be identified. Without clearly identified skills for each level of paper, it is possible that differences in scores could be a reflection of skills that have not been taught prior to testing rather than from true differences between the ability of students in the international school and the population for which the test is designed.

The ISA also tests core skills rather than being designed to measure the content of a particular curriculum. Again, broad categories that are used, such as 'Uncertainty

and data' and 'Shape and space'. This makes it difficult to know the exact skills that are included in the test's design and at which grade levels the various components that make up these categories will be tested. Using such a test for diagnostic purposes can be very difficult as when there is an area of weakness as it is impossible to ascertain whether the weakness is due to factors that require improvement within the school or because the skill has not been taught before it is being tested. Even when choosing a test with a defined content domain, such as is the case with MAP and Iowa testing, there is no guarantee that correct inferences will be made within the international school setting. The tests may not reflect the actual content that is taught or the grade at which it is taught within the school. Again, differences in attainment may reflect the differences in what the students are actually being taught and the age at which it is being taught rather than the students' failure to make progress in learning the elements contained within the test's content domain.

Iowa testing uses the US common core curriculum to define the content for testing. It will follow the scope and sequence defined within that curriculum. However this may not match the scope and sequence of the curriculum used within an international school which includes influences from the curricula designs of other nations. While English and mathematics are subject areas that are included in the curricula of most English-speaking countries, each country will have identified or prioritised different content for use in their curriculum. The sequence in which the skills are presented or categorised may differ within each of the countries. For example, in the US, the mathematics curriculum tends to be very topic based with students in different grade levels following courses with titles such as 'Geometry', 'Algebra 1' and 'Calculus'.

Australia and the UK follow more integrated course, with elements of geometry and algebra being included in the syllabus for each academic year of secondary study. Within these countries, there are also different expectations about the age at which a student is introduced to particular concepts. Therefore, skills included within a grade level could be very different and, as a result, testing coming out of that country could include questions based on a very different set of skills. Consequently, testing such as Iowa testing which is based on the US common core curriculum may not match the chosen elements of the curriculum which are included in the syllabus of the international schools in particular grade levels. This will again lead to difficulties in identifying the cause for differences in attainment within the school.

MAP testing does allow for different curriculum standards to be used in their design. Users can select the curriculum standards used by different US states. However, international schools who use MAP testing must choose tests from pre-designed options. They have to determine the standards which offer the best match to their curriculum. This can lead to similar concerns about the inferences that can be made as for the Iowa test.

Even where the content domain of the test matches the curriculum that is used in the international school, the order in which skills are taught within the international school may depend upon the particular nationality that is the main influence within the particular grade or school section, or it may be that changes in the pacing of the curriculum may need to occur because students are trying to develop English skills while also learning the curriculum content. Again, because large scale testing is not

sensitive to such changes, care needs to be taken when interpreting any inferences that are made about the rate of learning in the international school setting.

When the SAT was being redesigned, one of the ideas at the forefront of the minds of the consultants was to test the skills that would be imparted by the best teachers. It was expressed that such teachers in the US would be determined to give a deep knowledge and understanding of how the founding documents of the US have inspired conversations on dignity, justice and freedom. The technical manuals explaining the SAT assessments emphasise that the reading assessment will be based on the US founding documents such as the Declaration of Independence and the Bill of Rights, and the documents that they consider to have inspired conversations regarding civic life. However, international schools are unlikely to focus on US history in the way that schools based in the US would. Therefore, students in the US are likely to be at an advantage as they will have an opportunity to analyse such documents in their classes. This will give them more preparation for doing this in the test. Furthermore, the language used in such documents may be different in the level of formality or use more antiquated forms of English. While College Board have stated it is not necessary to have previous knowledge of the documents to be able to answer the questions that will be included in the test, students who have studied the documents in their classes may be advantaged by their familiarity with them.

This section highlighted the potential for differences in the content that is being tested and that which is taught in the international school context. This included the potential for the international population to be less familiar with sources used for

questions in one of the tests. The next section will consider concerns relating to the technical characteristics of the tests.

4.7.2 Technical characteristics

There are a number of technical features which international teachers need to be aware of when considering the results from standardised testing. Differences between the population that testing was designed for and the international school population in terms of factors such as linguistic and cultural background can be a source of construct irrelevant variance. The next section will discuss the measures that have been taken by the test developers to overcome these factors. Areas where these measures may be insufficient to prevent construct irrelevant variance to occur will be identified.

In norm-referenced testing, the standardisation sample is vitally important and therefore a responsibility is placed on test publishers to provide adequate information about the constitution of the standardisation sample. Part of the process of identifying suitable testing for an educational setting is making sure that the sample is adequate so that it is appropriate to use comparisons between test takers and the standardisation sample. Characteristics such as the demographics of the standardisation sample in terms of such things as age, gender and race (Reynolds et al., 2010) should be verified to make sure that they allow for appropriate use of the normative data that is available from the test (AERA et al., 2014). Where testing is designed for specific groups, construct-irrelevant variance can be introduced when that testing is used with a different group of students, and test inferences may not be

generalisable if the normative group is not appropriate in the new context (AERA et al., 2014).

The SAT suite of assessments, the Iowa assessments and MAP testing were originally designed for use in US schools and so the procedures that were used to validate the inferences were based on samples of US students. However, the demographic characteristics of the student population in the US differs markedly to the populations of students in international schools. The majority of students in the samples are first language learners who have been born and raised in the US. While there are second language learners included, they will be a small minority within the sample. They are learning in the dominant language of the country and their experiences of the language will not be solely based in the classroom but will include language learned as part of their socialisation. This is in contrast to students in international schools where many students will be bi- or multi-lingual learners. They come from families where the home language is not English and they use a different language in everyday social situations as English is not the primary language of the country that they are living in. It is likely that their only interaction with English will be within the classroom. The ELL population in the US are also more likely to be from low-income families and the impact of being a second language learner in assessment may be confounded with the effects of socioeconomic factors on learning (Schwabe et al., 2016). The ELL population in international schools are more likely to belong to higher socioeconomic groups who do not suffer from the lack of educational opportunities that may be experienced by ELL students in the US. Therefore, there are likely to be differences in the reasons for their mistakes in testing.

While there are demographic categories that international school students might identify with included in the sampling frame, they are not used with students outside of the US in mind. For instance, the technical manuals for both the SAT suite of assessments and the Iowa tests identify "Asian" as one demographic category that is used in sample analysis. However, this category is meant to identify US students who are of Asian heritage. These students may be the children or grandchildren of immigrants and therefore may only have experience of living in the US with English as their first language. International students who identify themselves as Asian are less likely to have experiences of living in the US but are more likely to have been born in an Asian city and to have lived in countries where English is not the primary language that is spoken.

The US based testing agencies include categories that allow comparisons on a more local context. For instance, the SAT suite of assessments includes analysis that compares the results of schools to that of both State and District, while MAP testing provides school level norms which allow schools to compare themselves to others within their district. This may reflect an understanding that there will be differences in the curriculum and the population demographics within the state or district which will make a difference to summary statistics and will mean that national statistics are not useful for comparison purposes. However, for international schools, most of the testing agencies do not provide any such breakdown and the comparisons are only with the overall samples used. MAP testing has aimed to provide international schools with more regional comparisons although this is not included in their reports. While they use US population percentiles in the results that they provide through the

score analysis part of their online portal, they do give schools who are outside the US access to comparison information which relates to test taking populations that may be more relevant for them. A secure website provides the means and standards deviations for test takers from different regional organisations such as EARCOS. They also provide this data for individual countries. However, the number of test takers from some countries is relatively small and consequently care must be taken by schools using the data as small samples are more likely to be biased as they are unlikely to include full representation of the population and so could be skewed by individuals or small groups within the sample. For instance, if a selective school is included in the sample and their students form a relatively large part of that sample, then their results will have an unduly large influence on the samples statistics leading to bias. This small sample size can mean that assumptions regarding the normality of the distribution are undermined and so being able to use the Normal distribution to establish percentile information within this group is not possible.

However, with the SAT suite of assessments, for instance, performance is given in terms of percentiles that compare the student to the whole group of test takers and to a nationally representative sample of US test takers, while Iowa testing compares to a National Percentile Rank. As has already been discussed, there are likely to be vast differences between the students included in these groupings and the international school population. Even with the ISA, which is designed with the international school population in mind, the comparison groups that are identified may still not prove useful for internationals schools. The comparison groups of 'like' schools are formed based on the percentage of students for whom English is not considered to be their first language. However, the criteria that are used to

define students according to their language status means that students who are only beginning to learn English are put into the same category as students who have been learning in English for several years as well as students who have one English speaking parent but for whom English is not the main language spoken at home. Thus, the schools that are classified to have the same percentage of students for who English is not considered to be their first language may be very different in terms of the profile of English language learners that they have. Consequently, the impact that this may have on their ability to adequately show their learning when taking part in the test.

The linguistic and cultural features of the test can also lead to construct irrelevant variance in the results that are achieved by students in international schools. While students may understand the constructs being tested, there may be difficulties caused by their ability in the language of testing that leaves them unable to demonstrate that understanding. There are measures allowed by the different testing agencies to support students who are English Language Learners.

In the case of the ISA, it is recommended that schools exclude students who have a low level of English proficiency and for whom taking the test would be distressing. ELL students taking the ISA are allowed to use a bilingual dictionary for the mathematics and science section of the testing only. They are not allowed to use any support materials for the reading and writing sections. For MAP testing, teachers can assign accommodations including text-to-speak which allows the students to have all testing components read to them. They can also use a bilingual dictionary to support them during testing. College Board have recently introduced additional

provision to support English language learners in some of the testing in the SAT suite in that they now provide translated instructions in a number of languages, the use of bilingual word to word dictionaries are permitted and learners are entitled to extra time to complete test sections. One prerequisite for this provision is that students must be enrolled in a school in the US or in one of its territories. Hence, it is acknowledged that ELL students may be able to produce more representative scores using this provision. However, this provision is not available for students in international schools, many of whom are ELL. This may mean that students are not able to show their true level of ability in the testing context. Iowa testing also includes simplified language tests and oral native language support accommodations for students with appropriate documentation in the US, although no information was found to indicate how these accommodations would apply to international schools. While accommodations are in place to support ELL students for all the testing programs listed, research has found that such accommodations are not enough to close the performance gap between ELL students and first language speakers (Schwabe et al., 2016) and consequently, these provisions may not be enough to exclude instances of construct-irrelevant variance from the results of ELL students. However, language is not the only source of construct-irrelevant variance. Cultural contexts can also affect the validity of inferences made.

The US tests include screening in their test development which is designed to make sure that the language used in testing is fair for all students. For instance, in the documentation regarding the SAT suite of assessments, it is acknowledged that there are regional differences in the vocabulary used to describe the same situation, for instance, they comment that ‘pop’ and ‘soda’ are regional variations that may be

used to describe a soft drink. They comment that generic terms should be used in preference to ones that are considered region-specific. However, regional language concerns could lead to issues for international school students on a number of levels which would not be considered when developing testing in the national setting. While these schools may all use the English language as their teaching medium, there are differences in the English language used by different nationalities and students' language development may be influenced by the country from which their curriculum materials are drawn. For instance, if students had more exposure to American culture, they would be used to hearing about the trunk of a car, the windshield, and the trashcan. However, if they are taking an Australian test, these would be referred to as the boot, the windscreen and the rubbish bin respectively. It is not just the cultural use of language that can lead to confusion, as subject-specific vocabulary can also be different in the national versions of the language. For instance, in the US, they would talk about exponents, trapezoids and variables. In other versions of English, this could be indices, trapezia and pronumerals. While it could not be expected that tests that are designed for a national population should include all variations or language for international use, it is important that educators seeking to make inferences about student attainment are aware of contexts that could affect the student's ability to show their learning. Question settings can also lead to construct-irrelevant variance.

Examples of ways in which cultural differences can impact test scores include questions that are set in the context of a sport such as cricket or baseball where the rules may not be understood by students of all nationalities. Currencies denominations which are used in questions may be familiar to some students but not

to others. For instance, testing involving questions on American or British currency may prove confusing for students who are not used to a currency that is based on more than one unit or for students who are unfamiliar with the values of specific coinage such as dimes and quarters. The ISA attempts to remedy any problems with understanding about currency be inventing their own denomination, the zed. This currency has 100 cents represent 1 zed which matches major currencies such as the Euro and the dollar. However, students based in Japan may have difficulty as the yen is the only denomination used and is not broken down into a smaller denomination and therefore the use of a decimal point and rounding to 2 decimal places is not a concept that Japanese based students are familiar with. Even when conversion between different units of currency is included in the curriculum objectives, there will be a difference in comprehension and application between students for whom this is a taught concept and those for whom it is a lived, everyday understanding. Again, US test include reviews to ensure that their test items are accessible to the population of interest. However, such reviews are generally carried out by groups who are unlikely to have an understanding of culture outside of the US. For instance, the SAT suite uses US public school teachers to check that answers to questions are not affected by factors that are unrelated to the constructs that are being measured. However, given that they are based in the US, they will only be considering factors that related to US culture. It is therefore possible that this review will fail to identify characteristics of the test that will have a negative impact on the results of students who do not have the experiences of US culture. Even things like punctuation can be culturally based. In the SAT suite of assessments, one of the skills tested is the use of punctuation. However, punctuation is not used in that same ways by all English-speaking countries and consequently

incorrect answers could be the result of cultural differences rather than from a lack of learning.

Culture may also influence the way that students answer in testing. Many US standardised tests consist of selected response items in which the question includes a number of responses including the one correct answer and the student is aiming to identify the correct response (Mertler, 2007). Of course, there is no limitation placed on how the student can answer. If they do not know the response, they may get the correct answer by guessing. Nowadays, there is no penalty for getting an answer wrong. Hence, students do not lose out if they guess and get the answer wrong. The SAT suite, Iowa and MAP testing all use multiple choice testing with no penalty is given for guessing. Students in the testing population may then see that guessing when they did not know an answer or when they could not exclude all of the incorrect alternatives as an appropriate testing strategy to employ. This would have an impact on the scores that were being used in the standardising sample. It is also an expectation in computer adaptive testing that students will get questions wrong and the documentation which accompanies MAP testing quotes that it is expected that students will get approximately 50% of the questions wrong. However in some cultures, getting an answer wrong is seen to result in losing face which is deemed unacceptable while other cultures may deem guessing as a form of cheating because it is not being truthful about what you know. The impact of this could be that students refuse to guess in multiple choice situations or that they take a much longer time to complete computer adaptive tests.

While lack of cultural understanding can have a minor impact the number of questions a student gets wrong, this impact is made more major when the impact that this can have on the score is considered. For instance, when the SAT suite of assessments is analysing a student's college readiness, getting two or three more questions correct can be the difference between being deemed as on track or being shown as being over a year behind in terms of the attainment to be considered college ready.

Scaled scores are used to represent student achievement in all of the tests identified in the sample. Teachers using reports from different providers need to be aware that these scaled values will have a different meaning within each testing format and there is unlikely to be any relationship between the meaning of the scaled scores given by the different testing providers. International School teachers may use different testing programs for different grade levels within the school and if they are trying to measure student progress, they will need to be aware that scaled scores from different tests will have different meanings. For instance, both the PSAT and the ISA use scaled scores in the range of 0 to 800, but even though the range is the same, there is no evidence to suggest that there is a relationship between the scores that students achieve in one testing system and what they would achieve in the other. When different testing programs are used for different grades, International School teachers may not be able to relate the measures given and they will need to find alternative ways to identify improvement in student has made in the periods between tests. With the wide variety of scales that they may be expected to become familiar with, it is possible that there will be confusion and frustration amongst teachers as they do not have the necessary understanding of the meaning that can be attached to

the scaled scores of different testing programs. As a result, teachers may choose not to attempt to understand all scales but to relate their understanding one scale that they are familiar and comfortable with.

To facilitate the process of identifying areas of relative strength and weakness in their curriculum and look for areas that needs improvement, tests are broken down into different subsections. Subtest scores may be given to show attainment in different skills within a content area. However, when the test is broken down, it must be remembered that there are only a small number of items testing each area so interpretations based on the score must be treated with caution. As values are at best approximations, this can lead to questions about the reliability of interpretations made and the implications for interventions made as a result of them. Subtest scores are frequently not equated from year to year and so using them to estimate progress from one year to another is questionable. With the potential for differences in language and culture impacting the scores of international school student, it is important that teachers in this setting recognise that there is even more reason to question the reliability of interpretations. As there is only a very small number of questions relating to each topic area, the impact of getting one question wrong on the interpretation of the sub-domain are likely to be large, and therefore errors resulting from cultural or linguistic factors could have major effects on the consequential inference.

Percentiles will be used across the different testing programs. Even within the same test, small changes in scores can result in large changes in percentiles, particularly for results at the extreme ends of the distribution. Percentiles used by the different

testing programs will frequently apply to different populations, so caution must be exercised if percentile are considered when looking at individual student results.

This section highlighted the potential differences between the composition of the standardisation sample and the international school population. Further, it considered that the measures taken to overcome cultural bias may not be sufficient to identify where that bias may affect the international school population. As the international school population has a significant number of ELL students, accommodations that were allowed by the testing agencies were also discussed. It was noted that one of the tests allowed accommodations for its intended population but that these were not made available for international school students. The potential impact of small differences on the summary statistics was also highlighted. The factors mentioned above will all affect the inferences that can be made from test scores. The next section will consider some more specific concerns related to the interpretation of career and college readiness from test results.

4.7.3 Interpretation of results

In this section, concerns regarding the interpretation of results will be considered. Following testing, schools will receive feedback in the form of score reports which include analysis of student achievement. Score reports tend to use a lot of statistical jargon in presenting their results. This jargon is not common language to most teachers but is rather the expert language of the educational measurement and testing community (Gotch & Roduta Roberts, 2018; Roduta Roberts et al., 2018). This vocabulary used can mislead teachers that there is certainty in what is being expressed when in fact no such certainty exists (Gotch & Roduta Roberts, 2018;

Roduta Roberts et al., 2018). Given that there are many areas where construct-irrelevant variance could be included in testing when it is used in international schools, it is important that teachers do not see interpretations as absolute but question what the results really mean within their school population. International school teachers also need to be aware of potential threats to the validity of inferences that they may make from elements contained within the score reports that they receive.

One area where teachers need to be mindful of this indication of certainty is in the interpretation of a student's path towards college and career readiness. As has been noted, the aim of the SAT suite of testing is to determine the progress a student is making towards readiness to succeed in college level and career readiness programs. Measures that are designed to provide information on a student's state of college and career readiness are also supplied by both Iowa and MAP testing. According to Iowa, their benchmarks are designed to indicate that a student is ready to enrol in first year credit-bearing post-secondary course while College Board (2017b) state that the benchmark used for judging a student to be college and career ready indicates that if they take a credit-bearing course in a related subject during their first semester at university, they have a 75% likelihood of attaining a grade of C or above. For College Board's (2017) EBRW benchmark, the related subjects are considered to be writing, literature, history or social studies, while the mathematics benchmark relates to courses in algebra, precalculus, calculus or statistics. According to ISC Research (2018), international school students generally out-perform their national counterparts in a range of examinations recognised by a majority of the world's universities while they are well prepared in terms of behavioural and attitudinal

attributes as many international schools foster approaches that develop skills independent learning, debate and collaboration. It may be that the boundaries given do not reflect the level of college and career readiness of the international school population.

In the case of international school students, the level college readiness will be dependent on the country in which they plan to pursue their university education, and while the US has recently been identified as the most popular destination for international school students, there are significant numbers who go on to attend universities in the UK, Canada and Australia while English programs available at European universities are also popular choices. Because of the different ways of applying to the universities of different countries and the differences in the requirements that those universities place on students when considering whether to offer places, it is impossible for a standardised test based on the processes of one country to determine the college readiness for all students in the international system.

4.8 Concluding comments

This chapter presented the results from the document analysis. There are warnings about appropriate use of testing included in each of the US testing manuals which tells the test user that the inferences for testing are only valid for the uses which are described in the manual. When these tests are used in international schools, they are being used outside of the situation for which it was designed in that the population characteristics in the international schools are significantly different to the population that the test is designed for. Even where testing is designed for the

international school population, there are concerns about how valid inferences can be made. However, the document analysis described in this chapter has revealed that there is the potential for significant discrepancies between the features described in testing used in national systems and the situations of international schools. For instance, the population characteristics of the standardisation sample are very different and the content domain that is being tested may not match the international school curriculum. So can international schools be sure that they are making valid inferences when using the information from testing data? What do they need to be aware of to ensure that they are making valid inferences from the results of standardised testing? The responsibility for the validation of any uses outside of those identified in the testing manual with the test user. Finally, what support do teachers and senior teachers need to make sure that they are able to interpret results appropriately? These questions will be discussed in chapter 6, when the above research will be related back the published literature on validity.

Chapter 5

Results – part 2

5.1 Introduction

This chapter will present and discuss the findings relating to the second research question. "How is the information from standardised testing currently being used as demonstrated by a sample of teachers drawn from four international schools based in the Kanto Plains region of Japan?". As explained in chapter 3 above, semi-structured interviews were conducted with a small group of international school teachers. A thematic analysis was carried out on the interview transcripts. Four general themes emerged; school factors that affect data use; factors relating to teachers that affect data use; issues specific to teachers in international schools; and changes that would facilitate data use. The analysis went on to identify categories and subcategories within each of the themes. The information relating to each of the categories and subcategories will be described at the beginning of each theme presentation below.

5.2 Schools factors that affect data use

This section looks at factors that affected data use within the schools in the sample. As stated, the first theme to emerge related to what was happening to increase data use within the schools. The first category within this theme considered the initiatives introduced by three of the schools to increase the amount of data use. These initiatives will be described in the first subsection. Each initiative resulted in different expectations about how data was to be used within the schools. Therefore, the second subsection will give details about the expectations for teachers to use data within each school.

5.2.1 Initiatives related to data use

As stated above, three of the schools had implemented initiatives to increase data use and comments relating to those initiatives will be given in this section. The initiative in school A, which centred around an action plan that was designed to encourage teachers to use data within their departments, will be described first. Details of school B's initiative will be given next. In this initiative, standardised testing data had been analysed by an administrator and a teacher in the school. The analysis was used to introduce a new scheme of work and teachers were assigned to Professional Learning Communities (PLCs). Finally, the initiative in school D will be described. This initiative centred around classroom data and also saw the introduction of a data team.

The administrator from school A acknowledged that international schools, like their national counterparts, have a wealth of data. However, as he states in the following comment, that data is frequently under-utilised.

Administrator 1: *The school does not suffer from a lack of data but much of what we have collected is not being used. There is not much specific analysis being done with the PSAT. The question is more about "what do you want from it?"*

The administrator explained that use of data was identified in an action plan at his school.

Administrator 1: *There are two school action plans which we hope will lead into each other; using data to make better decisions and improving English language proficiency. We hope this will lead to some action research.*

This initiative was introduced by showing a video about data use to the department chairs. This led to a discussion based on their school data regarding the kinds of

information they could get and what that information would mean. Teachers were given some rubrics about the data that included the percentages of students attaining specific levels in the PSAT. The teachers were asked to suggest interventions based on the data they had seen that would increase student progress so that students were reaching the levels that were appropriate for their age. The discussion considered areas where improvements could be made within departments. In this school, the initiative was aimed at getting teachers to increase their use of data. However, in a second school, the data was analysed by an administrator and a teacher and the results were used to implement new initiatives to improve learning.

In school C, teachers were not expected to review the data from the standardised testing. As is shown in the following statement, it was perceived that teachers did not have the skills necessary to analysis data.

Teacher 4: *So your average teacher, I mean different strengths in different areas. Perhaps analysing different tables isn't going to be their strength and searching for patterns in data isn't going to be their strength*

Instead, a comment from an administrator acknowledged that analysis of testing data had been done by a teacher who was interested in this area.

Administrator 2: *I know that (teacher 4) has done some really good work around data collection here over the last few years. ... I mean I think we're on that journey, we you know, we're going down that road and he probably deserves the lion's share of the credit for getting the school moving in that direction*.

When the administrator started at the school, he became interested in using the analysis that had been done as a catalyst for improvement. The interviewees used the data to identify curriculum areas where there were weaknesses with a view to introduce initiatives to strengthen their students' skills. Analysis suggested that students' reading skills were relatively weak compared to other skills that were being

tested when being compared to like groups. This is shown in the following comments:

Administrator 2: *Each time they were behind other like schools. So you know we were using the data to make some decisions about our need to really focus on strengthening the reading this year.*

Administrator 2: *as a result our school community is focused on reading and we know from the ISA data that (teacher 2) compiled over the last couple of years that reading is an area that requires strengthening at the school.*

Administrator 2: *And so you can see for (school name) for reading you know the kids were, our kids, this class in, so this was grade 3, grade 5, grade 7. ... and each time they were below other like schools. So you know we were using that data that to make some decisions about our need to really focus on strengthening reading this year.*

Having identified an area of weakness, the administrator commented that a presentation explaining the analysis was made to teachers.

Administrator 2: *So the head of school and myself did a presentation to all the teachers and it wasn't an uplifting presentation because in many ways, it was delivering some bad news that our reading results aren't as strong as we would like them to be and as a result we really need to focus on reading this year.*

The administrator said that the presentation was also used to introduce interventions including a new reading program. Teachers were trained to use the program. As the school wanted to make sure that teachers felt supported as they tried to raise attainment on reading, the teachers were invited to ask for other resources such as class libraries that would enhance learning within the program.

As the following comment shows, the school had also instituted PLCs which were designed to get teachers to talk about the work they were doing in their classrooms.

Administrator 2: *So they're operating this year, we've started the concept of Professional Learning Communities and all teachers are setting SMART goals around a learning area.*

They would share good practice and discuss what could be done to remedy situations where students were failing to make adequate progress. The administrator commented:

Administrator 2: *We've adopted the professional learning community model and the expectation that they operate in a PLC –professional learning community and that they are regularly meeting, they are looking at student learning data, they're looking at their assessments, they're looking at interventions and as a result they are looking at student learning data.*

They looked at ways of working within their teaching teams to discover what the data was telling them and then to translate this into actions that would improve the learning within their classrooms. He stated that:

Administrator 2: *They're often teaching reading collaboratively; they're working with each other's kids and they're having conversations about each other's kids and their learning.*

Sometimes this might mean that each teacher in the team would teach a skill to one class and then the classes would be swapped so that the teacher was teaching the same skill to a different class. The following comment demonstrated this.

Administrator 2: *So we've got to think about each, what are the strategies that kids need. They're working together with two classes and breaking them up based on needs and they're working with the learning support teacher or the EAL teacher to try and meet kids' needs with this particular goal. Then the really rich thing about is that is job invented professional learning. So they are learning from each other and, you know, learning how they all teach reading and then trying to employ the strategies they are learning.*

They were also able to use the learning support teachers in a more focused way. This initiative used the results from standardised testing to introduce curriculum changes designed to improve student attainment. However, in the third school, analysis was focused around classroom data.

The teachers in school D commented that an initiative that had been implemented there was designed to increase the analysis of data, but this was centred more around analysing data within their classrooms. An example of this is shown in the following comment.

Teacher 5: *And I know that the school is going towards, or is trying to move towards more data analysis within the classroom. It's got to do Visible Learning*.

Teachers had undergone training including materials designed to encourage them to collect and analyse data from classroom tests and assignments to confirm that students were making progress within their units of work. They had been taught to use a statistical calculation involving the means and standard deviations from pre- and post -tests, usually using a spreadsheet to find the effect size, with boundary values being given to identify when adequate progress was being made. This calculation was also being used on the results of standardised testing as a measure of progress, although it did not lead to identification skills where weaknesses existed and improvements could be made. This school had also started a data team to review the results of standardised testing as it came into the school and two of the teachers in the sample commented that they were members of this team.

Teacher 5: *Well the only way that I know that there is a movement towards looking at more data in the classroom and the results is that because I am on the data committee.*

The different initiatives led to variations in the expectations for data use between the schools and these will be considered in the next subsection.

5.2.2 What expectations do schools have for data use

This subsection will describe the different expectations that the schools placed on teachers to use data. Details of how teachers were required to use the data under

each of the initiatives will be given. School C did not have any data use initiatives. The teacher from school C felt that there was a different purpose for standardised testing data within their school and this will be explained.

In school A, the administrator stated that aim of the initiative was to get teachers to start looking at data. This would encourage them to talk about what they could see. The administrator stated that all data was shared. It was expected that all teachers should look at and think about what the data showed. However, as is shown by the following comment, there was no stipulations for how the data should be used.

Administrator 1: *All get the data but it is not certain how it is used. There is not a school policy which says it has to be used in a specific way.*

The plan was that by getting teachers to focus their attention on data and discussing the results with other teachers would give them the opportunity to become more confident. He commented that this, in turn, would result in them making better conclusions about what the data showed.

Administrator 1: *Getting people to look at and think about the data takes time. Everybody has a different capacity and different thoughts. Some people found it useful and the more they found it useful the more they used it.*

Following the implementation of the action plan, the administrator stated that teachers were talking more about the data and thinking more about what it was showing them. He stated that:

Administrator 1: *Talking to teachers makes things better – it is possible to see improvement. They develop better habits and start to make sure that they look at the data*.

The administrator highlighted differences that were seen in how departments responded to the initiative. He stated that some departments had their own individual discussions. Others departments had joined together to discuss what they

could understand from the data. There were one-to-one meetings with a focus on a particular data set such as the feedback received following AP examinations. The administrator in school A acknowledged that it would take time for teachers to engage with and use data, and that the capacity to use and understand data would be developed at different rates depending on the comfort level of the individual teacher. He went on to state:

Administrator 1: *As teachers were using the data more, they were finding it more useful and the more useful they found the data the more they continued to use it.*

In having conversations with teachers, the administrator had gained the impression that teachers were gaining in confidence and getting into better habits about using the data. As a result, they were seeing that data could be used successfully to inform their teaching.

In school C, teachers were required to set learning goals for individual students and for their class at the beginning of the initiative, as shown by the following statement.

Administrator 2: *So although they've set a grade level goal of around 85% of kids being K, then what they did was they set an individual growth goal for each child. So they are really individualising learning; thinking about where is the kid in September and where could they potentially be in May/June.*

Teachers were then required to collaborate with colleagues in their team. They were expected to work out the what changes to make to improve learning. This included identification of areas of weakness and changes that needed to be made to teaching strategies. The administrator commented:

Administrator 2: *What are the short-term strategies that we are working on, whether it's making inferences, predicting; so they are really thinking about and having rich conversations about how you can improve student learning around reading.*

Teachers were asked to provide regular feedback data from their classrooms showing how their students were responding to the new reading program. This included the results from assessments that were included within the program. Teachers were required to send an analysis of their students' results back to the administrator at least three times a year. The administrator stated:

Administrator 2: *What we've done is we've asked teachers to conduct the Fountas and Pinnell assessment at least three times a year…. And what teachers are doing is that they are conducting that assessment and then they are sending that data to me and so I'm compiling it.*

They were able to choose how they presented the data. It was noted that many different formats had been used, including varieties of colour coding, written formats, tables and numerical values. However, it was envisaged that a common format would be required in the future. Teachers were asked to review the data they were compiling and use it to inform their teaching by identifying students who were not considered to be making adequate progress. At the end of the year, teachers had to meet with the principal and vice principal to discuss how their students had progressed. As shown by the following comment, the objective was to get teachers to think about improving teaching.

Administrator 2: *For me, I don't really care what their goal was. It's that they are thinking about it. You know, they spent a bit of time thinking about it and tweaking it and wordsmithing it. At the end of the day, for me, grade 1 is focused on improving reading. That's all it is. How do we improve reading? I'm not interested in whether they make 85% or not.*

Teachers were not treated differently based on whether they had reached their target or not. He commented that:

Administrator 2: *So all of that is expected but we are not necessarily holding them accountable to "this child hasn't grown and you're responsible for that", you know. That's not, that wouldn't be fair.*

Teachers were also required to give a short presentation to their colleagues sharing how their students had performed relative to the targets they had set. This was seen as a chance for teachers to celebrate the progress that their students had made.

The results from standardised testing in the school were used by the data managers to triangulate the data received from the teachers and provide evidence to determine the success of the program that had been implemented. As shown by the following comment, the standardised testing results were used to triangulate the results that were obtained from the internal testing to identify anomalies and look for areas where further improvements could be made.

Administrator 2: *But what it does allow you to do is look for inconsistencies, have conversations with teachers to compare the ISA with what they are seeing within the classroom and then really think about how we can use all of that to help kids learn*.

Data from testing only came yearly, and there was a time lag between the students taking the test and the school receiving the results. Therefore it was not considered to the best data to be used to help in day-to-day teaching.

Teachers in school D reported uncertainty about what they were meant to do with the data for one test that they used. They reported that there was a push to make data more available within the school but that they were not really sure what they were expected to do with it. As shown by the following comment, they had looked at and discussed the data from the tests within their department.

Teacher 7: *I think we are meant to discuss it in our department but we don't generally discuss the ISA in our department. I think we tried for a while to take a look but I think that we found that it was kind of overwhelming.*

However, they stated that there were impediments that meant that they were not able to get useful information from the data for one of that standardised tests. Testing data was not used by all schools as a way to improve student learning. For some, the results were more for current or prospective parents.

The teacher from school B expressed the opinion that the purpose of standardised testing within the school was to provide comparisons to other similar schools. This was then used to market the school to prospective parents. The teacher stated that such data was not seen to be relevant to the classroom or to individual students' progress.

Teacher 3: *The school gets data from the ISA and PSAT but if we get feedback, I am not aware of it. It is used more to gauge the school relative to other international schools rather than to give feedback about the teaching and learning within the school.*

The feedback from the tests was not applied to learning within the school. Parents will use the results from standardised testing to confirm that their child is getting educational opportunities that are at least as good as they would receive in their home country, as is shown by the following comment.

Administrator 2: *We would show where our students are and our grade 7 students would often be higher that the OECD average for grade 9 and that was always very reassuring for our parent community. And then you can correlate that with PISA, and they show the breakdown by country, you know, the Netherlands or India and Malaysia and Australia, and you can show where your school is in relation to all other countries and that was always very validating for parents who were concerned about living abroad in an international community. ... They could see that their kids were doing better than the average child in that country.*

As the results from standardised testing were not shared with teachers in school B but did appear on the school website, this teacher interpreted that the results of such testing were not perceived to be important to the teaching and learning process but

were needed only for external monitoring and promotion purposes. Analysis for their teaching was expected to come more from the feedback they received from the other programs that they followed, such as the IBDP.

5.2.3 Summary of section 5.2

This section described the initiatives that were being used to increase data use within the schools in the sample. It can be seen that the schools adopted different methods to increase data use. In one school, the analysis was conducted by a teacher and an administrator. Here, the analysis of standardised testing led to the introduction of a curriculum initiative. The teachers were expected to implement that initiative. They then provided feedback about the progress of their students to the administrator. Teacher development was facilitated through the introduction of PLCs. The second school used Action Plans with the intention that all teachers should improve their analysis skills. At the time the interviews were conducted, the teachers were still familiarising themselves with the data that was available and working out how the data could be used. The third school introduced a Data Analysis Team but was more focused on teachers using analysis of classroom assessment. They implemented a program that was designed to help teachers to implement the classroom assessment. The schools all had different expectations of their teachers in terms of data analysis. The next section will describe how the teachers interpreted these requirements.

5.3 Factors relating to teachers' use of data

The next section will present information relating to the second theme which considered the factors that affected teacher's use of data. The section will start by describing teachers' understandings of how their schools expected them to use data.

It will also relate teachers' comments on their ability to use data. Descriptions of any training and professional development that teachers had undertaken to enable them to use data will then be given. Barriers that prevent teachers will then be discussed. This will start with by identifying concerns that teachers had about using the data in their practice. Factors such as confidence, access to data and time will also be discussed. These are the factors that affect teachers across national situations, according to the research literature on data use.

5.3.1 Teachers' perceptions of how they were expected to use data

The next subsection will present the information given when teachers were asked to describe the skills they had developed to interpret testing data. As identified in the literature review, it is necessary for teachers to undertake appropriate professional development to enable them to interpret data effectively. Details of any training that teachers had received will also be given in this subsection.

In school A, a head of department explained that while he was expected to have looked at the data and to be familiar with the information that was contained in it, there we no specific requirements about the timing. The teacher stated:

Teacher 1: *I am expected to have looked at it and I am expected to have thought about what it means for my instruction. I don't think I am expected to have spent like once a week, for half an hour or anything like that. There's no time expectation. But at least as a department chair I am expected to be familiar with what the results are. And then as a department chair, I expect the teachers have at least looked at it say once every semester or so - whenever it comes out. When they get a chance to see it. And at the very least, from having seen it they've thought about "Ok, this is something I need to change or this is something I'm encouraged about, that I can, that we can work with so we'll keep doing well."*

The initiative had resulted in more conversations about the data both between the heads of department and within each department. He commented:

Teacher 1: *But we have spent time as departmental chairs here just talking about what they mean, what kind of data we can get out of them. We have spent time as teachers looking at the data we collect and what does it mean, and what implications it has for what we do*.

In the same school, a second departmental chair emphasised the need to talk to members of her department about the data and to discuss how they might alter their teaching based on what the data was telling them. She said:

Teacher 2: *but as an initial thing just offering to the teachers "Hey I got this data. Would you like to meet just to talk with me about it? What you found encouraging, discouraging, you know, ways to get some traction on some things. And so I did that with two of the teachers.*

She had made sure to have conversations with each of their AP teachers to find out about issues that were identified in the feedback they had received and their response to it. In those conversations, it was found that the teachers had already reviewed their own data and had made changes to their teaching in response to the feedback they had received. However, interpreting the data requires knowledge of how to use the data. This led to comments from teachers on whether they could understand the data they were presented with. Their comments on their ability to interpret data will be given next.

There were 8 teachers in the sample who stated that they had the ability to do some analysis of the data that they received. Only three indicated that they could fully interpret the statistical data from all the testing their school did. A teacher who was just beginning to use the data that was being shared was uncertain about how to interpret the data correctly. She stated that:

Teacher 2: *We administer it and we get that information. I know I should use it and I don't ….it was interesting to look at the data for my class now and think about what that means for… I am not sure how to use that*.

Because the teacher was uncertain about how to get appropriate information from the data, she had previously chosen not to use it. Even though she was now starting to look at it, she stated that she still did not know how to use it.

As is illustrated by the following quote, where teachers did state that they were able to interpret the data they related their understanding to experiences of using statistics in other situations.

Administrator 1: *My father was a mathematics teacher so I had a general feel for the use of statistics.*

As shown by the following quotes, teachers attributed their understanding of statistics to the subjects they were teaching.

Teacher 1: *I think, as I mentioned with the studying of the statistics, as I mentioned, that has helped me to interpret the data better.*

Teacher 3: *Data analysis is part of my science training. I am using data all the time.*

They adapted that subject knowledge to help them. However, not all of the teachers who used statistics in other areas could use it to aid their understanding of statistical concepts used in the testing data they received. They expressed concern both in the interpretations they were making and their ability to support other colleagues who were less able to understand the data. For instance, the teacher who had been studying statistics commented that:

Teacher 1: *I probably need to have my vision broadened too about what the possibilities for data are too. … There are training issues. I wonder sometimes if I don't know enough about how to interpret the data and then how do I then train other people to be able to interpret the data in a way that is useful.*

So even where teachers have an understanding of statistical concepts, it may be that they would benefit from more specific training on how to relate this to interpreting testing data. Therefore, teachers were asked to describe any training that they had received in interpreting the data they received, and this information will be given next.

Most teachers reported that they lacked training and experience in the skills related to data analysis. Six of the nine interviewees stated that they had not received any training in analysing standardised testing data during their initial teacher training. Furthermore, they had not been involved in any in-service training courses specifically designed to develop their skills in such analysis even though their schools were expecting them to make more use of the data from standardised testing. The following quotes provide evidence that teachers had not received training in how to analyse data.

Administrator 1: *As a teacher who graduated in 1985, this was not part of what was in the teacher training experience.*

Teacher 1: *Ok. I don't think I have ever received any formal training into how to read them.*

Teacher 2: *What training? I have never had a class in statistics and I haven't really had any training in the subject*.

Teacher 5: *I have had absolutely no training whatsoever. Neither at teacher training, at previous schools, or even really here.*

Teacher 7: *I feel that when I am reading data, I am relying on math skills that I learned in high school*.

Only one teacher in the sample (Teacher 4) reported receiving training related to analysis of testing data while working at their current International School. This teacher had taken an online course in using assessment to inform teaching followed by an overseas course aimed at improving understanding of the data in the ISA. This

seemed to stem more from the teacher's personal interest than from any initiatives that the school was developing at the time. Following the training, the teacher worked with the data available in the school and shared his interpretations with other colleagues.

While there was little evidence that specific training had been given, some teachers did mention that they relied on experiences from previous schools as this had resulted in them developing some skills to interpret the data in standardised testing reports. As shown in the following quote, one teacher reported having experience of target setting using standardised testing while working at a local school in England.

Teacher 6: *And then just on-the-job training in a school in the UK and that was with the head teacher and with an LEA consultant – a Local Education Authority consultant – and the data manager of the school as well. Target setting for what they should get by the end of Key Stage 3 and what they should get by the end of Key Stage 4 and having those as minimum target grades for students so that everything was assessed on trying to push them towards those minimum targets at Key Stage 3 and Key Stage 4.*

The teacher was employed as head of the English department and was working in collaboration with a head teacher and a Local Education Authority adviser. They were involved in setting target levels for students to achieve in future assessments based on assessments that had been recently completed. Results from testing were not used to analyse the particular strengths and weaknesses of a student or class. The impact on the curriculum was therefore based on the predetermined skills defined by the levels that the students were working towards rather than any attempt to identify or remediate difficulties that the students were experiencing.

As mentioned above, school A, had introduced their initiative with a video presentation and follow-up discussion.

Administrator 1: *The department chairs were shown the video and felt they were being told nothing surprising.*

Teacher 1: *We have spent time as department chairs here just talking about what they mean; what kind of data we can get out of them.*

Following this, the departmental chairs then worked with their departments to continue analysing the data.

Both administrators expressed a view that there was a great variability amongst their teachers in terms of their ability to interpret testing data. Administrator 2 stated, "Everybody has a different capacity, different thoughts", while the other acknowledged:

Administrator 1: *Some people have trouble because they haven't used data before. Some people do have a proclivity towards data and they can move forward.*

The teachers that were interviewed also stated that there were training issues that needed to be addressed within their schools, either for themselves or for members of their department. When asked what they would do if they wanted to gain more understanding of the data, teachers suggested that they may ask a colleague for help.

Teacher 2: *Yeah, there are people I could go to ask the questions. I could. Our PSAT, the person in charge of the PSAT, she's more than happy to talk to anybody about any kind of data.*

Others commented that they thought that there might be support materials produced by the organisation that was responsible for writing the test but they were not easy to access. For instance, teachers commented:

Teacher 1: *And I am thinking that, probably somewhere they will actually tell you "ok, this is what a score will tell you and this is what a score won't tell you", but I haven't ever seen anything like that handy or in an obvious spot.*

Teacher 2: *And also, there might be those things out there somewhere. I am sure that they offer, I don't know, maybe they offer something like that but it's also low on the totem pole of felt needs.*

So if there were materials that could support teachers in their understanding of the data, they were not easy to find and the teachers stated that they did not have the time to search for them. Teachers did not indicate any knowledge of where the materials might be, for instance if they might have been sent to the school or were available on a website. It was not just the ability to use data that affected teachers' choices about using data. Teachers in the sample highlighted other potential barriers that prevented them using data in their practice and these will be described in the next section.

5.3.2 Teachers' concerns about using data

A number of concerns were expressed by teachers about using data and these will be discussed in this subsection. These concerns included fears that the data would impact their curriculum and lead to teaching to the test. Teachers' own belief about data use could also result in the data being used superficially. Use could also be restricted due to organisational factors both within the school and regarding the testing.

International schools have the freedom to choose their curriculum. A number of teachers stated that they did not want their curriculum to be determined by a test. Teachers wanted the curriculum to be determined first and then a test chosen that would give them information about the success of that curriculum. The following comments show that they did not want to choose a test and then build their curriculum around that test.

Administrator 1: *The testing should test the program it should not be used to decide the program*.

Teacher 3: *The general feedback is useful in terms of what the students are expected to do in the exam although the danger is it can result in two years of teaching to the test.*

Teacher 5: *I just don't want to end up seeing like teaching to a test or stuff like that. That would be my biggest fear.*

Their philosophy is that the test should measure what they are teaching and that it should not determine what is being taught.

The teachers interviewed did state that data analysis could be used to improve the educational achievement of their students. However, as shown by the following comments, they had concerns that they would not be able to use it successfully as they lacked the knowledge to apply it.

Teacher 5: *I think the data is very important. What I don't necessarily enjoy is trying to interpret it all.*

Teacher 2: *it was interesting to look at the data for my class now and think about what that means. I am not sure how to use that.*

An administrator expressed the view that teachers may not all be positive about using data. This is shown by the following comment.

Administrator 2: *different teachers are in different places about whether it's a good thing or not, you know. Well we're going through all that at the moment*.

This administrator also stated that teachers' willingness to use data could be affected by concerns about the implications of the results on their job. He commented that some teachers related the use of data within the school to initiatives such as NCLB in the US and were worried about how test scores would be related to their performance evaluation as teachers and ultimately to their job security. As one administrator stated

Administrator 2: *You know whether, there are some people who think well it's 'No Child Left Behind' and what is this all about and so where to begin.*

Data was not tied to job performance in any of the schools and no teacher related instances where data had been used in discussions relating to concerns about their performance. One school specified that the use of data had never been tied to teacher performance.

Administrator 2: *We're not using the data in that way though. We're not using it as part of supervision and evaluation. So the notion is that it's meant to be, the focus is really how do we improve student learning and part of that involves looking at student learning data and knowing who they are and where you want them to be.*

However, another teacher mentioned that he was aware of an international school where analysis of testing data was used to decide pay and to determine financial incentives. Teacher 4 stated "It was tied to their bonuses. It's the Singaporean approach to right." The teacher did comment that this might be more due to the influences of the culture of the particular country that the international school was based in rather than an indication that there was a movement towards payments by results within the international school system.

Teachers commented that they lacked the time to spend interpreting data. As shown by the following comments, they devoted time to other activities that they perceived as more likely to have a positive impact on their students' education or they found they had no time once they had fulfilled their day-to-day responsibilities.

Teacher 2: *we also have enough to do without looking for reports*.

Administrator 2: *I don't know that the school as a community has necessarily focused; has had the capacity or the time or whatever to focus on student learning data like the ISA and to really think about what it means in terms of student learning*.

Teacher 7: *There are members of my department that are much more adept at reading that kind of stuff but there's also the time factor of them then having to go through that and pull it apart, as well.*

They commented that there were more pressing issues to be dealing with and this meant that they were unable to devote time to analyse data.

Teachers commented that there were issues with gaining access to data within their schools. As shown by the following comments, teachers were not given full access to all available data.

Teacher 2: *The rate at which the school is proactive about sharing the data is irregular.*

Teacher 6: *It's not always centrally available.*

When access was given, it was patchy and inconsistent from year to year.

Teacher 2: *Sometimes it is sent out and made available to me and sometimes not until I track it down*.

Even where full access was given consistently, teachers reported that they did not know how or where to gain access.

Teacher 7: *Sometimes I am just told its available rather than where to actually go find that data*.

These inconsistencies can frustrate teachers who are trying to use the data and ultimately mean that they give up on trying to use the data in their practice. Even when data was made available, teachers commented that the time lag between testing and receiving information could also cause problems.

Teacher 2: *Well since it's given in October and I don't get it until you know probably January.*

Administrator 2: *You know what happens with the ISA. Its comprehensive but it takes so long to get the data, you know and so, and by then the kids have moved on.*

Many of the standardised tests used by the schools have at least two-month gap between when the test is taken and when the school receives the feedback.

One teacher commented that the amount of data that is available can be the cause of problems.

Teacher 2: *I know we have this huge pile of data; we must be able to do something with it…. I've gotten out the data every once in a while and tried to have a look at it and gotten so overwhelmed by the amount of it.*

Having so much data resulted in the teacher giving up because they didn't know how to begin processing the data into useful information. While many of these issues reflected the difficulties faced by teachers in national schools, some of the teachers highlighted factors relating to data use in the international school situation.

5.3.3 Summary of section 5.3

Many of the teachers in the sample expressed a lack of confidence in their ability to interpret data. They did not have previous training or experience that would help them to understand how to gain information from the data. Some in-school training was being done in the schools although this tended to be focused on the particular initiatives that the schools were implementing. As the schools were only at the beginning of implementing their initiatives, it is not known whether there were longer-term plans for professional development within the schools. The only teacher who had attended external professional development had done so at their own request. However, it was not just lack of training that prevented teachers from applying information from the data in their schools. The timeliness of receiving the data, teachers' access to the data and their own beliefs about using the data could also influence their decisions about trying to interpret data. The issues highlighted here are those that are identified in the literature as they affect teachers in national

situations. However, the teachers in the sample identified some concerns that related to their school populations and these will be described next.

5.4 Issues specific to teachers in international schools

International schools were the focus of the study and a theme emerged raising issues specifically related to the populations of international schools. The interviews with the international school teachers raised issues with data use that are different to the general experience of teachers in national school systems. Teachers in the sample identified other factors that may be more specific international schools. Therefore, the section will describe factors such as the English language ability of students in international school populations and the effects of the transient population on summary statistics. Finally, the impact of cultural issues was also raised.

The first factor to be described is the impact of students' language on testing. Teachers from three of the schools involved in the sample mentioned the disparity that was seen in the attainment in test scores for reading and writing when compared to those in mathematics. This was highlighted in the following comment.

Administrator 2: *Overall as a school, our reading was always a little bit behind and our mathematics was a little bit stronger. But it took a lot of people by surprise when the saw that. The idea was that there was always feeling that our maths was weak and our reading was strong. But when you see the cold hard facts. A lot of it may have to do with our EAL population.*

Teachers had perceived that students performed well in English while mathematics was the weaker subject in the school. As the following comment shows, because English is not the language that is used in the student's home environment, their use of English will be restricted to the times when they are at school.

Teacher 6: *performance that is less good as opposed to the rest of it, which tends to be a particular skill which is knowing the meaning of unfamiliar words. And that's something that I would expect from students who are not immersed in English all of the time.*

While students may appear to have some mastery when speaking the language in school, they may experience problems with written English in areas that would not be common to native speakers. As the next comment explained, students may not have acquired English vocabulary at the same rate as native speakers because they are operating in a different language outside of the educational environment.

Administrator 2: *It's not language acquisition. Its more analysis and so those are things that, and this is where any EAL student is going to have big; EAL students tend to be a lot more literal in their reading. They don't understand the nuances; they don't understand the jargon or the colloquialisms. So that's where they tend to struggle.*

This reduces their ability to use knowledge of the language structure when working out the meaning of words that are new to them. All of these schools have programs or initiatives that are designed to improve the skills of their second language students but there were concerns that the testing that was being used may not be helpful in measuring the success of these initiatives. In one school where they had implemented an initiative that was seen to have resulted in improvements of student attainment in vocabulary. As the following comment shows, there were no clear indications of improvement shown in the results of the standardised test that they used.

Administrator 1: *Although vocabulary was targeted, it was not tested appropriately. The program that was being used is a good program but improvements that are being made are not showing up and there is not another test that would measure the improvements.*

The areas of vocabulary that were targeted in the teaching initiative were not those that were being tested in the standardised testing and consequently the results of the

test did not show the improved attainment that was seen within the school. Another school had moved away from one test because of the prevalence of American language and the use of cultural settings that non-American students would not understand in the questions. A majority of the students in the school had not lived in the United States and the school's philosophy was that teaching should not favour one form of the English language or one particular culture. This was reflected in the fact that curriculum resources were drawn from many different English-speaking countries. As a consequence of reviewing questions in the test, the school decided that the results were unreliable because it was uncertain whether the mistakes made were because of deficiencies in the English skills that were being tested or if they were from an inability to understand the cultural references and culturally specific language that was being used in the testing. As the following comment shows, this disparity could affect the inferences from the standardised testing that was being used.

Administrator 1: *College Board state that there is a 65% chance that a student who scores 1550 in the SAT will get a B average in college. When the school looks at SAT data they see that many of the students who get this average do so because of their mathematics result. The two English scores are relatively weaker.*

For many students in the international system, English is not their mother tongue and they are learning in a second or third language. This leads to differences in the way they perform in the test. It calls into question the validity of the inferences that can be made from the testing. However, language was not the only factor highlighted by these teachers. Teachers also commented that the transience of the population could lead to issues with testing data and this will be discussed next.

Teachers highlighted the impact that the changes in the school population could have on the usefulness of the information that they received from testing. For instance, some teachers commented that testing data is most useful for tracking the academic growth of their students as they progress through the school.

Teacher 2: *I would be interested in looking at, in tracking them from 10th to 11th grade, what's the increase? Because it would mostly increase from 10th to 11th.*

Teacher 5: *What's useful to get from those presentation reports in particular ISA, is if we've had a student who's here long enough to track their progress or to see whether they have kind of remained the same.*

However, as the following comments show, international school populations tend to be transient and so tracking of individuals is made difficult.

Administrator 2: *Just another note is that before I got here, the school was also doing the ISA's 3, 5 and 7 only. And I did a disaggregation and tried to show how many students we could actually follow then from year to year and by the time we got to grade 7 we only had like 4 or 5 kids that were connected to grade 3.*

Teacher 2: *I feel like there should be a good thing to do like really in a school as small as we are, class score can vary wildly from year to year sort of, based on if you have a sample of 30 kids and five of them are gone one year and five new ones come that could radically change. You know, if the five who leave are native American, English speakers and the five who come are second language speakers. That could make a huge difference in the scores. And I feel like what I really, what I would almost like to be able to do. I wish someone would tell me what to do to track from year to year.*

In many cases, students will attend a given school for two or three years while their parents are on an overseas posting and then they will move to another international school or back to their country of origin. Students vary in terms of their educational background and their English language ability and it is unusual for the students leaving to be replaced with students who are of a similar ability or the same language background. Because international schools have relatively small year groups, small changes in the ratio of native to ELL students can have a big impact on summary

statistics. As a result, the composition of cohorts can change markedly from year to year and consequently, the tracking of longitudinal progress of a given grade can be quite challenging.

Curriculum evaluation is also made more difficult as it is impossible to separate whether changes in attainment are due to initiatives that have been introduced or the changes in the compositions of the grade.

Teacher 5: *You can see; usually, some years critical reading skills might be a little bit lower than other years but that's again you have to put that into context. Whether we have possibly more ELL students that year or and overall.*

Individual cohorts will also have very different composition in terms of language background meaning that initiatives that may be beneficial for one grade may not be necessary or helpful for another grade. Furthermore, teachers who take jobs in international schools tend to stay for shorter periods of time. Different issues may come to the fore depending on who is teaching a class, again making it more difficult to identify skills to focus on for improvement. However, it was not just language that caused concerns for these international school teachers. Some issues arose in relation to culture and these will be discussed next.

One teacher commented that sharing information from data was made difficult because of issues relating to culture. He had tried to share the results of analysis from standardised testing data. However, his findings were initially met with resistance.

Teacher 4: *trying to find a student, a learner-centred problem and got a lot of pushback from the teachers. It was the first time that anything like that had ever really been implemented here and it was kind of like, I'd only been here a year also, so "Hey, new guy" a little bit*

In Japan the senpai/kohai system demands that younger, newer members of staff show respect to those who are older or longer serving. He commented that the older members are seen as mentors, while the younger are expected to learn from them.

Teacher 4: *But you can't be the youngest one on the staff in Japan, in a Japanese dominated team and really lead it.*

This led to difficulty when a younger member of staff wanted to share evidence of student weaknesses and introduce curriculum change into a department which was predominantly made up of older, longer-serving teachers who are Japanese. Within international schools it is possible to have departments where the dominant nationality has a cultural practice that require sensitivity when introducing new initiatives.

Other teachers highlighted the impact of their workload on their ability to spend time working with data. Because international schools tend to be small, the workload attached to specific roles may be greater than would be the case in an equivalent school in national educational systems. One teacher commented that

Teacher 2: *So we teach two English AP classes, two Social Studies AP classes and I just found out that the Psychology AP class has been thrown into my pot even though we don't have any other connection with that teacher.*

Teacher 2: *The teacher is actually the school counsellor and that's his main department…But he is difficult to pin down on a meeting time for that reason.*

Teachers may have to take on multiple roles and therefore may be busier than their counterparts. This increases the effects of time pressures and making it more difficult to do things such as find time to schedule meetings to discuss data.

5.4.1 Summary of section 5.4

This section highlighted some of the concerns expressed by the teachers in the sample that related to the populations of international schools. Teachers commented that standardised testing was not sensitive enough to pick up changes in the attainment of their ELL students. Differences in the attainment of ELL students could also undermine inferences from testing such as those predicting future college success. Further, the transience of the student population made it difficult to know if changes were as a result of a change in the student cohort or if it had resulted from improvements in attainment. Finally, it was highlighted that teachers and administrators need to be sensitive to how changes are introduced in situations that involve teachers from multiple cultural backgrounds. Considering some of the difficulties that were highlighted, teachers were asked to suggest improvements that might allow them to make better use of the information from testing and this will be discussed in the next section.

5.5 Suggested changes to improve data use

The final theme that will be discussed covers teachers' suggestions for improvements that might make the information from testing easier to use. The first subsection will cover alternative forms of presentation. Teachers stated that if data was presented in different ways, they may be more able to make use of the information. The second section will relate comments on the categories and groupings that are used for analysis by the testing agencies. Finally, comments regarding the use of technology will be discussed. The thesis will continue by looking at comments on the different forms of presentation that could make it easier for teachers to interpret the data from standardised testing.

Many teachers struggled to suggest changes in the presentation of testing feedback that would enable them to analyse the data better. They commented that did not have suggestions for improvements because they did not have knowledge of other presentation formats that could be used. They lacked awareness of alternative types of statistical presentation and because they had not had exposure to feedback from other testing agencies, they could not suggest formats that would be more useful to them. Where suggestions were made, teachers expressed a preference for the feedback to include written summaries of what they should be able to deduce from the data. As shown by the following comment, one teacher's preference was that most, if not all, of the information should be in a written format.

Teacher 5: *For me personally, a written presentation is better. I guess in an ideal world, what I would really want is for someone to go through the information, take out the key points and say "This is where you need to improve. These are the particular issues and genres that you need to focus on or in class to help your students perform better and just help their overall knowledge". And then present that in a written form.*

If charts were used then it was suggested that some written explanation should accompany it. An administrator commented that:

Administrator 1: *It is a good idea to give a few sentences that include simple observations that help to point out trends*.

This administrator highlighted the need to cater for all predispositions by including a variety of display formats in the presentation.

Administrator 1: *Some people prefer to look at the numbers whereas some see better from the pictures. Summaries should be entered into the chart.*

So testing agencies should acknowledge that teachers will have the different preferences and therefore they should use multiple formats to express the same data.

As shown by the following comments, the administrator suggested that testing agencies needed to stick to good statistical practices when presenting their data.

Administrator 1: *Diagrams should include keys, colour, charts and numbers. A key should be used and the value of each data point should be specified.*

The views of the display need to be standardised and a standardised scale should be used to give a better picture.

Scales should help teachers get a sense of where things are in that SAT scales need to be shown from 0 to 2400 not from say 1500-2400, percentage scales should go from 0-100

The following comments show that teachers wanted to have access to students' test papers.

Teacher 1: *Unless you know, you see the kid's paper and know exactly how they did individually on their particular questions and what they did, there's no useful feedback you can get from it.*

Teacher 5: *With the actual skills, it could be so much more useful if we actually have the type of questions that went along with it. So we would be able to see, well ok, they have problems reading bar graphs and this kind of thing. What type of question was given to them and what type of skills can we glean from that type of question that we need then to focus on in teaching?..... They give the actual skill which is good but it doesn't put the skill into context. ... It's a lot easier, when you are looking at the data and when you see the students' critical reading skills are low, you know exactly what kind of format the question is and whether it's a short passage, a long passage or two passages together. You know what's going on there.*

Sometimes it was difficult to contextualise the cause of a student's difficulty because only a summary of the error had been provided or the general theme of the question was given and consequently it was impossible to identify the exact cause of the error that the student had made. Teachers had found that when they had access to the students' responses, they could see errors that the testing agency might not highlight and that they would not be aware of otherwise. Without being able to see the response, it was difficult to know exactly what sort of misconception the student had

and to identify the best remediation strategies to benefit the student or class. Student error may result from misunderstandings that do not relate to the skill being tested by the question. If access was given to how the students had responded then teachers stated that they would be more able to identify the source of the misconceptions that the student had and hence would be in a better position to remediate their problems.

One of the administrators in the sample stated that it was not the presentation that was important but rather that schools needed to consider the way the data was introduced to teachers. He commented that:

Administrator 1: *For those that don't it is better to keep charts simple. Don't give too much too soon. Don't give more than is asked for – wait until they want to know. Giving too much before people are ready may overwhelm them and prove counterproductive.*

This administrator commented that in the early stages of presenting data it was important to acknowledge the different levels of ability and confidence that teachers had and to give consideration to what data is presented, how it is presented and how much is presented at one time so that teachers' confidence is not further undermined. If teachers are not allowed to build up comfort with their skills, they may be reluctant or refuse to engage with data at a later stage. However, teachers did not just identify changes to presentation as ways to improve data use. They commented on changes that could be made in the way the data was analysed so that more relevant information could be obtained.

However, one teacher expressed concerns about how skills were analysed and particularly how they were being grouped together when feedback reports were being compiled. Teachers want to be able to track the improvements that their students are making either by tracking their progress from year to year or by

comparing the errors that students in a specific grade make. This will help them to judge the effectiveness of any interventions that they have made. However, one English teacher's highlighted groupings that they found particularly unhelpful.

Teacher 6: *Anything that is short term. Anything that is, that doesn't have recurring skills. So what I mean by that, one might say, so almost too specific, if that makes sense.*

When asked to describe the type of skill classifications that would help, he stated that he wanted more general categories than were being given in some of the current reports he had received, but found that some tests made the categories too general.

Teacher 6: *Specific enough that you know exactly what it means but not an individual skill for almost every single question.*

Teacher 6: *So, "identify and label three parts of a graph with correct information" comes up on one question but doesn't come up anywhere else. And then again it will come up "explain a whatever".*

If skills are too specific then it is impossible to make judgements because if there is only one question that tests a particular skill in one year and then it may not be included in the following year. Long term monitoring becomes impossible as testing cannot incorporate enough specific skills to test curriculum content every year without becoming unwieldy. However, other teachers commented that they found the breakdown of skills to be useful.

Teacher 1: *What I do like is when they go through topics and talk to us about what the school average was on say differential calculus or integral calculus and they'll show us that we seem to be doing that topic ok.*

Teacher 5: *Yes, they give the actual skills which is good.*

Teacher 5: *It's also useful when you can identify particular skill that you teach in the classroom or that you don't teach in the classroom.*

Hence there was not universal agreement on whether the breakdown of skills within tests were useful.

One teacher also stated that it was essential for the analysis to include comparisons to students in other educational institutions. He commented that:

Teacher 6: *But they still give, it still gives a better indication if you've got something else to compare with. Because you may think you are doing the right thing but if you are looking and you see that a like school is consistently, that like schools are consistently doing better at one skill then you are over three years or something, then you probably know that there's more that you can do in that skill, like being more over about teaching it.*

This teacher wanted the comparisons between schools because that allowed for the identification of skills that the school needed to work on. The comparison allowed teachers to see that they were teaching appropriate skills to similar levels as other students of similar ability. However, there was a note of caution about how those comparisons are made in that care should be taken to make sure that comparisons are between groups that were truly alike. The teacher mentioned that one test used made comparison based on the ability of the students by looking for schools whose overall scores were similar and breaking down relative attainment in sub skills. However, another test compared schools based on the proportions of students within the school for whom English was not their first language. In the following comment, they expressed concern that making comparisons with schools who have similar percentages of second-language learners was not helpful.

Teacher 6: *As long as the comparisons are fair, that's the only problem. Like our ELL students are not necessarily the same as other schools' ELL students.*

It was pointed out that even though schools may have the same proportions of students learning in their second language, they were not necessarily alike. Students who are classed as second language learners have a large spectrum of abilities. At one end of the spectrum are students who are just starting to learn the language and have only learnt a small amount of vocabulary, while at the other end are students

who have been educated in English for a large part of their life or have lived in an English-speaking country thus giving them skills that would be at near native level. Some schools may have a majority of students who are at one end of the spectrum while other schools placed in the same category may be anywhere between these two extremes. The school with higher proportions of students who have been studying English for a longer time may outperform other schools based solely on the fact that their students have had more opportunity to develop their language skills. Differences in scores may reflect the difference in language acquisition within schools rather than the differences in actual progress that students are making. These comparisons may not be very helpful in judging how well the school is really doing or in highlighting areas of the curriculum for which there is a relative weakness and this may be a reason for questioning the validity of comparisons that are based on using the number of second language learners as a way of determining like schools. Teachers also commented on that the provision of appropriate technology could make the use of data easier.

5.5.1 Technological requirement

Teachers in international schools reported that they would like to have the ability to drill down into their testing data. As the following comment shows, they favoured presentations that allowed them to look deeper into the data rather than simply presenting them with the results.

Teacher 6: *maybe if you could, if you click each student and it shows you like the relative level that they are at in terms of, let's go back to those PSAT skills, right, you know "Understanding an author's craft", "identifying the meaning of words". If you could do that, like click each student and it shows you maybe where they're at on a bar chart compared to the school and then compared to like schools. That might help.*

One teacher commented on a feature he liked in one of the testing formats he received.

Teacher 1: *Theirs is quite good because they give you the chance to sort and filter the data. You can look at a subsection of students, you can compare grades, you can look at boys versus girls, you can look at non-English speakers versus English speakers. You can drill down to an individual student. There are a number of ways that you can format the way it looks.*

But he stated this was not available in all testing formats.

Teacher 1: *There is just one kind of report that they give me and I can't fiddle with the data and make it into some other shape. So the flexibility to do searches or to export it and to put it into an Excel document would be handy.*

One teacher did report that they had an internal system that had been developed by the Information Technology technician in their school. He reported that:

Teacher 1: *I know that the internal one that we have for gathering grade data here is fairly flexible and you can search it across by standards or down by grade or in departments or between departments. It's not too bad that way. But then its written by one poor programmer who's writing all our, he's looking after all of our computing needs so.*

The school's technician had developed the system in spare time using skills developed for other purposes. As there is no dedicated staffing for development of this system, it had taken a number of years to develop and did not have all the full capabilities of a commercially produced system but it was seen to be useful for the needs of this teacher.

5.5.2 Summary of section 5.5

Many of the teachers were unable to offer suggestions for improvements to the presentation of data as they lacked knowledge of different presentation types. Where suggestions were given, they were for summaries or information to be given in

written form rather than using graphs and statistics. It was suggested that the use of appropriate technology could increase teacher's ability to interpret data.

5.6 Discussion

All of the teachers in the sample stated that their schools wanted data to be used more. They reported that there were developments within their schools that were specifically aimed at getting them to work more with data. However, there seems to be differences in perceptions across the school about the amount of data use that is expected and in the methods that are being used to implement data use. Most of the teachers interviewed stated that they are using data within their schools but they did not have the ability to analyse data adequately to fulfil the requirements that were part of their role. Even those teachers who commented that they were able to do some data analysis found that there were barriers that prevented them from using data adequately; either they found that they were unable to get information from specific tests or they could not help colleagues in gaining understanding in interpreting the data for their classes. Some teachers expressed doubts about finding any usefulness for the testing data they received, restricting themselves to information gained only from their own teaching.

Teachers expressed a preference that data should be analysed for them and written information be given to them about the results of the analysis. However, such analysis may prove too generalised to be of any use to them. For instance, it may be difficult to sort the data so that they were just looking at the information regarding the students that they teach. Where subtest scores were used to identify areas of weakness in the score report, it was found to be difficult to identify which students

had difficulties in which skill areas and this may result in students having remediation they don't need or missing out on remediation that they do need.

There were differences in the perceptions regarding the purposes of the testing with some teachers in the sample stating that it did not give information that would help to inform their teaching but rather that it should be used to monitor the progress of their students. Teachers also identified some areas of concern that were particular to their situation in international schools. They commented that students were relatively less successful in English skills as they were not working in their first language and were still learning the language. Some of the mistakes that they tended to make were because there were concepts that were acquired by students who had been raised in a fully English background while as English learners, they had to take time to learn these skills. As has been mentioned earlier, the transience of students also led to problems both in monitoring the growth of individual students and in program evaluation. It was seen by one of the teachers that testing was not for use by teachers but was done for marketing purposes and to reassure parents regarding the standard of the education their children were receiving.

As it seems that there will always be the expectation that standardised testing will be use in international schools, it is important that there is a valid purpose for their use as otherwise the use of testing will serve only to waste student's time and reduce their educational opportunities and this in itself would be a misuse of testing. However, there remain a number of open questions such as:

> Are the validated uses and inferences of large-scale standardised tests transferable to international schools?

How can teachers be supported so that they can make better use of the data that they receive from standardised testing?

These questions will be discussed further in chapter 6.

Chapter 6

Discussion

6.1 Introduction

This study aimed to identify potential challenges that teachers in international schools need to be aware of when trying to interpret data to make decisions about practices in their schools. To do this, it analysed the technical manuals provided by agencies for the testing that is used by international schools. It then explored the challenges faced by a sample of teachers who work in international schools when trying to use the data from standardised testing to inform their practice. It considered the expectation that schools placed on these teachers to use data and the factors that influenced their decisions regarding data use. This chapter will collate the main findings of the research and relate them back to the research questions.

6.2 What large-scale standardised tests are used by international schools?

The majority of the standardised tests that were used in the international schools were developed by testing agencies based in the US. These tests were the Iowa Assessments, the SAT suite of assessments, and the Measures of Academic Progress. The International Schools Assessment was the only test that was develop by an agency outside of the US. This test is specifically designed for use by international schools and comes from a testing agency in Australia. Three of the four schools in the sample used more than one test. They used one test in the earlier years and then switched to a different test as students entered the final years of schooling. The tests used in later years generally including some measure of college and career readiness. As will be discussed below, these measurements relate to US universities. Students and teachers need to be aware of this when students plan to attend university

programs outside of the US. According to Dunbar et al., (2015) no test can be equally suited to all school situations. Differences exist in terms of curriculum standards, instructional emphasis and characteristics of the student population. These factors may impact test scores even though they are irrelevant to the constructs being tested. Therefore, these factors need to be considered when making evaluations about the validity of testing inferences in the new population. Consequently, the first research question looked at the test manuals to identify elements that had the potential to introduce construct-irrelevant variance when these tests are used with international school populations.

6.3 How are the characteristics of international schools reflected in the validity arguments of large-scale standardised tests?

The use of testing from the US in international schools is an example of testing that is being used outside of the purpose it was designed for. Such tests are described by Oliveri et al., (2015), as exported tests. When a test is used in such a way, there are two potential sources of construct-irrelevant variance that need to be considered. The potential mismatch between the curriculum that is taught in school and the content domain of the test is the first source that must be considered. The second relates to population characteristics such as language and cultural background. These potential threats will be considered below. The next section will discuss how curriculum differences can affect the validity of inferences made from testing

6.3.1 How do the standardised tests that are used reflect the curricula used in international schools?

We are told by Kane, (2006, 2016b) that establishing validity requires questioning how well a test covers the content domain. According to Dunbar et al., (2015), when

deciding to adopt testing into a particular context, it is essential to establish alignment between local educational standards and the testing materials. They acknowledge that this is a time-consuming process, but they state that this is the only way that differences between testing and local curriculum standards can be discerned. Three of the tests used are from the US. The IOWA test uses the Common Core standards. While MAP testing allows international schools the possibility of choosing a test that is better matched to their curriculum, there are a limited number of choices and most relate to US curricula. As has been pointed out by Thomas & Goldstein, (2008) education systems differ in terms of content and aims. Tests will reflect the content that is considered important within a nation and Catling (2017) acknowledges that there are only superficial similarities in curricula across different contexts, including in international schools. Because there is a strong likelihood that there are differences between the content that is being tested and the curriculum content that is taught in international schools, it is difficult to be certain of the validity of any inferences that could be made. It is therefore important that the validity arguments relating to content are evaluated before such testing is implemented into an international school. The Standards (AERA et al., 2014) tell us that test developers work from specifications of the content domain. Yet two of the tests, the PSAT and the ISA, did not list their curriculum standards. The ISA states that it is designed to measure core skills rather than a particular curriculum. It is argued by Nardi (2008) and Thomas & Goldstein, (2008) that it is not possible to develop a test without a curriculum. Meanwhile, Thomas & Goldstein, (2008) question whether it would be desirable for a test not to be associated with a curriculum. According to Bates, (2011) the ability for such tests to claim validity across the international schools system with its complex curricular differences is a

"moot point" (p. 429). Dunbar et al., (2015) comment that "content quality is thus the essence of arguments for test validity" (p. 22). Understanding how content domain relates to the circumstances of an individual school means that there needs to be a detailed list of the content that is included in testing. Schools will struggle to identify the impact that curricula differences have on their student's scores. Where the testing agencies are drawing the questions from a pre-determined list of skills, it would be helpful if those skills were included in the documentation that is shared with test users.

All of the tests give subtest score information that are designed to give information about different content domains. Because of the diverse curricula that are used in international schools, this leads to questions about the ability of such tests to provide valid information within those subdomains. It was acknowledged by an English teacher in the sample that such feedback should focus on wide skills. However for a mathematics teacher, being told that a student is a year behind in their "Heart of Algebra" skills raises questions about the skills that are required by a student under this subtest score. The only way to determine this is to search through the test, which might then lead to a form of teaching to the test which again results in threats to validity. It would be helpful to have more definition on the skills that are included in testing. However, differences in curriculum specifications are not the only potential source of construct-irrelevant variance. Population differences can also lead to concerns

6.3.2 Population differences

As stated by Kane (2006), when the testing population differs from that used in the validation studies, there is the potential for difference characteristics in the new population to undermine validity arguments. Sears (2015) highlights the diversity of the international school population with schools that may have students from over 60 nationalities. These students have a wide variety of linguistic and cultural backgrounds. When teachers are looking at performance of international schools on standardised tests, both language ability and cultural background may reduce the validity of inferences that can be made. These differences will be considered below.

6.3.3 What are the potential threats to validity that teachers need to be aware of when using large-scale standardised tests in international schools?

One identified concern that was commented on by two teachers in the interview sample was that students' answer to questions in standardised testing were impacted by differences in language ability. As noted by Kieffer et al., (2009) there are unique challenges when ELL students take standardised assessments. All tests require a certain amount of language ability and as Abedi, (2002, 2008) states, differences in performance by ELL students may be due to lack of English proficiency rather that deficits in the content being tested. Three of the tests included ELL students in their standardising procedures, but in different ways. With ISA and MAP, questions are pre-tested using previous testing cohorts which will include ELL students. However, there is no specific mention of procedures that check for differences in item functioning for ELL students. The IOWA testing does mention the inclusion of students who are ELL in its development guide (Dunbar et al., 2015). Here, students could be excluded from testing or accommodations could be provided. However, again there is no mention of procedures to identify

differential functioning of questions based on language status. They focus more on the provision of accommodations to support ELL students. With SAT, there was no specific mention of inclusion rates of ELL in standardising samples. Kieffer et al's., (2009) work highlights the need to ensure that interpretations that are validated for other learners also apply to ELL students. Research is needed to identify the processes necessary to ensure that ELL students are able to demonstrate their learning in standardised testing.

Ensuring the validity of testing for the ELL population is highlighted as a pressing challenge for test developers (Schwabe et al., 2016). As Staehr Fenner (2016) points out, it is important that ELL students are provided with appropriate testing accommodations when completing testing. While all of the tests used offered accommodations for ELL students, the amount of support varied with the different tests. The most extensive support is offered by MAP testing which gives schools the opportunities to allow their students to use bilingual dictionaries as well having a setting which can provide text-to-speak for all testing components. The ISA allows the use of bilingual dictionaries but restricts their use to mathematics and science components. While the SAT suite of assessments does identify that ELL students may have different needs, it essentially offers no support to students attending international schools. Bilingual glossaries are offered but these are only available to students who are taking tests in the US. They do provide text-to-speak options but again these are limited to students who have identified additional needs rather than for ELL students. It is suggested by David (2011) that students may benefit from different accommodations at different stages of learning English. It is essential that testing agencies provide all necessary accommodations to ELL students so that they

can demonstrate their understanding in the standardised test. All ELL students should have access to glossaries and dictionaries, as there is research that suggests these may be beneficial to them (Kieffer et al., 2009). However, there is debate about which other accommodations are actually helpful to ELL students, (Kieffer et al., 2009). The population of ELL students in the US is described as large and increasing (Young, 2009). There has also been an increase in the size of the international population taking standardised tests designed for US populations (Oliveri & von Davier, 2016). There is a need for more research to discover which accommodations work and the stage of language development each would be most beneficial for. This should include research that separates students by language background. For instance, are there differences in how helpful an English dictionary is if the student's primary language does not use the Roman alphabet? However, as noted by Oliveri, Lawless, & Mislevy, (2019), language is not the only potential source of construct irrelevant variance when tests are used in new situations. Cultural differences can also result in threats to the validity of inferences.

6.3.4 How do the standardised procedures for the tests that are used reflect the population in international schools in terms of culture?

Cultural differences will be reflected in the assessments from individual countries (Brown, 2002). As Hayden (2006) tells us, there are differences to the way in which a student will respond to test questions which may be as a result of their cultural background rather than their ability in the subject being tested. The results of such tests can be confounded by the lack of cultural understanding, and that lack of understanding may lead to invalid scores and it is important for teachers to be aware of this when examining testing data. Each of the development procedures for US tests includes readings that are specifically looking check for cultural bias, the

people carrying out these checks are focused on US culture. As noted by Valdes & Figueroa, (1996) tests are culturally biased towards those who have the appropriate cultural capital. Where it is acknowledged that a significant number of test takers will not have experience of US culture, there needs to be the opportunity to identify where cultural references could have an impact on test results. Hayden (2006) asks if the inferences made from tests which include questions with settings that are unfamiliar to some students are equally valid for all groups of students. However, as is acknowledged in the technical manuals for the ISA (ACER, 2019b), designing a test that is truly culturally neutral is likely to be impossible. The US tests all carry out DIF analysis to check for on difference in performance amongst ethnic groups.

As pointed out by Oliveri & Lawless (2018), there can be challenges when using DIF classifications with exported testing. The groupings used in US tests may reflect identifiers that members of the international school population would use but there is a mismatch in the way they are being used. The classifications used by the test developers are for US citizens who have given cultural heritages. They will have experienced living in the US and will have some knowledge of the culture. There is a high likelihood that international school students who use these identifiers are relating it to their nationality and they will have had no experience of living in the US or its culture. The cultural understandings of students in the international populations will be different and therefore the DIF procedures used may not identify questions that may be problematic for these students. The research went on to consider if teachers in international schools used the results from standardised testing. It is acknowledged by Hayden (2006) that there is more awareness of the impact of cultural and linguistic differences on the results of testing and so testing

agencies are doing much to minimise the effects of these variables. However, she goes on to say that it would be remiss to claim that the problems relating to validity for students of different linguistic and cultural backgrounds does not remain an issue.

6.3.5 Measures of college and career readiness

The US tests all include measures of academic readiness but the measures relate to students' readiness to take on programs in the US university system. As with other factors already discussed, there are significant cultural differences in the relationship between secondary and tertiary education and this leads to differences in the requirements that countries place on students who are seeking to gain a place in university. In the case of US universities, Bates (2011) comments that admission is based on a process of testing that is unrelated to the curriculum or the pedagogical practices used in schools but instead is based on standardised testing in which performance is compared to a large population of students. However, Wiliam (2010) acknowledges a difference in the perspective of the purposes of schooling between the US and European countries. The European tradition was that education beyond the age of 15 was only for the 5% to 10% who were planning to pursue post-secondary education, while in the US there has been a belief that education is valuable as preparation for adulthood and so all young people were expected to remain in education until the age of 18. This historical difference has resulted in a difference in both the format of education for students between the ages of 15 and 18 and the requirements for entry into university. Few UK universities will consider the SAT in their admissions decisions but instead will require students to gain passes in curriculum-based subject qualifications such as A-level, AP or IBDP. US universities are themselves starting to turn away from using the results of admissions

tests as a part of their requirements, with more universities making the requirement to submit results optional for applicants. Furthermore, in the UK, students' courses are determined by the major they choose to study at university. Hence, they may not go on to take courses in the subjects that are indicated in the college readiness measures. Meanwhile, in US universities, students are likely to have a general education component with a requirement to take courses in mathematics, English, social studies and science, regardless of the major they are taking

According to Chester (2018), it is not possible to capture all the dimensions of college and career readiness accurately as it is a multidimension construct and academic preparation is only one consideration. For instance, Conley (2007) recognises that students need to possess behavioural attributes such as time management and appropriate study skills, attitudinal qualities such as persistence and an awareness of their level of performance. He goes on to highlight that there is a cultural component to college readiness that means that students from some communities have possession of certain privileged information that means they are better placed to understand such things as how to apply to college, they know the difference in the requirements for success at college as opposed to high school, and they know how to react with the educational professional and peers while there. This cultural component may vary according to the country the university is based in. As many of the students in the Japanese international school system will go on to higher education programs in many countries around the world they may find that meeting the requirements for readiness in these standardised tests is not enough to show they are ready for other country's university systems.

6.4 How is data from standardised testing currently being used as demonstrated by a sample of teachers drawn from four international schools based in the Kanto Plains region of Japan?

As identified by researchers such as Hattie, (2005); Mandinach, (2012) and Schildkamp & Ehren, (2013), there has been a drive for schools to use data to inform practice in many national settings. However, there is little if any research about the position in international schools. The second research question sought to find out about use of standardised testing data in a small sample of schools. The question was separated into several sub-questions and so each of these sub-questions will be addressed individually.

6.4.1 What requirements do these schools place on teachers to use data from tests?

Like schools in national contexts (Pierce & Chick, 2011b; Schildkamp & Ehren, 2013), these international schools were trying to make more use of data. It was indicated by teachers from three of the schools represented in the sample that their schools were in the process of introducing initiatives to increase the amount of data use. However, as Coburn & Turner, (2011) found when researching national contexts, data use was introduced into schools using a variety of initiatives.

In one of the schools, they acknowledged that teachers had different capacities to work with data (Datnow & Hubbard, 2016). Two teachers who had a particular interest led an initiative to increase data use. They had analysed data from standardised testing and identified an area where students were underperforming. As stated by Gotch & Roduta Roberts (2018), data is analysed with the aim of improving outcomes for students. The analysis in this school led to the introduction of a new teaching program aimed at improving attainment. As Bruniges, (2011) and

Ingvarson (2005) warn, teachers can lack the pedagogical information required to help students close gaps. To facilitate teacher collaboration in identifying teaching strategies that would improve student learning, the school introduced Professional Learning Communities (PLC). According to Dunlap & Piro, (2016) teachers prefer to socially construct knowledge to identify effective teaching strategies and the establishment of PLCs enabled them to do this. The program involved regular testing and teachers were required to set targets for their students. While classroom teachers were not required to analyse the standardised testing data, they were required to present a written analysis of student results based on the program's assessments back to the senior teacher. The standardised testing data was used to monitor the performance of the students results after the implementation of the curriculum program.

Research by Means et al., (2011) found that teachers were more confident and more likely to use data if they could work with colleagues. An example of this was seen in the second school. Here, a less structured approach was adopted in which teachers were encouraged to explore using data to inform their practice. However, there were no specifications as to how teachers show do that. Teachers were encouraged to work together and as a result, analysis was going on both within and between departments in the school. The informal procedures were aimed at introducing a more formalised policy once teachers had developed confidence by working together.

As Supovitz (2012) acknowledges, teachers have access to a wide variety of data from assessments the carry out in their classrooms. The initiative in the third school

was designed to get teachers looking at that data. It was based on Hattie's (2010) meta-analysis of educational research and the identification of practices that could improve student learning. While analysis focused on classroom-based assessments, there was an expectation that standardised testing data should be analysed to support their classroom practice.

The schools all seemed to be very much at the beginning introducing their data initiatives. As Stoelting's (2019) experience highlights, international schools have a transient teaching population and initiatives can stall as the teaching population changes. In light of this, the long-term success both in terms of using data and the impact on learning in the school would be areas for further investigation. One of the teachers who was a major driver of the initiative in the school C has already left. It is possible his departure could have had serious implications on the continuation of the initiative. Given that the use of data-driven methods is considered to be an important measure to improve school attainment, and that the schools will have invested time and resources into developing appropriate measures within their schools, it is necessary for international schools to have procedures in place to ensure important initiatives are continued even after the staff who are responsible for their implementation have moved on. Teachers do experience challenges when initiatives to use data are introduced so these will now be considered.

6.4.2 What challenges do these teachers identify when trying to use data from tests?

Like teachers in the research carried out by Means et al., (2010) and Schildkamp et al., (2014), the international school teachers reported that they did not have enough time to analyse data adequately. The reasons given, such as having time to complete

their other responsibilities were also similar to those in national situations (Kerr et al., 2006). However, the teachers also commented on additional pressures that could result because teachers had to take on more roles as their international schools were small. For instance, one teacher also had the role of college counsellor, which would be a full-time assignment in larger schools. Teachers also reported time lags between when testing was taken and when the results were received. As Wasson (2009) points out, the usefulness of diagnostic information is questionable if it is received six months after the test has taken place. Accessibility was also identified as problematic by these teachers. They agreed with the conclusions of Lachat & Smith (2005) that it was important that data was stored centrally and teachers could access it easily. According to Kerr et al., (2006) data use is facilitated by ensuring teachers have easy access to the data and providing computer applications to support teachers in using the data. Teachers here also commented that data use would be easier if they had access to technology that would allow them to interrogate the data. However, the schools could not implement the sort of technology that is discussed by Wayman & Stringfield, (2006b) which is generally designed for school districts. There is a need for the development of software that will help individual small schools to sort and interrogate their data. The international school teachers also identified challenges caused by their school populations which are not identified specifically in the research on data use.

As Durán's (2008) work identifies, differences in the performance of ELL students in standardised tests can result from their differing language profile rather than from difference in ability. The teachers in the interview sample mentioned that the language background of their students could cause additional problems when trying

to analyse data. Where the questions from the testing could be analysed, mistakes that could be identified included those that were made because students were ELL and had not been immersed in English in the same way that students living in the US would be. Where initiatives were implemented to improve the English skills of the students in one of the schools, it was found that students did not show improvements when taking the tests even though improvements could be seen in the classroom. Given that tests are designed with a specific purpose in mind (Shepard, 2016), it may be that the changes in class were not reflected in the test's purpose. Improvement of teachers' assessment literacy skills (Popham, 2009) may result in finding measures that would better suit the purpose of monitoring the effects of initiatives in their student population.

As Walker (2017) tells us, there are frequent changes in the international school population with many students only attending the school for a short time, maybe two or three years. Teachers commented that transience makes tracking student progress difficult. The ELL population is heterogenous (Lane & Leventhal, 2015). The small year groups in international schools also meant that changes in the group composition in terms of the number and type of ELL students could bring about significant changes in group statistics which would be confounded with any changes in results due to progress within the student population. With the move to releasing more score reports in electronic formats, it would be useful to include options that allowed for schools to look at the progress made by students who could be tracked across different testing sessions. However, if data protection concerns could be resolved, it would also be useful if testing results were able to travel with students so

that schools could receive results from previous testing sessions that were carried out at previous schools.

Even when testing is designed with international school populations in mind, teachers commented that the heterogeneity of the population (Lane & Leventhal, 2015; Oliveri et al., 2019) could cause problems when group statistics were given. Because student composition could be very different even where the percentage of ELL students was similar, it was felt that using this for comparison was not helpful. One area of concern regarding teachers' ability to read data is their access to training and professional development.

6.4.3 What training and support have these teachers received to enable them to understand the score reports from the standardised testing that they receive?

Research, such as that by Mandinach & Gummer, (2012) highlights the need for teachers to receive professional development in data use. As Pierce & Chick, (2010) state, interpreting testing reports is probably beyond the ability of teachers who have not studied any statistical concepts. Yet eight of the nine teachers commented that they had not received specific training to help them to understand the data from standardised testing. For many, the only time they had been taught to use statistical concepts was during their own time at school. They found that the skills that they had acquired at that stage were not sufficient to help them understand the standardised testing reports that they were expected to interpret. Where training had been received by a teacher in the sample, it was at the request of the teacher rather than because the school had identified it as a priority. As Datnow & Hubbard, (2016) tell us, professional development is also important to address teachers'

attitudes and fears regarding data use. Having attended training, one teacher experienced resistance both to the result of the analysis and use of testing to help inform curriculum change. Given that most of the schools were at the beginning of their time using data, it is not known what future plans the schools had for providing professional development for their teachers. However, as Pierce, Chick, & Wander, (2012) comment, it is important that training is not a one-off event as teachers benefit more when they have access to long-term professional development. Given that the Standards (AERA et al., 2014) tell us that the validity of inferences is enhanced by making sure that those who are responsible for using the data have sufficient skills to interpret the results, it is essential that schools work on developing the assessment literacy skills of their teachers (Popham, 2018). This training should also aim to improve teachers' abilities to read the score reports that are sent to their schools. Current research has not yet identified the most successful methods to support teachers in developing the necessary skills for data use (Poortman, Schildkamp, & Lai, 2016; Sun, Przybylski, & Johnson, 2016). It is essential that more research to be done to establish the conditions under which teachers make the most progress in being able to use the data from testing.

Of course, one place that information is shared is in the technical manuals that test developers produce in response to AERA et al's., (2014) requirements to explain the process that they have gone through to validate appropriate uses for the testing they have developed. The documentation that is produced is designed for measurement professionals and is very technical in terms of the vocabulary used and the level of understanding that is required. In the case of international schools, the educators who are responsible for the implementation of testing policies are unlikely to be

measurement professionals. Their lack of understanding of educational data is likely to mean that they are unable to interpret these score reports and apply the information to their own situation. This means that there is a gap between the needs of the users who are trying to make decisions regarding the use of testing and the appropriateness of the inferences that they are looking to make and the documentation that is supplied. As Goldstein (2015) acknowledges, the users of tests are far removed from the developers of those tests. Connections need to be established between the measurement professionals who are responsible for deciding upon the presentation formats that are used in score reports and the teachers who need to use them. This connection could lead to greater understanding of the gap between the skills that teachers need to have to interpret the reports and the skills they actually have. This could lead to the production of more user-friendly reports or to the provision of better support materials. It may also be possible to identify where connections could be made between the language of measurement professionals and that of educational professionals.

6.5 Significance of the study

As has been stated, by Oliveri & Lawless (2018) and Oliveri et al., (2015), the use of standardised testing outside of the population for which it was designed is becoming a more common phenomenon. The study sought to consider some of the challenges that happen when tests are exported. As highlighted by Wendler & Powers, (2009) when a test is used outside of questions of validity arise because of differences between the intended and new populations. This is the case for many of the standardised tests used by international schools. There were clearly stated aims and warnings against uses outside of those that had been validated during the

development process in each of the technical documents released by the US testing agencies that were reviewed during this study. These tests were not designed with international school populations in mind. There are likely to be differences in the curriculum used and the test's content domain. Further, the linguistically and culturally diverse population does not match the population that the test is designed for.

The Standards (AERA et al., 2014) state that validity is the joint responsibility of both the test developer and the test user and they are very clear in the delineation of the responsibilities of both sets of users. While test developers are required to publish enough information to support any intended score interpretations in their documentation, the responsibility to determine the validity of inferences in a given situation is clearly assigned to the test user. It is the test user who will be aware of their own individual situation. Therefore, they will be in the better position to identify the potential threats to validity and to evaluate the consequences that could result from the interpretations made. In the case of international schools, this responsibility falls to the educators in the school. However, the study highlighted that the ability of these educators to fully understand the potential threats to validity in their situation is not guaranteed. The research demonstrated that teachers in the interview sample expressed reservations about their ability to understand data related to testing. This inability to understand the statistics related to testing data is likely to extend to knowledge regarding appropriate uses of tests and the effects of population differences on the validity of inferences that can be made. This highlights a need for educators to have a higher level of assessment literacy (Popham, 2018). This will mean they are better able to make decisions identify challenges to validity in their

particular situation and to make judgements about the appropriateness of the inferences they can make for the testing that they use.

The complexity of testing individuals who are culturally and linguistically diverse and the need to enhance validity of testing for this population has also been highlighted by Schwabe et al., (2016). The international school population is an example of a culturally and linguistically diverse population. As has been stated in chapter 2, it is a large and rapidly increasing population. However, with changing demographics across the world brought about by immigration, linguistically and culturally diverse populations are also to be found in many national settings. It is hoped that the analysis given in the research will give an insight into the challenges faced by some teachers and educational professionals who are trying to use exported tests

6.6 Limitations of the study

The study was carried out in a very small number of international schools which were all located within Japan. Even within that constraint, it was difficult to gain access to teachers. Schools reported that they received many requests for permission to conduct research so they had restrictions on who would be granted access. Consequently, there is no way that the research could be considered representative of all international schools.

Furthermore, the analysis is based mainly on testing from the US. This was the testing that was identified by schools in the sample. However, other sources of

testing exist. It is acknowledged that there is not a full picture of all testing formats that could be used within the international schools around the world.

6.7 Directions for future research

The thesis was focused on the use of standardised testing in international schools. Concerns regarding the use of such testing with linguistically and culturally diverse populations were highlighted. While the has been a lot of research looking into the educational needs of ELL students, there needs to be more research into how language background impacts their performance in large-scale standardised testing. The research should use a block design as differences in language structure and written script may alter the results of the research. That design should also categorise students by their language profile. For instance, it should consider the balance between the student's first and second language. It should also consider whether the language of testing is the language used in the country of residence. Further, there needs to be more research into the provision of accommodations for ELL students. This should identify the types of provisions that do support students and the point in their language development that the different provisions should be given. Again, the research needs to consider if differences in the structures or script of the student's first language alter the recommendations for the type of accommodation that would be most beneficial to the student.

In spite of the amount of research that has been carried out into how teachers can be supported in using data to inform practice, there is still no conclusive advice on the most appropriate ways to improve teachers' ability to reliably analysis and use the information from the data that is available in their schools. There is a need for more

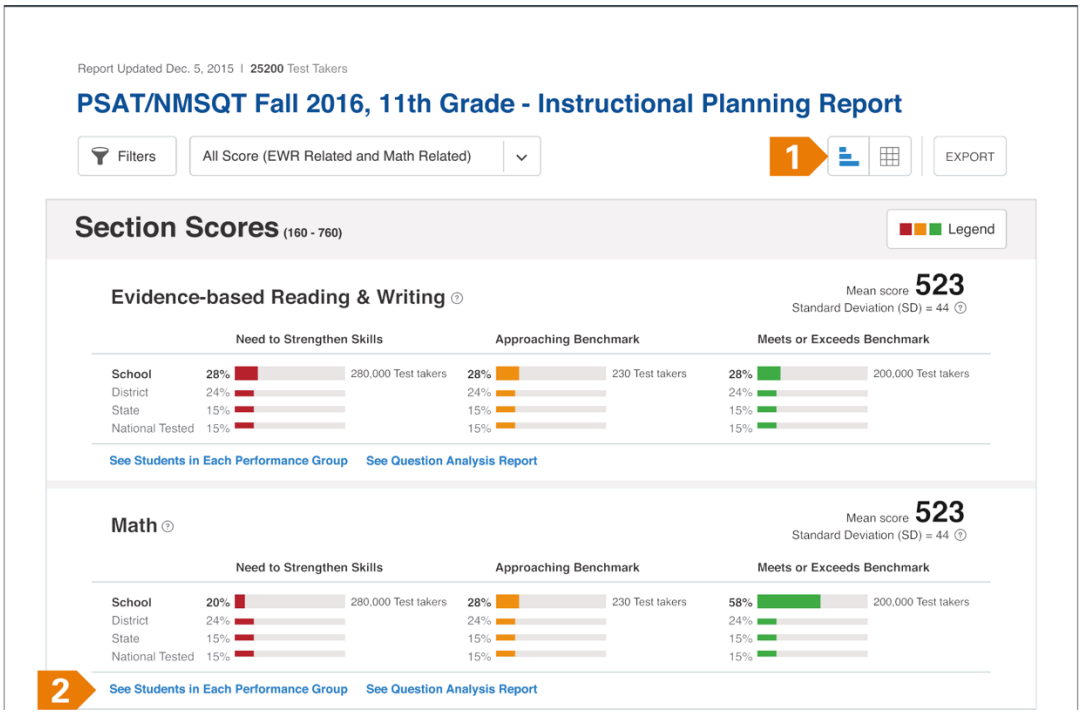
quantitative analysis that identifies best practice in introducing concepts and supporting teachers in developing their skills in the longer term. Further, given that the advice is that professional development should be long-term, research needs to be carried out to identify the best way to provide such structured training to teachers who are transient and unlikely to have access to long-term provision through their educational institution. Research could also be carried out that identifies the expectations that test developers have of test users' knowledge and skills. This could be used to identify what knowledge test users are missing and inform those who are responsible for developing professional development.

Appendix 1

Example of score report graphics

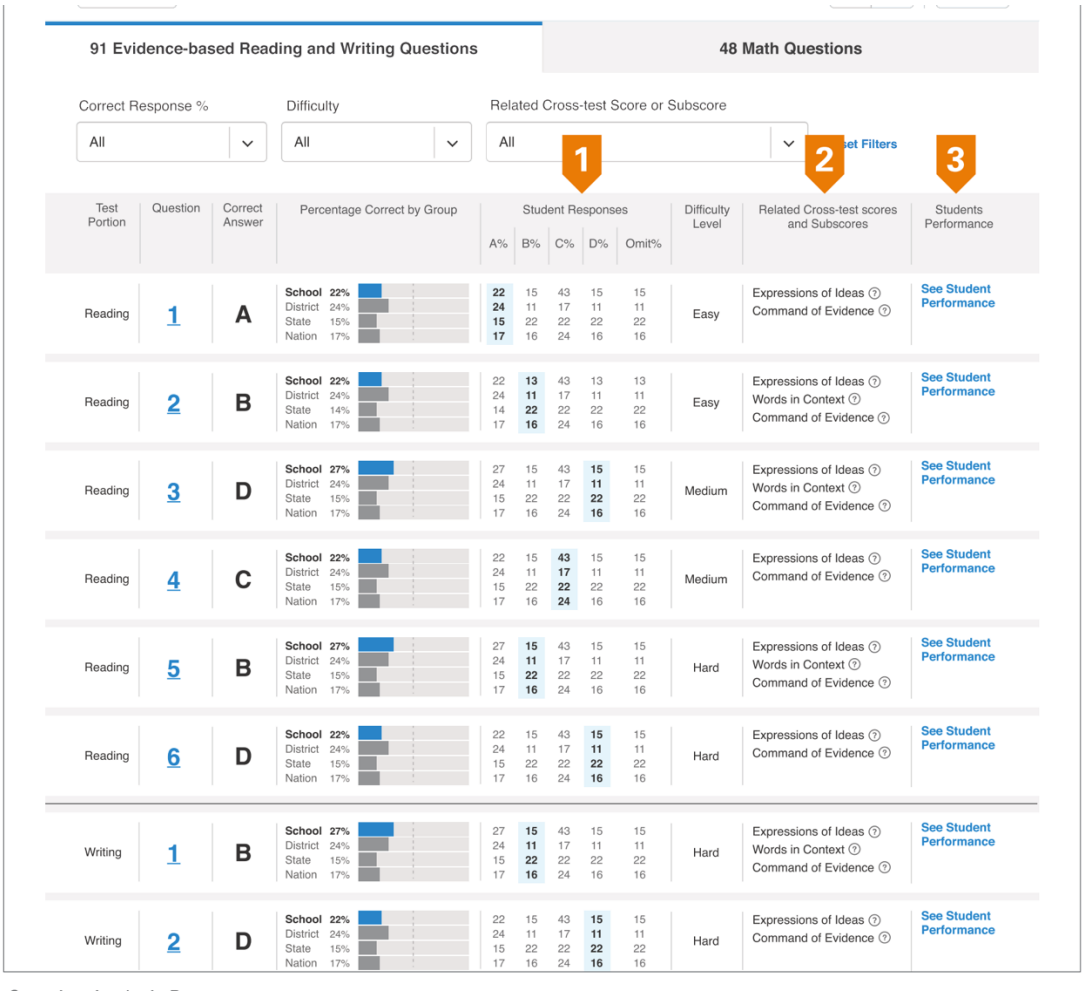
Samples from PSAT

Instructional Planning Report



(College Board, 2015d)

Question analysis report



91 Evidence-based Reading and Writing Questions | 48 Math Questions

Correct Response % | Difficulty | Related Cross-test Score or Subscore

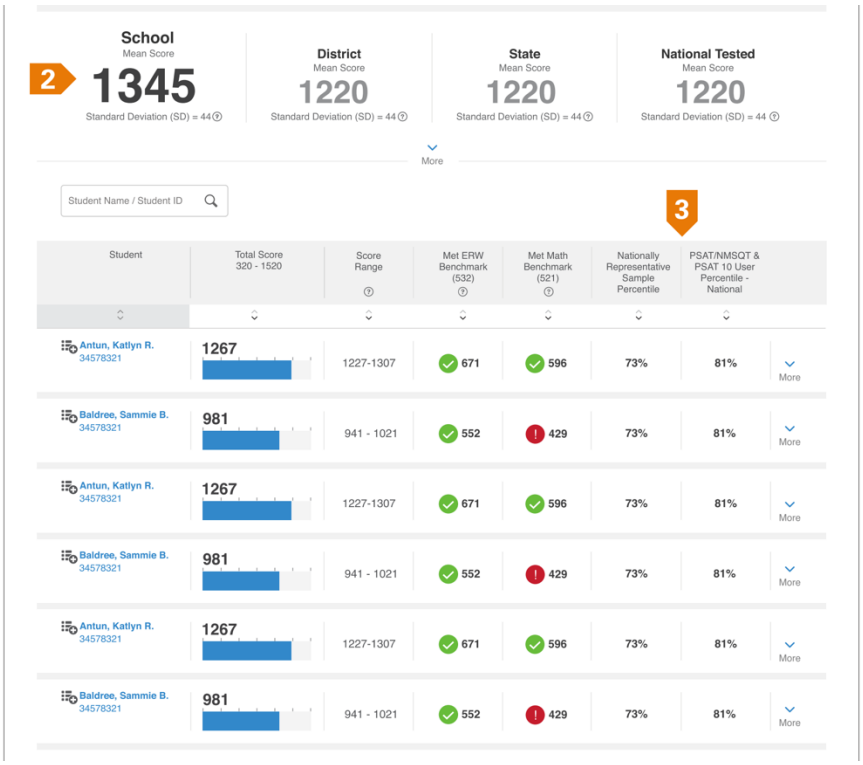
All | All | All

1 2 3

Test Portion	Question	Correct Answer	Percentage Correct by Group	Student Responses A%	B%	C%	D%	Omit%	Difficulty Level	Related Cross-test scores and Subscores	Students Performance
Reading	1	A	School 22%, District 24%, State 15%, Nation 17%	22, 24, 15, 17	15, 11, 22, 16	43, 17, 22, 24	15, 11, 22, 16	15, 11, 22, 16	Easy	Expressions of Ideas, Command of Evidence	See Student Performance
Reading	2	B	School 22%, District 24%, State 14%, Nation 17%	22, 24, 14, 17	13, 11, 22, 16	43, 17, 22, 24	13, 11, 22, 16	13, 11, 22, 16	Easy	Expressions of Ideas, Words in Context, Command of Evidence	See Student Performance
Reading	3	D	School 27%, District 24%, State 15%, Nation 17%	27, 24, 15, 17	15, 11, 22, 16	43, 17, 22, 24	15, 11, 22, 16	15, 11, 22, 16	Medium	Expressions of Ideas, Words in Context, Command of Evidence	See Student Performance
Reading	4	C	School 22%, District 24%, State 15%, Nation 17%	22, 24, 15, 17	15, 11, 22, 16	43, 17, 22, 24	15, 11, 22, 16	15, 11, 22, 16	Medium	Expressions of Ideas, Command of Evidence	See Student Performance
Reading	5	B	School 27%, District 24%, State 15%, Nation 17%	27, 24, 15, 17	15, 11, 22, 16	43, 17, 22, 24	15, 11, 22, 16	15, 11, 22, 16	Hard	Expressions of Ideas, Words in Context, Command of Evidence	See Student Performance
Reading	6	D	School 22%, District 24%, State 15%, Nation 17%	22, 24, 15, 17	15, 11, 22, 16	43, 17, 22, 24	15, 11, 22, 16	15, 11, 22, 16	Hard	Expressions of Ideas, Command of Evidence	See Student Performance
Writing	1	B	School 27%, District 24%, State 15%, Nation 17%	27, 24, 15, 17	15, 11, 22, 16	43, 17, 22, 24	15, 11, 22, 16	15, 11, 22, 16	Hard	Expressions of Ideas, Words in Context, Command of Evidence	See Student Performance
Writing	2	D	School 22%, District 24%, State 15%, Nation 17%	22, 24, 15, 17	15, 11, 22, 16	43, 17, 22, 24	15, 11, 22, 16	15, 11, 22, 16	Hard	Expressions of Ideas, Command of Evidence	See Student Performance

(College Board, 2015d)

Scores by institution



(College Board, 2015d)

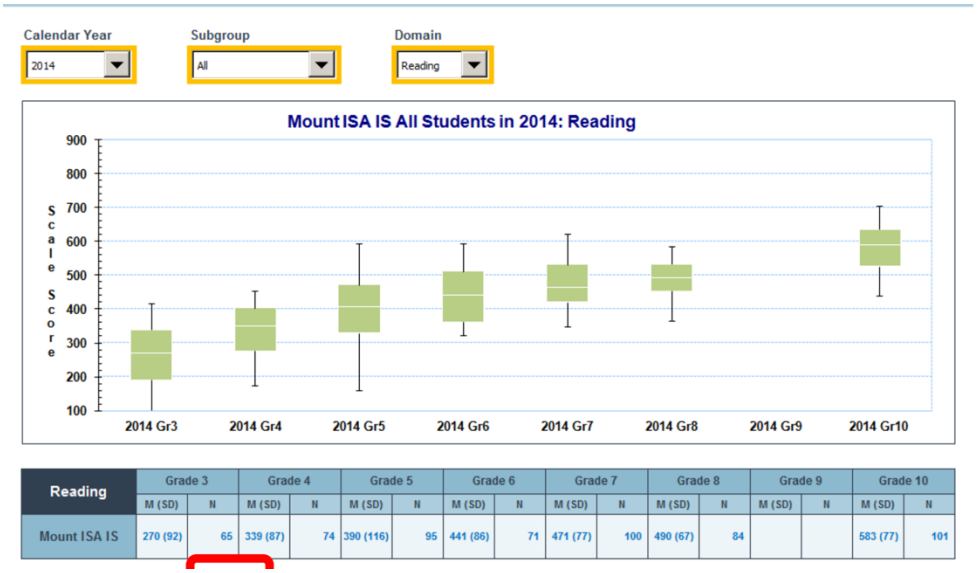
Samples from ISA

School Report

Domain		Mathematical Literacy				Reading			
		n[3]	*mean*[4]	*S.D.*[5]	*significance*[6]	*n*	*mean*	*S.D.*	*significance*
All	This school [1]	37	499	(66)		40	424	(105)	
	All other schools	3766	499	(80)	N	3901	423	(94)	N
	Other like schools [2]	550	503	(76)	N	637	446	(88)	N
Males	This school	21	517	(66)		21	466	(96)	
	All other schools	1955	508	(82)	N	2022	416	(95)	Y
	Other like schools	298	511	(75)	N	333	441	(87)	N
Females	This school	16	476	(59)		19	378	(98)	
	All other schools	1808	490	(77)	N	1876	431	(93)	Y
	Other like schools	252	494	(76)	N	304	452	(88)	Y
English speaking background	This school	27	502	(69)		30	431	(105)	
	All other schools	1475	501	(77)	N	1531	456	(91)	N
	Other like schools	375	502	(75)	N	419	462	(84)	N
Non-English speaking background	This school	10	n/a	n/a		10	n/a	n/a	
	All other schools	2291	498	(82)	n/a	2369	402	(90)	n/a
	Other like schools	175	506	(78)	n/a	217	415	(86)	n/a

(Australian Council for Educational Research (ACER), 2009)

Overall - Displays performance by grade of all students in a given calendar year.



Reading	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		Grade 9		Grade 10	
	M (SD)	N	M (SD)	N	M (SD)	N	M (SD)	N	M (SD)	N	M (SD)	N	M (SD)	N	M (SD)	N
Mount ISA IS	270 (92)	65	339 (87)	74	390 (116)	95	441 (86)	71	471 (77)	100	490 (67)	84			583 (77)	101

(Australian Council for Educational Research (ACER), n.d.)

Snapshot - Highlights strong or weak performance by grade level of all students

Mount ISA IS All Students in ISA 2013 (Mean Z-Score)				
Grade	Mathematical Literacy	Reading	Narrative/Reflective Writing	Expository/Argumentative Writing
Grade 3	-0.2	0.4	0.1	0.3
Grade 4	0.2	0.2	0.0	0.0
Grade 5	0.1	0.1	-0.1	-0.1
Grade 6	0.3	0.3	0.1	0.2
Grade 7	0.0	0.1	0.2	0.2
Grade 8	0.4	0.2	0.3	0.2
Grade 9	0.2	0.5	0.6	0.4
Grade 10	0.2	0.6	0.3	0.2

(Australian Council for Educational Research (ACER, n.d.)

Appendix 2

Documentation used for analysis of testing

Test	Documentation
SAT suite of assessments	SAT Suite of assessments technical manual: Characteristics of the SAT (College Board, 2017b) SAT technical manual: Appendixes (College Board, 2017c) SAT technical manual: Appendix B (College Board, 2017d) PSAT/NMSQT Understanding scores 2015 (College Board, 2015c) PSAT 8/9 Understanding scores 2016 (College Board, 2017a) Content alignment – SAT suite of assessments (College Board, 2019a) Key features – SAT suite of assessments (College Board, 2019b) Math content alignment – SAT suite of assessments (College Board, 2019c) Reading content alignment – SAT suite of assessments (College Board, 2019d) SAT suite of assessments: Educator Guide (College Board, 2019e) Writing and language content alignment – SAT suite of assessments (College Board, 2019f) Counselor resources for the redesigned SAT (College Board, 2015a) Official educator guide to the PSAT/NMSQT and PSAT related assessments (College Board, 2015b) PSAT/NMSAT Understanding scores 2015 (College Board, 2015c) The SAT suite of assessments: Using scores and reporting to inform instruction (College Board, 2015d) Test specifications for the redesigned SAT (College Board, 2014)
ISA	About the ISA (ACER, 2019a) ISA 2018-2019 Information handbook (ACER, 2019b) ISA quick guide (ACER, 2019c) ISA 2019-2020 School Coordinator's Handbook (ACER, 2019d) ISA International Schools' Assessment Program (ACER, 2015) Assessment and student learning: Collecting, interpreting and using data to support learning (ACER, 2009)
IOWA	Technical summary for Form F of the IOWA Assessments (Welch, Dunbar, & Fina, 2018) Forms E and F research and development guide (Dunbar et al., 2015) Measuring growth with the IOWA Assessments: A black and gold paper (Welch, Dunbar, & Rickels, 2014)

	Content validity for large-scale assessment (University of Iowa, n.d.-a) Interpreting standard reports from the IOWA assessments (University of Iowa, n.d.-b)
MAP	Technical manual for Measures of Academic Progress (MAP) and Measures of Academic Progress for Primary Grades (MPG (NWEA, 2011) NWEA 2015 MAP norms for student and school achievement status and growth (Thum & Hauser, 2015) MAP College readiness benchmarks: A research brief (Thum & Matta, 2015)

Appendix 3

Ethics Approval

Ethics approval: T Pomroy

Wed 18/12/2013 09:10

To: POMROY P.J. <p.j.pomroy@durham.ac.uk>
Cc: COE R.J. ED-PGTSTUDENTS E. <ed.pgtstudents@durham.ac.uk>

Dear Trish
I am pleased to inform you that your application for ethical approval in respect of 'Enabling teachers to improve instruction by increasing their ability to use testing data' has been approved by the School of Education Ethics Committee.
May we take this opportunity to wish you good luck with your research.
Best wishes.

Sheena Smith
Research Office
School of Education
Durham University

Appendix 4

Interview Schedule

Explanation of the study and request to fill in permission form

Background Information
Subject taught, grade levels.

Personal experience with using testing data in education
Use of testing data at previous schools.
Any training courses attended.

School's policy on use of testing data
How does the school expect data from these tests to be used?
What access do teachers have to the reports from these tests?

Investigation into teacher's use of data
What do you hope to gain from the reports?
What is your purpose in using them?
If you don't use them, why not?

For each type of report used-
How do you use the reports that are supplied?
What do the reports tell you? How do you know?
Any comments on the information you can get from them.

What information in the reports is useful? What is not useful?
What information is missing?
Which types of presentation help you? Which types of presentation do you not understand?

How could the reports be improved?

What would help you make better use of the data given in these reports?

Appendix 5

Consent from Head Teachers/Principals

Dear

I am a teacher at an international school in Tokyo and a part-time student at Durham University.

I am requesting permission to interview teachers at your school for the research project which forms part of my EdD course at Durham University, UK.

The Research

Teachers are expected to use an increasing amount of data to evaluate their practices and monitor their students' progress. Score reports from standardised assessments from external testing organisations (e.g. College Board, IB, ACER) are one source of such data. These reports provide analysis of the patterns of student responses and are aimed at providing teachers with information about areas of strength and weakness in their curriculum. I want to find out if the information given in these reports is appropriate and whether teachers understand the formats used to present the data.

Methodology

In the first part of this project, I would like to interview Curriculum Coordinators and teachers of mathematics, English, science and social studies/humanities to find out their opinions of these reports.

As part of my investigation I would ask teachers explain about the types of reports that they use and what information they can gain from them. Although I do have some examples of the types of reports that are produced, it would be helpful if teachers can bring some samples of the types of report that they have access to. I understand that this information is sensitive and therefore specifically request your permission to do this.

Issues of Privacy and Confidentiality

The project is being carried out in accordance with the ethical guidelines of the British Education Research Association and under the supervision of the School of Education at Durham University. The Ethics Committee at Durham University has approved the research methodology. All information received will be confidential. The school and any teachers interviewed will not be named in any papers that may be produced from this work.

I would be grateful if you could send a brief response confirming that it is acceptable for me to interview teachers from your school and advising me as to whether I may ask teachers to bring examples of score reports to that interview.

Please contact me if you require further information. My email addresses are p.j.pomroy@durham.ac.uk

Thank you for your assistance.
Kindest Regards
Patricia Pomroy

Appendix 6

Participant Information Sheet

Enabling teachers to improve instruction by increasing their ability to use data from standardised assessment.

You are invited to participate in a research project which forms part of my Ed.D course at Durham University, UK. I want to investigate how teachers use the data from standardised testing reports.

Teachers are expected to use an increasing amount of data to evaluate their practices and monitor their students' progress. Score reports from standardised assessments from external testing organisations are one source of such data. These reports provide analysis of the patterns of student responses and are aimed at providing teachers with information about areas of strength and weakness in their curriculum.

In the first part of this project, I would like to interview teachers to find out their opinions of these reports and see how much useful information they can get from them. I want to find out if the information given in these reports is appropriate and whether teachers understand the formats used to present the data.

The interview will last about 45 minutes. It will be recorded and the data will be stored securely. As a participant in this project, you have the right to privacy and any data collected will be treated in the strictest confidence. Your name will not be used in any publication of the results from this project.

Participation in the project is completely voluntary and participants are at liberty to withdraw at any time without prejudice or negative consequences.

Appendix 7

Participant Consent Form

Enabling teachers to improve instruction by increasing their ability to use data from standardised assessment.

Researcher: Patricia Pomroy

- I have read the Participant Information Sheet and have had opportunities to ask questions about the study.
- I understand the research project and my involvement in it and I consent to take part in it.
- I understand that I will be digitally recorded during the interview and that hard and electronic copies will be made but that access will be restricted to the researcher and the supervisor.
- I understand that information gained in the interview may be published but that my results will remain anonymous and confidential.
- I understand that I am free to withdraw from the study at any time without prejudice or negative consequences

Name of Participant……..

Signed …….. **Date** ……

Contact Details:
Patricia Pomroy - p.j.pomroy@durham.ac.uk

Bibliography

Abedi, J. (2002). Standardized achievement tests and English language learners: Psychometrics issue. *Educational Assessment*, *8*(3), 231–257. https://doi.org/10.1207/S15326977EA0803

Abedi, J. (2006). Psychometric issues in the ELL assessment and special education eligibility. *Teachers College Record*, *108*(11), 2282–2303. https://doi.org/10.1111/j.1467-9620.2006.00782.x

Abedi, J. (2008). Classification system for English language learners: Issues and recommendations. *Educational Measurement: Issues and Practice*, *27*(3), 17–31. https://doi.org/10.1111/j.1745-3992.2008.00125.x

ACER. (2009). Assessment and student learning: Collecting, interpreting and using data to support learning.

ACER. (2015). ISA International Schools' Assessment Program.

ACER. (2019a). About the ISA.

ACER. (2019b). ISA 2018–2019 Information Pack.

ACER. (2019c). *ISA quick guide*.

ACER. (2019d). *ISA School Coordinator's Handbook*.

Adams, R. (2019). Secondary teacher recruitment in England falls short of targets. *The Guardian*.

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Alhamdan, B., Al-Saadi, K., Baroutsis, A., Du Plessis, A., Hamid, O. M., & Honan, E. (2014). Media representation of teachers across five countries. *Comparative Education*, *50*(4), 490–505. https://doi.org/10.1080/03050068.2013.853476

Anagnostopoulos, D. (2007). The New Accountability and Teachers' Work in Urban High Schools in the USA. *Oxford Studies in Comparative Education*, *17*(1), 119–135. https://doi.org/10.1177/0895904803254481

Angelico, T. (2005). An evidence approach to improvement: A case study of the Victorian Catholic sector. In *Using data to support learning (Conference Proceedings).*

Aragon, S. (2016). *Teacher Shortages: What We Know. Teacher Shortage Series*.

Au, W. (2007). High-Stakes Testing and Curricular Control: A Qualitative Metasynthesis. *Educational Researcher*, *36*(5), 258–267. https://doi.org/10.3102/0013189x07306523

Australian Council for Educational Research (ACER). (n.d.). How to use the ISA Interactive Tracking Report.

Australian Council for Educational Research (ACER). (2009). *A guide to interpreting the ISA data for School Leaders and Administrators*.

Avalos, B. (2011). Teacher professional development in Teaching and Teacher Education over ten years. *Teaching and Teacher Education*, *27*(1), 10–20.

https://doi.org/10.1016/j.tate.2010.08.007

Baird, K. (2017). *Trapped in mediocrity : why our schools aren't world-class and what we can do about it*.

Baker, E., Barton, P., Darling-Hammond, L., Haertel, E., Ladd, H., Linn, R., … Shepard, L. (2010). *Problems with the use of student test scores to evaluate teachers*.

Balf, T. (2014). The Story Behind the SAT Overhaul. Retrieved October 24, 2019, from https://www.nytimes.com/2014/03/09/magazine/the-story-behind-the-sat-overhaul.html

Ball, S. (2008). The education debate. Bristol: Policy Press.

Bartlett, K. (1998). International curricula: more or less important at the primary level? In M. Hayden & J. Thompson (Eds.), *International education: principles and practice*. London: Taylor & Francis.

Bates, R. (2011). Assessment and international schools. In R. Bates (Ed.), *Schooling internationally globalisation, internationalisation, and the future for international schools*. Milton Park, Abingdon, Oxon: Routledge.

Becker, D., & Pomplun, M. (2006). Technical reporting and documentation. In S. Downing & T. Haladyna (Eds.), *Handbook of test development* (1st ed.). Mahwah: Erlbaum.

Berliner, D. (2011). Rational responses to high stakes testing: The case of curriculum narrowing and the harm that follows. *Cambridge Journal of Education*, *41*(3), 287–302. https://doi.org/10.1080/0305764X.2011.607151

Bernhardt, V. (2004). Continuous Improvement: It takes more than test scores. *ACSA Leadership*, (December), 16–19. https://doi.org/10.1016/B978-0-12-801764-7/00009-7

Bernhardt, V. (2013). *Data analysis for continuous school improvement*.

Black, D., & Armstrong, P. (1995). Some aspects of staff development in international schools. *The International Journal of Educational Management*, *9*(5), 27–33. https://doi.org/10.1108/09513549510088426

Blazer, C. (2011). *The unintended consequences of high-stakes testing*. https://doi.org/10.5860/choice.41-6045

Bolhuis, E., Schildkamp, K., & Voogt, J. (2016). Data-based decision making in teams: enablers and barriers. *Educational Research and Evaluation*, *22*(3–4), 213–233. https://doi.org/10.1080/13803611.2016.1247728

Booher-Jennings, J. (2005). Below the Bubble: "Educational Triage" and the Texas Accountability System. *American Educational Research Journal*, *42*(2), 231–268. https://doi.org/10.3102/00028312042002231

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2003). The Theoretical Status of Latent Variables. *Psychological Review*, *110*(2), 203–219. https://doi.org/10.1037/0033-295X.110.2.203

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

Boudett, K. P., City, E., & Murnane, R. (2005). *Data wise: a step-by-step guide to using assessment results to improve teaching and learning*.

Bowen, G. (2009). Document Analysis as a Qualitative Research Method. *Qualitative Research Journal*, *9*(2), 27–40. https://doi.org/10.3316/QRJ0902027

Bowman, E. (2018). Reticence to Change in Education. *Journal of Thought*, *52*(1–2), 48–65.

Braun, V., & Clarke, V. (2006). Qualitative Research in Psychology Using thematic analysis in psychology Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101.

Brennan, R. (2006). *Educational measurement* (Fourth Edi). Westport, CT: Praeger Publishers.

Brennan, R. (2013). Commentary on "Validating the Interpretations and Uses of Test Scores." *Journal of Educational Measurement*, *50*(1), 74–83. https://doi.org/10.1111/jedm.12001

Brimijoin, K. (2005). Differentiation and High-Stakes Testing: An Oxymoron? *Theory into Practice*, *443*(3), 254–261. https://doi.org/10.1207/s15430421tip4403_10

Brown, G., & Hattie, J. (2011). Communicating Test Scores to Teachers: Moving from statistics to use., 1–8.

Brown, R. (2002). Cultural dimensions of national and international assessment. In M. Hayden, J. Thompson, & G. Walker (Eds.), *International education in practice: dimensions for schools and international schools.* [Place of publication not identified]: Routledge.

Bruniges, M. (2011). Developing and Implementing an explicit school improvement agenda, 1–13.

Bryman, A. (2015). *Social Research Methods.* Oxford: Oxford University Press ;

Buckendahl, C., Plake, B., Impara, J., & Irwin, P. (2000). Alignment of standardized achievement tests to State content standards: A comparison of publishers' and teachers' perspectives. In *Annual meeting of the National Council of Measurement in Education*.

Busher, H., & James, N. (2007). The ethical framework of research practice. In A. R. J. Briggs & M. Coleman (Eds.), *Research methods in educational leadership and management*. Los Angeles: Sage Publications.

Cambridge, J. (2011). International curriculum. In R. Bates (Ed.), *Schooling internationally: globalisation, internationalisation, and the future for international schools*. London: Routledge.

Carnoy, M., & Loeb, M. (2002). Does External Accountability Affect Student Outcomes ? A Cross-State Analysis Author. *Educational Evaluation and Policy Analysis*, *24*(4), 305–331.

Catling, S. (2017). Developing the curriculum in international schools. In S. Blandford & M. Shaw (Eds.), *Managing international schools.* [Place of publication not identified]: ROUTLEDGE.

Chandler, J. (2010). The role of location in the recruitment and retention of teachers in international schools. *Journal of Research in International Education*, *9*(3), 214–226. https://doi.org/10.1177/1475240910383917

Chester, M. (2018). Foreward: "A toast to Sofia and Hector's future." In K. L. McClarty, K. D. Mattern, & M. N. Gaertner (Eds.), *Preparing students for college and careers: theory, measurement, and educational practice*.

Cheung, F. M. (2004). Use of Western and Indigenously Developed Personality Tests in Asia. *Applied Psychology*, *53*(2), 173–191. https://doi.org/10.1111/j.1464-0597.2004.00167.x

Chick, H., Pierce, R., & Wander, R. (2014). Sufficiently assessing teachers' statistical literacy (Vol. 9, pp. 1–6).

Clark, N. (2014). The Booming International Schools Sector. *World Education News & Review*, (2012), 1–10.

Clarke, M., Shore, A., Rhoades, K., Abrams, L., Miao, J., & Li, J. (2003). Perceived effects of state-mandated testing programs on educators in low, medium and high stakes states. *National Board on Educational Testing and Public Policy, Chestnut Hill, MA.*

Coburn, C., & Talbert, J. (2006). Conceptions of evidence use in school districts: Mapping the terrain. *American Journal of Education*, *112*(4), 469–495.

Coburn, C., & Turner, E. (2011). Research on Data Use: A Framework and Analysis. *Measurement: Interdisciplinary Research & Perspective*, *9*(4), 173–206. https://doi.org/10.1080/15366367.2011.626729

Cohen, L., Manion, L., & Morrison, K. R. B. (2011). *Research methods in education. Seventh edition.*

Cohen, L., Manion, L., & Morrison, K. R. B. (2018). *Research methods in education*. London; New York: Routledge.

College Board. (2014). *Test Specifications for the Redesigned SAT*.

College Board. (2015a). *Counselor resource guide to the redesigned assessments*.

College Board. (2015b). *Official educator guide to the PSAT/NMSQT and PSAT-Related Assessments*.

College Board. (2015c). *PSAT/NMSQT Understanding scores 2015*.

College Board. (2015d). *The SAT suite of sssessment: Using scores and reporting to Inform Instruction*.

College Board. (2017a). *PSAT 8/9 Understanding Scores*.

College Board. (2017b). SAT suite of assessments technical manual: Characteristics of the SAT.

College Board. (2017c). SAT suite of assessments technical manual appendixes, (December).

College Board. (2017d). *SAT technical manual: Appendix B*.

College Board. (2019a). Content alignment - SAT suite of assessments. Retrieved September 9, 2019, from https://collegereadiness.collegeboard.org/about/alignment

College Board. (2019b). Key features - SAT suite of assessments. Retrieved October 5, 2019, from https://collegereadiness.collegeboard.org/about/key-features

College Board. (2019c). Math content alignment - SAT suite of assessments. Retrieved October 5, 2019, from https://collegereadiness.collegeboard.org/about/alignment/math

College Board. (2019d). Reading content alignment - SAT suite of assessments. Retrieved October 5, 2019, from https://collegereadiness.collegeboard.org/about/alignment/reading

College Board. (2019e). *SAT suite of assesments: Educator guide*. *New York*. https://doi.org/10.1186/1471-2458-13-275

College Board. (2019f). Writing and language content alignment - SAT suite of assessments. Retrieved October 5, 2019, from https://collegereadiness.collegeboard.org/about/alignment/writing-language

Confrey, J., Makar, K., & Kazak, S. (2004). Undertaking data analysis of student outcomes as professional development for teachers. *Zdm*, *36*(1), 32–40. https://doi.org/10.1007/BF02655755

Conley, D. (2007). *Toward a More Comprehensive Conception of College Readiness*.

Council of International Schools. (n.d.-a). Partners - CIS Council of International Schools. Retrieved February 9, 2020, from https://www.cois.org/for-schools/international-accreditation/partners

Council of International Schools. (n.d.-b). Rubric for Domain D.

Creswell, J. W. (2012). *Educational research: planning, conducting, and evaluating quantitative and qualitative research*. Boston, Mass.: Pearson.

Crooks, T., Kane, M., & Cohen, A. (1996). Threats to the valid use of assessments. *International Journal of Phytoremediation*, *21*(1), 265–286. https://doi.org/10.1080/0969594960030302

Darling-Hammond, L., & Rustique-Forrester, E. (2005). The Consequences of Student Testing for Teaching and Teacher Quality. *Yearbook of the National Society for the Study of Education*, *104*(2), 289–319.

Datnow, A., & Hubbard, L. (2016). Teacher capacity for and beliefs about data-driven decision making: A literature review of international research. *Journal of Educational Change*, *17*(1), 7–28. https://doi.org/10.1007/s10833-015-9264-2

David, J. (2011). Research says.../ High-stakes testing naroows the curriculum. *Educational Leadership*, *68*(6), 78–80.

Deakin, H., & Wakefield, K. (2014). Skype interviewing: reflections of two PhD researchers. *Qualitative Research*, *14*(5), 603–616. https://doi.org/10.1177/1468794113488126

Dee, T. S., & Jacob, B. A. (2010). The impact of no child left behind on students, teachers, and schools. *Brookings Papers on Economic Activity*, (2), 149–194. https://doi.org/10.1353/eca.2010.0014

Denscombe, M. (2017). *The Good Research Guide for small-scale social research projects*. London: Open University Press.

Diamond, J. B. (2012). Accountability policy, school organization, and classroom practice: Partial recoupling and educational opportunity. *Education and Urban Society*, *44*(2), 151–182. https://doi.org/10.1177/0013124511431569

Doerr, H., & Jacob, B. (2011). Investigating secondary teachers' statistical understandings. *Proceedings of the Seventh Congress of …*.

Drennen, H. (2002). Criteria for curriculum continuity in international education. In M. Hayden, J. Thompson, & G. Walker (Eds.), *International education in practice: Dimensions for schools and international schools*. [Place of publication not identified]: Routledge.

Dunbar, S., Welch, C., Hoover, H., Forsyth, R., Frisbie, D., & Ansley, T. (2015). *Forms E and F Research and Development Guide*.

Dunlap, K., & Piro, J. S. (2016). Diving into data: Developing the capacity for data literacy in teacher education. *Cogent Education*, *3*(1132526). https://doi.org/10.1080/2331186X.2015.1132526

Durán, R. (2008). Assessing English-Language Learners' Achievement. *Review of Research in Education*, *32*(1), 292–327. https://doi.org/10.3102/0091732X07309372

Earl, L. (2005). From Accounting to Accountability: Harnessing Data for School Improvement. In Australian Council for Education Research (ACER). (Ed.), *Using data to support learning (Conference Proceedings).*

Earl, L., & Katz, S. (2002). Leading Schools in a Data Rich World. In K. Leithwood & P. Hallinger (Eds.), *Second International Handbook of educational leadership and administration* (pp. 1003–1024). Dordrecht, Netherlands: Kluwer.

Eduers. (2009). The History of SAT. Retrieved September 9, 2019, from https://www.eduers.com/sat/history_of_sat/

European Comission. (2009). *National Testing of Pupils in Europe : Objectives , Organisation and Use of Results*. *Lifelong Learning*.

Ferrara, S. (2014). Formative Assessment and Test Security: The Revised Standards Are Mostly Fine; Our Practices Are Not. *Educational Measurement: Issues and Practice*, *33*(4), 25–28. https://doi.org/10.1111/emip.12050

Ferrara, S., & Lai, E. (2016). Documentation to support test score interpretation and use. In S. Lane, M. Raymond, & T. Haladyna (Eds.), *Handbook of test development* (2nd ed.). New York: Routledge.

Figlio, D., & Loeb, S. (2011). *School Accountability*. *Handbook of the Economics of Education* (1st ed., Vol. 3). Elsevier B.V. https://doi.org/10.1016/B978-0-444-53429-3.00008-9

Findlay, L. (2002). Negotiating the swamp: the opportunity and challenge of reflexivity in research practice. *Qualitative Research*, *2*(2), 209–230.

Firestone, W. A., & González, R. A. (2007). Culture and Processes Affecting Data Use in School Districts. In P. A. Moss (Ed.), *Evidence and decision making* (Vol. 106, pp. 132–154). USA: Wiley-Blackwell. https://doi.org/10.1111/j.1744-7984.2007.00100.x

Firestone, W. A., Mayrowetz, D., & Fairman, J. (1998). Performance-Based

Assessment and Instructional Change: The Effects of Testing in Maine and Maryland. *Educational Evaluation and Policy Analysis*, *20*(2), 95–113.

Fitzgerald, T. (2007). Documents and documentary analysis. In A. R. J. Briggs & M. Coleman (Eds.), *Research methods in educational leadership and management*. Los Angeles: SAGE Publications.

Forte, E. (2010). Examining the Assumptions Underlying the NCLB Federal Accountability Policy on School Improvement. *Educational Psychologist*, *45*(2), 76–88. https://doi.org/10.1080/00461521003704738

Gallagher, L., Means, B., & Padilla, C. (2008). *Teachers' Use of Student Data Systems to Improve Instruction, 2005 to 2007*. *US Department of Education*.

Garcia, E., & Weiss, E. (2019). *The teacher shortage is real, large and growing, and worse than we thought.*

Gipps, C. (2012). *Beyond testing: towards a theory of educational assessment*. Abingdon [Royaume-Uni]: Routledge.

Gipps, C., & Stobart, G. (2009). Fairness in Assessment. In *Educational Assessment in the 21st Century: Connecting Theory and Practice* (pp. 105–118). https://doi.org/10.1007/978-1-4020-9964-9

Goldstein, H. (2015). Validity, science and educational measurement. *Assessment in Education: Principles, Policy and Practice*, *22*(2), 193–201. https://doi.org/10.1080/0969594X.2015.1015402

Goldstein, H., & Thomas, S. (2008). Reflections on the international comparative surveys debate. *Assessment in Education: Principles, Policy and Practice*, *15*(3), 215–222. https://doi.org/10.1080/09695940802417368

Good, T. L., Wiley, C. R. H., & Sabers, D. (2010). Accountability and Educational Reform: A Critical Analysis of Four Perspectives and Considerations for Enhancing Reform Efforts. *Educational Psychologist*, *45*(2), 138–148. https://doi.org/10.1080/00461521003720171

Gotch, C. M., & Roduta Roberts, M. (2018). A Review of Recent Research on Individual-Level Score Reports. *Educational Measurement: Issues and Practice*, *37*(3), 46–54. https://doi.org/10.1111/emip.12198

Gulek, C. (2003). Preparing for High-Stakes Testing. *Theory into Practice*, *42*(1), 42–50.

Gumbrecht, J. (2014). Major changes coming to 2016 SAT test. Retrieved September 9, 2019, from https://edition.cnn.com/2014/03/05/living/sat-test-changes-schools/index.html

Hannaway, J., & Hamilton, L. (2008). *Performance-based Accountability Policies : Implications for School and Classroom Practices*. https://doi.org/10.1037/e722482011-001

Hardman, J. (2017). Improving recruitment and retention of overseas teachers. In S. Blandford & M. Shaw (Eds.), *Managing international schools.* [Place of publication not identified]: Routledge.

Harris, D. M. (2012a). Postscript: Urban schools, accountability, and equity: Insights regarding NCLB and reform. *Education and Urban Society*, *44*(2), 203–210. https://doi.org/10.1177/0013124511431571

Harris, D. M. (2012b). Varying teacher expectations and standards: Curriculum differentiation in the age of standards-based reform. *Education and Urban Society*, *44*(2), 128–150. https://doi.org/10.1177/0013124511431568

Hattie, J. (2005). What is the nature of evidence that makes a difference to learning? In *Using Data to Support Learning* (pp. 11–21).

Hattie, J. (2009). Visibly Learning from Reports: The Validity of Score Reports. *Meeting of the National Council on Measurement in …*.

Hattie, J. (2010). *Visible learning: a synthesis of over 800 meta-analyses relating to achievement*.

Hattie, J. (2014). The Last of the 20th-Century Test Standards. *Educational Measurement: Issues and Practice*, *33*(4), 34–35. https://doi.org/10.1111/emip.12053

Hayden, M. (2006). *Introduction to international education: international schools and their communities*. London: Sage Publ.

Hayden, M., & Thompson, J. (2011). Teachers for the international school of the future. In R. Bates (Ed.), *Schooling internationally globalisation, internationalisation, and the future for international schools*. Routledge.

Hess, F. (2008). The new stupid. *Educational Leadership*, *66*(4), 12–17.

Hochberg, E. D., & Desimone, L. M. (2010). Professional Development in the Accountability Context: Building Capacity to Achieve Standards. *Educational Psychologist*, *45*(2), 89–106. https://doi.org/10.1080/00461521003703052

Holderness, J. (2002). The role of continuing professional development in the improvement of international schools. In M. Hayden, J. Thompson, & G. Walker (Eds.), *International education in practice: Dimensions for schools and international schools*. London: Routledge. https://doi.org/10.4324/9780203416983_chapter_7

Ingersoll, R., & Perda, D. (2009). The mathematics and science teacher shortage : Fact and myth, 45. https://doi.org/10.1037/e546742012-001

Ingvarson, L. (2005). Getting profressional development right. In Australian Council for Educational Research (ACER) (Ed.), *Using Data to Support Learning*.

International Reading Association, & National Council of Teachers of English. (2010). *Standards for the Assessment of Reading and Writing (Revised Edition)*. Newark, DE: International Reading Association/National Council of Teachers of English. https://doi.org/10.2307/358885

International Test Commission. (2019). ITC Guidelines for the Large-Scale Assessment of Linguistically and Culturally Diverse Populations. *International Journal of Testing*. https://doi.org/10.1080/15305058.2019.1631024

ISC Research. (2018). Pathways from K-12 English-medium International Schools to University.

Jennings, J., & Bearak, J. M. (2014). "Teaching to the Test" in the NCLB Era: How Test Predictability Affects Our Understanding of Student Performance. *Educational Researcher*, *43*(8), 381–389. https://doi.org/10.3102/0013189X14554449

Kane, M. (2001). Current Concerns in Validity Theory. *Journal of Educational Measurement*, *38*(4), 319–342.

Kane, M. (2006). Validity. In R. Brennan, E. National Council on Measurement in, & E. American Council on (Eds.), *Educational measurement*. Westport, CT: Praeger Publishers.

Kane, M. (2013a). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, *50*(1), 1–73.

Kane, M. (2013b). Validation as a Pragmatic, Scientific Activity. *Journal of Educational Measurement*, *50*(1), 115–122. https://doi.org/10.1177/1074248417724871

Kane, M. (2016a). Explicating validity. *Assessment in Education: Principles, Policy and Practice*, *23*(2), 198–211. https://doi.org/10.1080/0969594X.2015.1060192

Kane, M. (2016b). Validity as the evaluation of the claims based on test scores. *Assessment in Education: Principles, Policy and Practice*, *23*(2), 309–311. https://doi.org/10.1080/0969594X.2016.1156645

Kerr, K., Marsh, J., Schuyler Ikemoto, G., Darilek, H., & Barney, H. (2006). Strategies to Promote Data Use for Instructional Improvement: Actions, Outcomes, and Lessons from Three Urban Districts. *American Journal of Education*.

Keuning, T., Van Geel, M., & Visscher, A. (2017). Why a Data-Based Decision-Making Intervention Works in Some Schools and Not in Others. *Learning Disabilities Research and Practice*, *32*(1), 32–45. https://doi.org/10.1111/ldrp.12124

Kieffer, M. J., Lesaux, N. K., Rivera, M., & Francis, D. J. (2009). *Accommodations for english language learners taking large-scale assessments: A meta-analysis on effectiveness and validity*. *Review of Educational Research* (Vol. 79). https://doi.org/10.3102/0034654309332490

Kim, J. (2018). School accountability and standard-based education reform: The recall of social efficiency movement and scientific management. *International Journal of Educational Development*, *60*, 80–87. https://doi.org/10.1016/j.ijedudev.2017.11.003

King, N., & Horrocks, C. (2010). *Interviews in qualitative research*. London: SAGE.

Kivunja, C., & Kuyini, A. B. (2017). Understanding and Applying Research Paradigms in Educational Contexts. *International Journal of Higher Education*, *6*(5), 26. https://doi.org/10.5430/ijhe.v6n5p26

Koretz, D. (2005). Alignment, High Stakes, and the Inflation of Test Scores. *Yearbook of the National Society for the Study of Education*, *104*(2), 99–118. https://doi.org/10.1111/j.1744-7984.2005.00027.x

Koretz, D. (2009). *Measuring up: what educational testing really tells us*. Cambridge, Massachusetts: Harvard University Press.

KPASS. (2019). Member Schools. Retrieved June 18, 2020, from https://sites.google.com/site/kpassp2/home/about-kpassp

Krouwel, M., Jolly, K., & Greenfield, S. (2019). Comparing Skype (video calling) and in-person qualitative interview modes in a study of people with irritable

bowel syndrome-an exploratory comparative analysis. *BMC Medical Research Methodology*, *19*, 1–9. https://doi.org/10.1186/s12874-019-0867-9

Kvale, S. (2013). *Doing interviews*. Los Angeles: Sage.

Kvale, S., & Brinkmann, S. (2009). Interviews: learning the craft of qualitative research interviewing. Los Angeles: Sage Publications.

Lachat, M. A., & Smith, S. (2005). Practices that support data use in urban high schools. *Journal of Education for Students Placed At Risk (JESPAR)*, *10*(3), 333–349. https://doi.org/10.1207/s15327671espr1003_7

Lai, M. K., & Schildkamp, K. (2013). Data-based Decision Making: An Overview. In *Data-based Decision Making in Education* (pp. 9–21). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-4816-3_2

Lane, S., & Leventhal, B. (2015). Psychometric Challenges in Assessing English Language Learners and Students With Disabilities. *Review of Research in Education*, *39*(1), 165–214. https://doi.org/10.3102/0091732X14556073

Leighton, J. (2020). Cognitive diagnosis is not enough: The challenge of measuring learning within classroom assessments. In S. Brookhart & J. McMillan (Eds.), *Classroom assessment and educational measurement*.

Letukas, L. (2014). *Nine Facts About the SAT That Might Surprise You*.

Linn, R. (1989). *Educational measurement*. Phoenix, AZ: Oryx Press.

Linn, R. (2000). Assessments and Accountability. *Educational Researcher*, *29*(2), 4–16. https://doi.org/10.3102/0013189x029002004

Linn, R. (2005). Issues in the Design of Accountability Systems. *Yearbook of the National Society for the Study of Education*, *104*(2), 78–98. https://doi.org/10.1111/j.1744-7984.2005.00026.x

Lo Iacono, V., Symonds, P., & Brown, D. (2016). Skype as a tool for qualitative research interviews. *Sociological Research Online*, *21*(2). https://doi.org/10.5153/sro.3952

Maguire, M., & Delahunt, B. (2017). Doing a thematic analysis: A practical, step-by-step guide for learning and Ttaching scholars. *All Ireland Journal of Teaching and Learning in Higher Education*, *3*(335), 1–14. https://doi.org/10.1109/TIA.2014.2306979

Makar, K., & Confrey, J. (2002). Comparing two distributions: investigating teachers statistical thinking. In *ICOTS6* (pp. 1–4).

Mandinach, E. (2012). A Perfect Time for Data Use: Using Data-Driven Decision Making to Inform Practice. *Educational Psychologist*, *47*(2), 71–85. https://doi.org/10.1080/00461520.2012.667064

Mandinach, E., & Gummer, E. (2012). A systematic View of implementing Data Literacy in Educator Preparation. *Educational Researcher*, *42*(1), 30–37.

Mandinach, E., & Honey, M. (2008). *Data-driven school improvement: linking data and learning*. *Technology, education-connections, the TEC series*.

Mandinach, E., Honey, M., & Light, D. (2006). A theoretical framework for data-driven decision making. In *AERA, San Francisco* (pp. 1–18).

Mandinach, E., & Jackson, S. (2012). *Transforming teaching and learning through data-driven decision making*.

Mandinach, E., & Jimerson, J. B. (2016). Teachers learning how to use data: A synthesis of the issues and what is known. *Teaching and Teacher Education*, *60*, 452–457. https://doi.org/10.1016/j.tate.2016.07.009

Mandinach, E., Rivas, L., Light, D., Heinze, C., & Honey, M. (2006). The impact of data-driven decision making tools on educational practice: a systems analysis of six school districts (pp. 1–23).

Mandinach, E., & Schildkamp, K. (2020). Misconceptions about data-based decision making in education: An exploration of the literature. *Studies in Educational Evaluation*, (September 2019), 1–10. https://doi.org/10.1016/j.stueduc.2020.100842

Masters, G. (2001). *Educational Measurement*. Victoria: ACER.

Matters, G. (2006). Using data to support learning in schools: Students, teachers, systems. In *Australian Education Review*.

Matthews, J., Trimble, S., & Gay, A. (2007). But What Do You Do with the Data?. *Principal Leadership*, (May), 31–33.

McClelland, R. (2017). Managing assessment in the international school. In S. Blandford & M. Shaw (Eds.), *Managing international schools.* (pp. 48–62). London and New York: ROUTLEDGE.

McCulloch, G. (2004). *Documentary research in education, history, and the social sciences*. London; New York: RoutledgeFalmer.

Means, B., Chen, E., DeBarger, A., & Padilla, C. (2011). *Teachers' Ability to Use Data to Inform Instruction: Challenges and Supports*. *Washington, DC*.

Means, B., Gallagher, L., & Padilla, C. (2007). *Teachers' use of student data systems to improve instruction*. *US Department of Education*.

Means, B., Padilla, C., DeBarger, A., & Bakia, M. (2009). *Implementing Data-Informed Decision Making in Schools - Teacher Access, Supports and Use. US Department of Education*.

Means, B., Padilla, C., & Gallagher, L. (2010). *Use of Education Data at the Local Level: From Accountability to Instructional Improvement. US Department of Education*.

Merriam, S. B. (1998). *Qualitative research and case study applications in education*. San Francisco, Calif.: Jossey-Bass.

Merriam, S. B., & Tisdell, E. J. (2016). Qualitative research: a guide to design and implementation.

Mertler, C. A. (2007). *Interpreting standardized test scores: strategies for data-driven instructional decision making*. Los Angeles: SAGE.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (Third, pp. 13–103). Phoenix, AZ: Oryx Press.

Moore, R., & Shaw, T. (2017). *Teachers' Use of Data: An Executive Summary*.

Morris, A. (2011). *Student Standardised Testing Current Practices in OECD*

Countries and a Literature Review. *OECD Education Working Papers*. https://doi.org/10.1787/5kg3rp9qbnr6-en

Murphy, B. (2017). Performance-Based Assessment for English Language Learners. *Cornell University ILR School DigitalCommons*, 1–5.

Murphy, E. (1998). International accreditation: who needs it? In M. Hayden & J. Thompson (Eds.), *International education: principles and practice*. London: Taylor & Francis.

Murray, J. (2013). Critical issues facing school leaders concerning data-informed decision-making. *School Leadership & Management*, *33*(2), 169–177. https://doi.org/10.1080/13632434.2013.773882

Nardi, E. (2008). Cultural biases: a non-Anglophone perspective. *Assessment in Education: Principles, Policy & Practice*, *15*(3), 259–266. https://doi.org/10.1080/09695940802417467

National Research Council. (1999). *Testing, Teaching, and Learning*. *Testing, Teaching, and Learning: A Guide for States and School Districts*. Washington DC: The National Academies Press. https://doi.org/10.17226/9609

Nehls, K., Smith, B. D., & Schneider, H. A. (2014). Video-conferencing interviews in qualitative research. In S. Hai-Jew (Ed.), *Enhancing Qualitative and Mixed Methods Research with Technology* (pp. 140–157). Hershey, PA: IGI Global. https://doi.org/10.4018/978-1-4666-6493-7.ch006

Newton, P., Shaw, S., Lagrange, M., & Robinson, N. (2014). Validity in educational and psychological assessment.

Nichols, P., & Williams, N. (2009). Consequences of test score use as validity evidence: Roles and responsibilities. *Educational Measurement: Issues and Practice*, *28*(1), 3–9. https://doi.org/10.1111/j.1745-3992.2009.01132.x

Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic Analysis: Striving to Meet the Trustworthiness Criteria. *International Journal of Qualitative Methods*, *16*(1), 1–13. https://doi.org/10.1177/1609406917733847

Noyes, A. (2007). Mathematical marginalisation and meritocracy: Inequity in an English classroom. *The Montana Mathematics Enthusiast*, *1*(0), 35–48.

NWEA. (2011). *Technical manual for Measures of Academic Progress (MAP) and Measures of Academic Progress for Primary Grades (MPG)*. Portland, OR.

Oakland, T. (2004). Use of Educational and Psychological Tests Internationally. *Applied Psychology*, *53*(2), 157–172. https://doi.org/10.1111/j.1464-0597.2004.00166.x

Odland, G., & Ruzicka, M. (2009). An investigation into teacher turnover in international schools. *Journal of Research in International Education*, *8*(1), 5–29. https://doi.org/10.1177/1475240908100679

Oliveri, M. E., & Lawless, R. (2018). The Validity of Inferences From Locally Developed Assessments Administered Globally. *ETS Research Report Series*, *2018*(1). https://doi.org/10.1002/ets2.12221

Oliveri, M. E., Lawless, R., & Mislevy, R. J. (2019). Using Evidence-Centered Design to Support the Development of Culturally and Linguistically Sensitive

Collaborative Problem-Solving Assessments. *International Journal of Testing*, *19*(3), 270–300. https://doi.org/10.1080/15305058.2018.1543308

Oliveri, M. E., Lawless, R., & Young, J. (2015). A validity framework for the use and development of exported assessments. *ETS Report*.

Oliveri, M. E., & von Davier, A. (2016). Psychometrics in Support of a Valid Assessment of Linguistic Minorities: Implications for the Test and Sampling Designs. *International Journal of Testing*, *16*(3), 220–239. https://doi.org/10.1080/15305058.2015.1069743

Park, J., Kitayama, S., Markus, H. R., Coe, C. L., Miyamoto, Y., Karasawa, M., … Ryff, C. D. (2013). Social status and anger expression: The cultural moderation hypothesis. *Emotion*, *13*(6), 1122–1131. https://doi.org/10.1037/a0034273

Phelps, R. (2000). Trends in large-scale testing outside the United States. *Educational Measurement: Issues and Practice*, *19*(1), 11–21. https://doi.org/10.1111/j.1745-3992.2000.tb00018.x

Phelps, R. (2008). *Standardized testing primer*. New York: Peter Lang.

Phelps, R. (2015). Educational Achievement Testing: Critiques and Rebuttals. In *Correcting Fallacies about Educational and Psychological Testing.* Washington: American Psychological Association.

Pierce, R., & Chick, H. (2010). Interpreting literacy and numeracy testing reports: What do teachers need to know? In *ICOTS8* (Vol. 8, pp. 8–11).

Pierce, R., & Chick, H. (2011a). Reacting to quantitative data: Teachers' perceptions of student achievement reports. *23rd Biennial Conference of The Australian …*, 631–639.

Pierce, R., & Chick, H. (2011b). Teachers' intentions to use national literacy and numeracy assessment data: a pilot study. *Australian Educational Researcher*, *38*(4), 433–447. https://doi.org/10.1007/s13384-011-0040-x

Pierce, R., & Chick, H. (2012a). Improving Teachers' Statistical Literacy. In *Ozcots2012* (pp. 1–6).

Pierce, R., & Chick, H. (2012b). Workplace statistical literacy for teachers: interpreting box plots. *Mathematics Education Research Journal*, *25*(2), 189–205. https://doi.org/10.1007/s13394-012-0046-3

Pierce, R., & Chick, H. (2013). Statistical literacy: a professional skill for today's teachers. *Curriculum & Leadership Journal*, *11*(14).

Pierce, R., Chick, H., & Wander, R. (2012). Improving Teachers' Professional Statistical Literacy Evidence Base For Teachers' Statistical Literacy Workshop. In *OZCOTS2012* (pp. 1–16).

Pitts, R. T., & Naumenko, O. (2016). The 2014 Standards for Educational and Psychological Testing: What Teachers Initially Need to Know. *Working Papers in Education*, *2*(1), 1–6.

Plake, B., & Wise, L. (2014). What Is the Role and Importance of the Revised AERA, APA, NCME Standards for Educational and Psychological Testing? *Educational Measurement: Issues and Practice*, *33*(4), 4–12. https://doi.org/10.1111/emip.12045

Polikoff, M. (2012). The Association of State Policy Attributes With Teachers' Instructional Alignment. *Educational Evaluation and Policy Analysis*, *34*(3), 278–294. https://doi.org/10.3102/0162373711431302

Polikoff, M., Porter, A., & Smithson, J. (2011). How well aligned are state assessments of student achievement with state content standards? *American Educational Research Journal*, *48*(4), 965–995. https://doi.org/10.3102/0002831211410684

Poortman, C., & Schildkamp, K. (2016). Solving student achievement problems with a data use intervention for teachers. *Teaching and Teacher Education*, *60*, 425–433. https://doi.org/10.1016/j.tate.2016.06.010

Poortman, C., Schildkamp, K., & Lai, M. (2016). Professional development in data use: An international perspective on conditions, models and effects. *Teaching and Teacher Education*, *60*, 363–365. https://doi.org/10.1016/j.tate.2016.07.029

Popham, W. J. (1999). Why standardized tests don't measure educational quality. *Educational Leadership*, *56*(6), 8–15.

Popham, W. J. (2009). Assessment literacy for teachers: Faddish or fundamental? *Theory into Practice*, *48*(1), 4–11. https://doi.org/10.1080/00405840802577536

Popham, W. J. (2011). Assessment literacy overlooked: A teacher educator's confession. *Teacher Educator*, *46*(4), 265–273. https://doi.org/10.1080/08878730.2011.605048

Popham, W. J. (2018). *Assessment literacy for educators in a hurry*. Alexandria, VA: ASCD.

Reynolds, C. R., Livingston, R. B., & Willson, V. L. (2010). *Measurement and assessment in education*. Upper Saddle River, NJ: Pearson.

Roduta Roberts, M., & Gotch, C. M. (2019). Development and Examination of a Tool to Assess Score Report Quality. *Frontiers in Education*, *4*(March), 1–10. https://doi.org/10.3389/feduc.2019.00020

Roduta Roberts, M., Gotch, C. M., & Lester, J. N. (2018). Examining Score Report Language in Accountability Testing. *Frontiers in Education*, *3*(June), 1–17. https://doi.org/10.3389/feduc.2018.00042

Roulston, K. (2014). *Reflective interviewing: a guide to theory and practice*. Los Angeles: SAGE.

Rubin, H. J., & Rubin, I. S. (2012). *Qualitative interviewing: the art of hearing data*. Thousand Oaks, Calif.: SAGE.

Schildkamp, K., & Ehren, M. (2013). From "Intuition"- to "Data"-based Decision Making in Dutch Secondary Schools? In *Data-based Decision Making in Education* (pp. 49–67). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-4816-3_4

Schildkamp, K., Karbautzki, L., & Vanhoof, J. (2014). Exploring data use practices around Europe: Identifying enablers and barriers. *Studies in Educational Evaluation*, *42*, 15–24. https://doi.org/10.1016/j.stueduc.2013.10.007

Schildkamp, K., Lai, M. K., & Earl, L. (2013). *Data-based decision making in education: challenges and opportunities*. Springer.

Schleicher, A. (2015). Are American students overtested? Listen to what students themselves say - OECD Education and Skills Today. Retrieved July 23, 2019, from https://oecdedutoday.com/are-american-students-overtested-listen-to-what-students-themselves-say/

Schraw, G. (2010). No School Left Behind. *Educational Psychologist*, *45*(2), 71–75. https://doi.org/10.1080/00461521003720189

Schwabe, F., von Davier, A., & Chalhoub-Deville, M. (2016). Language and culture in testing. In F. T. L. Leong, D. Bartram, F. Cheung, K. Geisinger, & D. Iliescu (Eds.), *The ITC international handbook of testing and assessment*.

Scott, J. (2014). A Matter of Record Documentary Sources in Social Research. New York, NY: John Wiley & Sons.

Sears, C. (2015). *Second language students in English-medium classrooms : a guide for teachers in international schools*. Bristol: Multilingual Matters.

Shepard, L. A. (2016). Evaluating test validity: reprise and progress. *Assessment in Education: Principles, Policy and Practice*, *23*(2), 268–280. https://doi.org/10.1080/0969594X.2016.1141168

Skelton, M. (2002). Defining 'international' in an international curriculum. In M. Hayden, J. Thompson, & G. Walker (Eds.), *International education in practice: Dimensions for schools and international schools*. https://doi.org/10.4324/9780203416983_chapter_4

Smith, M. (2005). Getting SMART with data in schools: Lessons in NSW. In *Using data to support learning (Conference Proceedings).* (pp. 38–45).

Smith, W. (2014). The Global Transformation Toward Testing for Accountability Education. *Education Policy Analysis Archives*, *22*(116), 1–34.

Smith, W., & Benavot, A. (2019). Improving accountability in education: the importance of structured democratic voice. *Asia Pacific Education Review*, *20*(2), 193–205. https://doi.org/10.1007/s12564-019-09599-9

Spillane, J. (2012). The more things change, the more things stay the same? *Education and Urban Society*, *44*(2), 123–127. https://doi.org/10.1177/0013124511431567

Staehr Fenner, D. (2016). Fair and square assessments for ELL. *Educational Leadership*, *73*(5).

Stickler, L., & Breland, N. (2007). A Critical Review of the Sat : Menace or Mild-Mannered Measure ? *TCNJ Journal of Student Scholarship*, *IX*, 1–8.

Stiggins, R. (1991). Assssment Literacy. *The Phi Delta Kappan*, *72*(7), 534–539.

Stoelting, L. (2019). How I Went from MAP Novice to MAP Coordinator at My School. Retrieved August 9, 2019, from https://www.nwea.org/blog/2019/how-i-went-from-map-novice-to-map-coordinator-at-my-school/

Sullivan, J. R. (2012). Skype: An Appropriate Method of Data Collection for Qualitative Interviews? *The Hilltop Review*, *6*(1), 54–60.

Sun, J., Przybylski, R., & Johnson, B. J. (2016). A review of research on teachers' use of student data: from the perspective of school leadership. *Educational Assessment, Evaluation and Accountability*, *28*(1), 5–33.

https://doi.org/10.1007/s11092-016-9238-9

Supovitz, J. (2012). Getting at Student Understanding — The Key to Teachers' Use of Test Data. *Teachers College Record*, *114*(11), 1–29.

Sutcher, L., Darling-Hammond, L., & Carver-Thomas, D. (2016). *A Coming Crisis in Teaching? Teacher supply, demand, and shortages in the U.S.*

Te Kete Ipurangi. (n.d.). Assessment online: Reliability and validity. Retrieved February 1, 2020, from https://assessment.tki.org.nz/Using-evidence-for-learning/Working-with-data/Concepts/Reliability-and-validity

The Graide Network. (2019). *Importance of validity and reliability in classroom assessments*.

Thomas, S., & Goldstein, H. (2008). International comparative studies of achievement – re-examining the issues and impacts. *Assessment in Education: Principles, Policy & Practice*, *15*(3), 211–213. https://doi.org/10.1080/09695940802417277

Thum, Y. M., & Hauser, C. (2015). NWEA 2015 MAP Norms for student and school achievement status and growth, 431.

Thum, Y. M., & Matta, T. H. (2015). MAP College Readiness Benchmarks: A research brief, 1–25.

UNESCO. (2017). *Accountability in education: Meeting our commitments*.

University of Iowa. (n.d.-a). *Content validity for large-scale assessment*.

University of Iowa. (n.d.-b). *Interpreting standard reports from the Iowa Assessments*.

Valdes, G., & Figueroa, R. (1996). *Bilingualism and testing: a special case of bias*. Norwood, N.J: Ablex.

Valentine, J. (1961). The College Entrance Examination Board. *National Council of Teachers of English*, *12*(2), 88–92.

Valli, L., & Buese, D. (2007). *The Changing Roles of Teachers in an Era of High-Stakes Accountability*. *American Educational Research Journal* (Vol. 44). https://doi.org/10.3102/0002831207306859

Vanlommel, K., & Schildkamp, K. (n.d.). How Do Teachers Make Sense of Data in the Context of High-Stakes Decision Making? *American Educational Research Journal*, *56*(3), 792–821. https://doi.org/10.3102/0002831218803891

Vanlommel, K., & Schildkamp, K. (2019). How Do Teachers Make Sense of Data in the Context of High-Stakes Decision Making? *American Educational Research Journal*, *56*(3), 792–821. https://doi.org/10.3102/0002831218803891

Vanlommel, K., Van Gasse, R., Vanhoof, J., & Van Petegem, P. (2017). Teachers' decision-making: Data based or intuition driven? *International Journal of Educational Research*, *83*(March 1994), 75–83. https://doi.org/10.1016/j.ijer.2017.02.013

Vanlommel, K., Vanhoof, J., & Van Petegem, P. (2016). Data use by teachers: the impact of motivation, decision-making style, supportive relationships and reflective capacity. *Educational Studies*, *42*(1), 36–53. https://doi.org/10.1080/03055698.2016.1148582

Vogler, K. (2003). An Integrated Curriculum Using State Standards in a High-Stakes Testing Environment. *Middle School Journal*, *34*(4), 5–10. https://doi.org/10.1080/00940771.2003.11495383

Walker, J. (2017). Norm-referenced standardised testing for institutional evaluation in the international school context. In M. Hayden & J. Thompson (Eds.), *Perspectives on assessment and evaluation in international education*. Woodbridge: John Catt Educational Ltd.

Wasson, D. (2009). Large cohort testing: How can we use assessment data to effect school and system improvement. In *Assessment and student learning: Collecting, interpreting and using data to support learning* (pp. 47–56). ACER.

Wayman, J., & Stringfield, S. (2006a). Data Use for School Improvement: School Practices and Research Perspectives. *American Journal of Education*, *112*(4), 463–468. https://doi.org/10.1086/505055

Wayman, J., & Stringfield, S. (2006b). Technology-supported involvement of entire faculties in examination of student data for instructional improvement. *American Journal of Education*, *112*(4).

Welch, C., Dunbar, S., & Fina, A. (2018). *Technical Summary for Form F of the Iowa Assessments*.

Welch, C., Dunbar, S., & Rickels, H. (2014). Measuring growth with the Iowa Assessments.

Wellington, J. J. (2016). *Educational research: contemporary issues and practical approaches*. London: Bloomsbury Academic.

Wendler, C., & Powers, D. (2009). What Does It Mean to Repurpose a Test? *R&D Connections*, *9*.

Western Association of Schools and Colleges. (2017). *Focus on Learning International Edition*.

Wiliam, D. (2010). Standardized Testing and School Accountability. *Educational Psychologist*, *45*(2), 107–122. https://doi.org/10.1080/00461521003703060

Wiliam, D. (2014a). Teacher expertise: why it matters and how to get more of it. In J. Hallgarten, L. Bamfield, & K. McCarthy (Eds.), *Licensed to create: Ten essays on improving teacher quality* (pp. 27–36). London: Rsa Action and Research Centre.

Wiliam, D. (2014b). What Do Teachers Need to Know About the New Standards for Educational and Psychological Testing? *Educational Measurement: Issues and Practice*, *33*(4), 29–30. https://doi.org/10.1111/emip.12051

Williamson, P., Bondy, E., Langley, L., & Mayne, D. (2005). Meeting the Challenge of High-Stakes Testing While Remaining Child-Centered: The Representations of Two Urban Teachers. *Childhood Education*, *81*(4), 190–195. https://doi.org/10.1080/00094056.2005.10522271

Willis, J., Dumont, R., & Kaufman, A. (2012). Individual norm-referenced standardised assessment. In B. J. Irby, G. Brown, R. Lara-Alecio, & S. Jackson (Eds.), *The handbook of educational theories*. Charlotte, N.C.: Information Age Pub.

Young, J. (2009). A framework for test validity research on content assessments

taken by English language learners. *Educational Assessment*, *14*(3–4), 122–138. https://doi.org/10.1080/10627190903422856

Zenisky, A., & Hambleton, R. (2012). Developing Test Score Reports That Work: The Process and Best Practices for Effective Communication. *Educational Measurement: Issues and Practice*, *31*(2), 21–26. https://doi.org/10.1111/j.1745-3992.2012.00231.x