Issues in Large-scale Assessments

Keynote address presented at PROMS 2009, July 28-30, 2009, Hong Kong.

Margaret Wu

University of Melbourne

Email address: m.wu@unimelb.edu.au

Phone: +61 3 8344 0963

Fax: +61 3 9348 2753

Postal address:

Assessment Research Centre,

234 Queensberry Street,

Faculty of Education,

University of Melbourne,

Victoria,

Australia, 3010

Keywords: Measurement error, sampling error, equating error, large-scale assessments

Executive Summary

Over the past ten years, large-scale assessments have gained more popularity and publicity than ever before. These include international surveys as well as national surveys such as Programme for International Student Assessment (PISA) (http://www.pisa.oecd.org) conducted by the Organisation for Ecomonic Co-operation and Development (OECD), Trends in Mathematics and Science Study and Progress in International Reading Literacy Study (http://timss.bc.edu/index.html) conducted by the International Association for the Evaluation of Educational Achievement (IEA). At the national level, in the United States, the No Child Left Behind (NCLB) Act 2001 stipulated the requirement of annual assessments at the state level. In Australia, NAPLAN (National Assessment Program – Literacy and Numeracy) began in 2008. In addition, a number of sample-based Australian national assessment programs are also being implemented for Science, Civics and Citizenship, and Information and Communication Technology Literacy (ICT).

Clearly, each assessment design needs to be built to match the objectives of the assessment. However, valid and reliable assessments are not easily developed. The validity and reliability of a large-scale assessment is under threat from multiple sources including measurement error, sampling error, measurement disturbances and administrative challenges. This paper critically evaluates the extent to which large-scale assessment programs meet their objectives. There are two main reasons that such an evaluation is called for. First, it is becoming evident that some assumptions of large-scale survey methodologies are violated, leading to invalid results (Mazzeo & von Davier, 2009; Monseur & Berezner, 2007), or, at best, caveats attached to the results (Wu, 2009a). Second, judging from reports in the media, there is a large section of the public who misquote or misuse large-scale assessment results. In recent months, the Australian media reported that the government has plans to release school level national assessment results to the public. Such results can be easily misinterpreted by the layperson. Most of all, such assessments have a high level of inaccuracies at the individual student level, and the degree of confidence with which the results are released is often brushed aside and ignored.

The key findings in this paper are:

- At individual student level, the measurement error associated with a 40-item test is more than half a year's growth in school. With annual testing using a 40-item test, it is expected that 16% of the students would appear to regress "backward" when they have actually made one year's progress.
- At the class level, with a 40-item test, the 95% confidence interval of the class average is over half a year's growth. This uncertainty is quite large. Any suggestion that teacher performance should be determined by such measurement of student growth is of serious concern.
- For sample-based surveys, the sampling error is large if two-stage sampling is conducted: schools are selected and then students within schools are selected. If the effective sample size is fewer than 400, there is little chance that trends can be measured. Typical differences between the mean scores of Australian states are not likely to be detected.
- Equating error is by far the largest source of error. Test items often work differently for different groups of students (e.g., across countries, states and gender groups). Item

positions in a test booklet also have a significant impact on item difficulty. Equating error can be as large as half a year's growth, throwing trend estimates completely off the track.

- There are many challenges in building a vertical scale: a common scale for several grade levels. Threats to equating tests across grades include different curricula across grades and item position effect in the test booklets. The validity of vertical scales is often questionable.

The following table summarises characteristics and issues for different types of assessment.

Type of assessment	Example	Assessment Characteristics and Issues
Test consisting of one instrument (e.g., 40 items) where every student takes the same instrument.	• NAPLAN	• Large measurement error at individual student level. Large error even at class level. No high-stakes decision should be made based on these results (e.g., teacher performance should not be judged by student scores in these tests). • No power to measure growth between two time points for individual students or at class level. • Equating is difficult because of fixed item positions and differential item functioning. • Test validity is not well-established due to the lack of coverage of content areas.
Test consisting of a number of rotated test forms, where each student only takes a small number of items.	• PISA • TIMSS	• Typically, results at individual student level are not of interest. • Item position effect can be moderated through a balanced test booklet design. • For international surveys, item by country interaction is very large. This threatens the validity of the results. If the number of common items for equating is small, equating error could be very large.
Sample-based assessment.	• NAP-Civics	• If two-stage sampling is used (selecting schools and then selecting students within schools), sampling efficiency could be reduced by a large factor. The result is that the sample may not be sufficiently large to provide any useful information.
Assessment across a number of grade levels.	• NAPLAN	• Vertical equating is difficult due to item position effect, curriculum differences across grades, and the lack of coverage of content areas.

Solutions to the problems?

To overcome the problems with measurement error, we can administer tests on multiple occasions. However, increasing the number of test administrations and student sample size will increase the cost and resources. So we must find cost-effective ways of testing students. The use of technology is part of the solution. Computer-delivered tests can reduce the cost considerably, as objective items can be marked automatically. Further, instant feedback can be provided to students and teachers. The current paper-and-pen format such as in NAPLAN has many drawbacks, in addition to the unreliability of the test results. To improve assessments, it is essential to utilise technology to build electronic item banks, and move towards computer-adaptive testing of students. Under such a system, students can be tested on multiple occasions throughout the year, the tests can be calibrated centrally, feedback can be provided immediately, and, above all, teachers can be relieved of many hours of marking. Computer-adaptive tests can reduce measurement error. Once such a system is built, a large number of students can be included in the assessment without too much additional cost so that sampling error will be reduced. Student progress can be monitored regularly under this system because there is longitudinal data at multiple time points throughout the year. Finally, with a wide coverage of content areas and moving away from fixed item positions in a test, equating error will also be reduced.

In terms of assessment design, one recommendation is to avoid constructing an assessment system that can “solve it all”. Focused studies may well be more suited to establish the effectiveness of a particular intervention, or a particular policy change. Smaller, purposeful and targeted assessment programs may achieve a narrow but well-defined set of objectives rather than a large-scale assessment system that does not provide any useful data.

To summarise, the mere implementation of an assessment program and production of a report will not serve the community any good unless valid results are produced. There are still many assessment programs that do not meet their objectives. A common scenario is that a government sets aside a budget for conducting an assessment program, and a contractor is appointed to conduct the program. The contractor designs a program based on the budget, and not on sound statistical basis. The program is completed and a report is produced. The usefulness of the results is not critically examined, and the validity of the results is not critically established. Nevertheless, accountability seems to have been achieved, with the money spent. This paper provides some rationale to call on those involved in assessments to stop implementing large-scale assessment programs for the sake of implementing them. Rather, the benefits of implementing an assessment program need to be thoroughly reviewed, before so much public money is invested.

Abstract

In large-scale assessments such as state-wide testing programs, national sample-based surveys and international comparative studies, there are many steps involved in the measurement and reporting of student achievement. Typically, the steps involved include the following:

- Items are developed and test instruments are assembled;
- Student samples are selected (for sample-based surveys);
- Tests are administered;
- Students' responses are marked (if there are open-ended items);
- Test results are scaled and equated (if trend analysis is of interest);
- Indicators are estimated (at individual and/or group levels);
- Reports are produced and conclusions are drawn.

The steps outlined above require expertise from different fields, including subject specialists, sampling statisticians, psychometricians and administrators. The steps are complex, and there is on-going research in all areas. Nevertheless, there are always sources of inaccuracies in each of the steps. It is of interest to identify the factors in the steps that may threaten the validity of the final results. Furthermore, an assessment of the relative magnitude of the threats will place the threats in order of importance. Survey designers can then improve the survey quality by focusing on areas that pose the highest threats to the validity of results.

This paper examines a number of factors that can lead to invalid results. In particular, sources of systematic errors pose the greatest threats to surveys. For example, a sample that is not sufficiently representative of the population will give biased results no matter how accurate the measurement process is. On the other hand, measurement disturbances such as differential item functioning may lead to serious bias in the results no matter how representative the sample is (e.g., even if the population is surveyed).

This paper discusses the magnitude of various sources of errors with reference to the objectives of assessment programs. A number of examples from large-scale surveys are used to illustrate threats to the surveys and the impact of the threats. The paper concludes by making a

number of recommendations that could lead to an improvement of the validity of large-scale survey results.

1. Introduction

Over the past ten years, large-scale assessments have been on the increase, and gaining more popularity, and publicity, than ever before. These include international surveys as well as national surveys. For example, the OECD's Programme for International Student Assessment (PISA) (http://www.pisa.oecd.org) began with 32 countries in the 2000 study. In 2003, 41 countries participated in PISA. In 2006, 57 countries participated in PISA. The International Association for the Evaluation of Educational Achievement (IEA) also conducts international studies in Mathematics, Science (TIMSS) and Reading literacy (PIRLS) (http://timss.bc.edu/index.html). Other cross-national studies include SACMEQ (The Southern and Eastern Africa Consortium for Monitoring Educational Quality) and PASEC (Programme d'Analyse des Systèmes Educatifs de la CONFEMEN). At the national level, in the United States, the No Child Left Behind (NCLB) Act 2001 stipulated the requirement of annual assessments at the state level. In Australia, NAPLAN (National Assessment Program – Literacy and Numeracy) began in 2008. In addition, a number of sample-based Australian national assessment programs are also being implemented for Science, Civics and Citizenship, and Information and Communication Technology Literacy (ICT).

No doubt, each large-scale assessment has its own objectives. The objectives generally range from profiling students' levels of achievement, to informing policy directions. In some cases, feedback to teaching and learning is also an objective of the assessment. Of course, there is no 'one-size-fits-all' assessment. Clearly, each assessment design needs to be developed to match the objectives of the assessment. Valid and reliable assessments are not easily developed. In a large-scale assessment, the threat to reliability and validity comes from multiple sources

including measurement error, sampling error, measurement disturbances[1] and administrative challenges. Cross-sectional studies are further limited in the type of research questions that can be answered, in comparison to longitudinal studies. This paper attempts to critically evaluate the extent to which large-scale assessment programs meet their objectives. There are two main reasons that such an evaluation is called for. First, it is becoming evident that some assumptions of large-scale survey methodologies are violated, leading to invalid results (Mazzeo & von Davier, 2009; Monseur & Berezner, 2007), or, at best, caveats attached to the results (Wu, 2009a). Second, judging from reports in the media, there is a large section of the public, including politicians, who misquote or misuse large-scale assessment results. In recent months, the Australian media reported that the government has plans to release school level national assessment results to the public. Such results can be easily misinterpreted by the layperson, as the assessment process is complex and the interpretations of the results need to be made with extreme caution. Most of all, such assessments have a high level of inaccuracies at the individual student level, and the degree of confidence (or rather the lack of confidence) with which the results are released is often brushed aside and is not well understood by the layperson. This paper attempts to delineate, in perspective, the degree of accuracies of various assessment results, so that the results can be used in informed ways.

2. Sources of errors associated with assessments

First, it is essential to recognise that assessment of student achievement has errors associated with it. Assessments, particularly in the large-scale case, are typically carried out by administering tests to a group of students. As there is only limited information obtained from each student (e.g., a multiple-choice test consisting of 40 items is administered), measurement

[1] I use the term “measurement disturbances” to indicate violations to the assumptions of the measurement model used in analysing the survey data.

error at the student level is large. If we are interested in reporting individual student results, then measurement error needs to be taken into account. Section 4 discusses the magnitude of measurement error in relation to test length.

If an assessment is concerned with estimating the average proficiency at a cohort level where sampling of students takes place, then sampling error is typically a main source of inaccuracy. Sampling error is discussed in Section 5.

Further, when trend estimates are reported, the equating process between administrations of tests has many sources of error. In some cases, the equating error far outweighs sampling and measurement error at the cohort level (Monseur & Berezner, 2007). Equating errors are discussed in Section 6.

There are other sources of error such as test administration procedures, marker reliability, item and test bias against particular groups of students, and the appropriateness of the statistical analyses carried out. The errors compound, so that the use of the results must take into account the confidence level with which the results are produced.

The recognition that assessment results contain error is the first step to help the layperson in interpreting assessment results.

3. Setting acceptable criteria for the accuracies of assessment results

Before discussing various sources of error, we should think about how to make judgments on the appropriateness of accuracies. First, in evaluating whether large-scale assessment programs meet their objectives, we need to determine an acceptable degree of accuracy of the assessment results *with respect to the objectives of the assessment*. Typically, it does not make sense to set a level of accuracy that is independent of the objectives. The following is an example. If I enter a weight loss program and expect to lose half a kilogram after

one week, but the scale I use to measure my weight is only accurate up to one kilogram, then there is little chance for me to have the confidence to find out whether the weight loss program is effective for me. In this case, I would want to use a scale that can measure more accurately. However, if I am interested whether the weight loss program is useful, on average, for a group of people, given that there are variations across people in terms of their weight loss, I could possibly establish an average loss for a large group of people using a somewhat less accurate scale for individuals. For example, using the scale that is accurate to 1 kg, I may find that, out of 1000 people, 500 people lost 1 kg after one week, and 500 people lost 0 kg. So, on average, the weight loss is 0.5 kg. That is, while an individual person's weights are not measured so well, we have a large group of people to establish an average. The more participants there are, the more accurate the average can be established. On the other hand, if I am interested in measuring the average difference in weight loss between males and females, and I expect the difference to be less than 300 g, then I would need a more accurate scale and/or many more participants. Otherwise, my survey would just be a waste of time as I would not expect to detect differences between males and females, not because there is not any difference, but because my measuring process is not accurate enough to detect that difference of interest.

Similarly, in measuring student achievement, the degree of accuracy required depends on the purpose of the survey. If we want to make high-stake decisions about individual students, then we need to reduce measurement error so that individuals can be measured well. If we are only interested in estimating the average achievement in a cohort, then we need to ensure that sampling error is small (if the survey is sample-based). If we want to know about differences between males and females, we typically need a larger sample to detect a small difference. For

example, if the expected difference between males and females is less than 0.1 in effective size[2], and our measures are accurate to 0.1 of a standard deviation (as is often the case in large-scale assessments, e.g., Mullis, et al (2003)), then we would not be able to detect any difference that is statistically significant between males and females simply because our measures are not accurate enough to detect that magnitude of difference. That is, the conclusion that the average achievement of males is not statistically significantly different from the average achievement of females can be largely expected without carrying out the survey. In this sense, the survey would be a waste of time and money. Of course, there may be an accuracy beyond which we are just not interested. For example, we may decide that we are only concerned if the difference between males and females is greater than an effect size of 0.1. In that case, we will decide on a sample size that can detect a difference of 0.1 effect size.

If our goal is to estimate trends of achievement levels across time, then our measures should be accurate enough to detect the expected change (which is typically small) across time. If our measures are not accurate enough to detect typical changes across time, then the survey is not appropriate for measuring trend. It will simply be a waste of time and money to conduct the survey.

The accuracy of our measures depends on the size of measurement error, sampling error, equating error, and other measurement disturbances. To set criteria to evaluate whether a survey is appropriate for a purpose, we will examine the size of the error of an estimate to see if the estimate is sufficiently accurate to answer the questions set out in the objectives of the survey. But before doing so, let us look at some expected growth rates in relation to the spread (standard

[2] In this paper, effect size is defined as a difference in scores divided by a standard deviation.

deviation) of student achievement from past surveys, so we can at least have a ballpark figure for what accuracies we are aiming for.

Figure 1 shows student achievement distributions for Grades 3, 5, 7 and 9 in reading for the Australian NAPLAN 2008 assessment (MCCETYA, 2008). The box for each grade shows the range mean±2×standard deviation (i.e., 95% of the student scores are expected within this range).

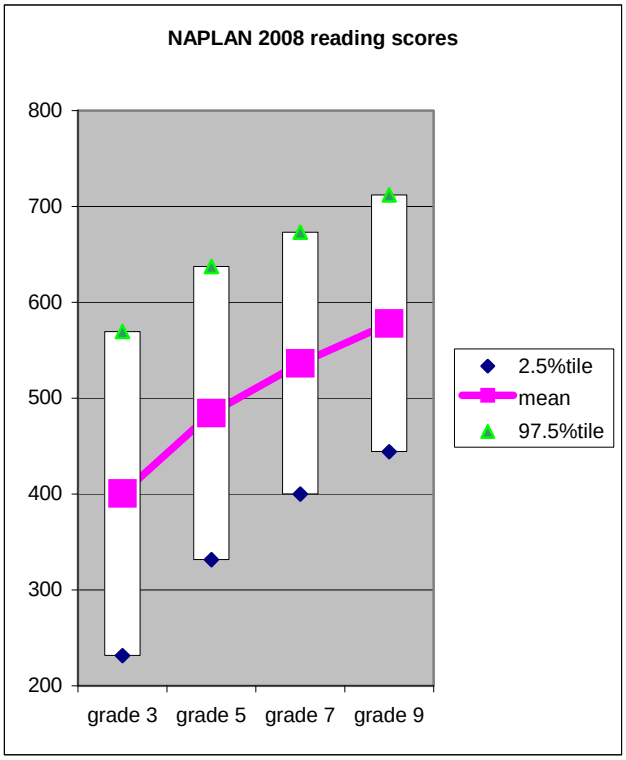

Figure 1 NAPLAN reading proficiency distributions in 2008

Table 1 shows the mean and standard deviation for each grade level.

Table 1 Mean and Standard Deviation of NAPLAN 2008 Reading Scores by Grade

NAPLAN reading	Grade 3	Grade 5	Grade 7	Grade 9
Mean	400.5	484.4	536.5	578.0
Standard deviation	84.5	76.5	68.2	67
Growth rate (per grade) in effect size	0.5	0.34	0.3	-

A number of observations can be made from Figure 1 and Table 1. First, the student proficiency distribution is wide in comparison to the average growth between grades. For example, students in Grade 3 have a wide range of abilities, covering the range of mean scores of several grades. Second, the growth rate is the largest for lower grades, and diminishes for higher grades. Third, the growth rate *per grade* is around 0.5 of a standard deviation for primary school years, an effect size[3] of 0.5 (half a standard deviation). This forms a frame of reference for evaluating the accuracies of achievement estimates. For example, we would not expect changes (trends) from one calendar year to another to be much more than one or two months of growth (effect size of 0.1), if any. (One would be working miracles to increase a cohort's result by one year's growth as compared to the previous year's cohort.) According to figures for NAEP long term trends (Mazzeo & von Davier, 2009), on average the change across time is typically very small. Table 2 is an excerpt from Mazzeo and von Davier (2009), Tables 1 and 2, page 13. For NAEP, in Reading, there appears to be no significant change over time. In Mathematics, there appears to be an increase over time, although small. The largest change is between 1999 and 2004 with an effect size of 0.15. During this period, the No Child Left Behind Act came into effect (although this Act seems not to have an impact on Reading).

[3] Note that this is an average effect size. In other studies, we have found that the growth between Grades 3 and 4 is around an effect size of 1, and the growth diminishes as Grade increases to around 0.3 effect size for Grades 8 and 9.

Table 2 Change Measured by Effect Size from NAEP Long Term Trend Study (Mazzeo and von Davier, 2009)

	Reading	Mathematics
Year	Change (measured by effect size)	Change (measured by effect size)
1990	-0.02	0.04
1992	0.08	0.08
1994	-0.05	0.04
1996	0.00	0.00
1999	0.04	0.05
2004	-0.02	0.15

According to the above information, any estimates of trends that have a 95% confidence interval of more than 0.1 effect size[4] (standard error of around 0.05 effect size or higher) would seem to have insufficient power to measure trends. The information provided in this section will serve as a reference for evaluating various sources of error in relation to the purpose of the surveys.

The following sections discuss three different sources of error in a survey and their impact on survey results.

4. Measurement Error

Measurement error usually refers to inaccuracies associated with a measuring instrument. In the case of measuring student achievement, the instrument is usually a test. For example, if two *parallel*[5] tests are administered, we would not expect everyone to obtain exactly the same score on the two tests. If the test is short (i.e., with few items), one would expect the measurement error to be large. Measurement error diminishes as the number of items increases

[4] This is a conservative estimate. As comparisons of trends involved two estimates, the 95% confidence interval for each estimate will need to be even smaller to lead to statistical significance of an effect size of 0.1.

[5] In this paper, I use the term "parallel tests" to mean tests that tap into the same construct and have the same level of difficulty.

in a test. The size of the measurement error can be estimated if a number of assumptions are made. For example, if the item response data follow a Rasch item response model (Rasch, 1960, 1977), one can obtain estimates of the sizes of measurement error, as shown in Table 3.

Table 3 Measurement Error and Test Length

Test Length	Measurement error (standard error in logits)[6]	Number of score points equivalent to one grade level (0.5 logit)[7]
20 items	0.51	2
30 items	0.43	4
40 items	0.36	5
60 items	0.30	7

To place the magnitude of the measurement error in perspective, let us take a look at a typical student achievement distribution from a large-scale survey conducted in an Australian state. Figure 2 shows the mathematics ability distribution for a random sample of Grade 4 students in a large-scale survey. The horizontal axis shows student ability in logits from Rasch model analysis. Mean achievement for each grade level is also marked on the horizontal axis. The size of the measurement error for a test with n items (where n takes values 20, 30, 40 and 60) is displayed by a horizontal line with arrows on the ends, showing a 95% confidence interval for an ability estimate of a student taking a test of length n items.

A number of observations can be made about Figure 2. First, it is noted that the ability distribution for one grade level is wide in comparison to the differences between mean

[6] The derivation of these figures is given in the Technical Appendix at the end of this document.
[7] These can be estimated from a score equivalence table produced by ConQuest, for example. They can also be estimated approximately, as shown in the Technical Appendix.

achievement across grades. This is in agreement with the results shown in Figure 1. That is, Y3 mean to Y6 mean all seem to be within the normal range of Grade 4 abilities. Second, the average growth per grade is about 0.5 logit[8] which is about half of a standard deviation. This, again, is in agreement with results shown in Table 1. Third, the measurement error for individual students taking one test of 60 or fewer items is very large in relation to the spread of the ability distribution, and in relation to the differences in mean scores across grade levels. What this means is that the measurement of each student is not at all precise. The 95% confidence interval of individual ability estimate covers several grade means. If our goal is to place a student at a grade-equivalent level, a test with 60 or fewer items will not do a good job. Further, the last column in Table 3 shows the number of score point difference equivalent to one grade level. For example, if a student obtained 11 out of 20, and another student obtained 9 out of 20, their ability difference in logits (an IRT measurement unit) will be around 0.5 logit, placing them one grade apart.

If our goal is to see if a student has grown from one year to another, there is little chance that the expected growth for one grade (around 0.5 logit) can be detected with such large uncertainty around the ability estimate from a 40-item test, a typical test length for state-wide or nation-wide tests.

[8] Again, note that the growth is more than 0.5 logit for Grades 3 and 4, and diminishes for higher grades.

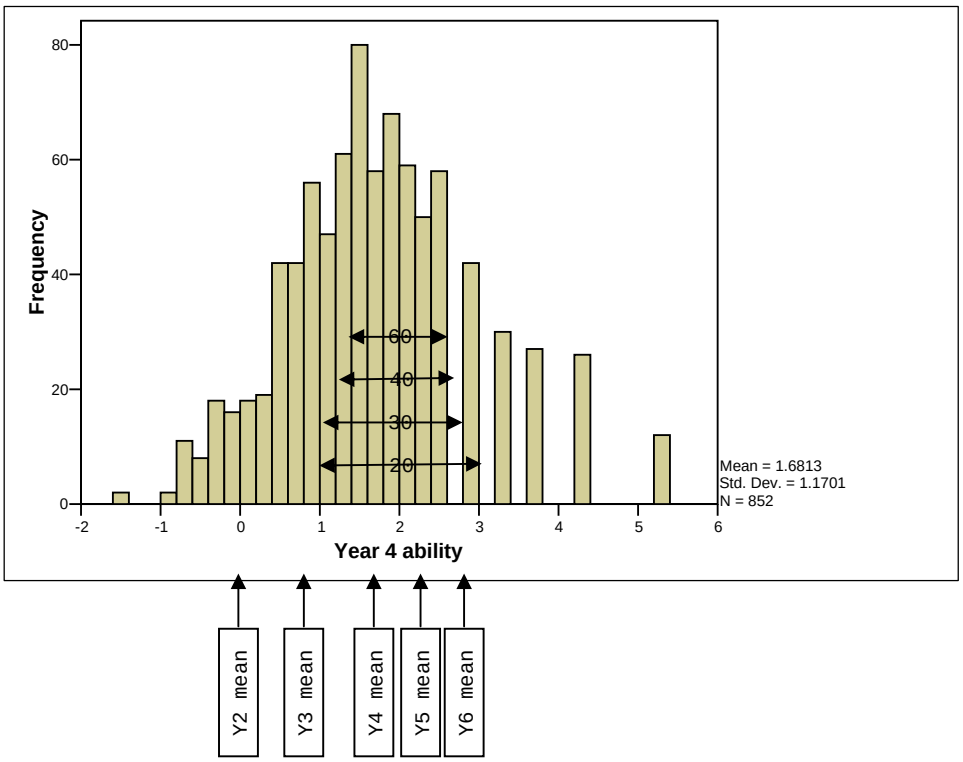

Figure 2 Grade 4 students' mathematics ability distribution in a large-scale survey

In fact, in tracking individual students' progress, it has often been noted that a number of students "go backwards" from one year to another when annual state-wide or nation-wide tests are conducted. Given the magnitude of the measurement error, it is not surprising that some students appear to perform worse than they did in the previous year. With a 40-item test, it is expected that 16%[9] of the students would appear to regress "backward" when they have actually made one year's progress (assumed to be 0.5 logit). This percentage estimate is conservative, as it does not take into account other sources of error such as the assumption that two tests are assessing the same content. In summary, any tracking of individual students' progress cannot be achieved reliably with two 40-item tests conducted at one year intervals because of the size of the measurement error and other measurement disturbances. The size of the measurement error

[9] The derivation of this percentage is provided in the Technical Appendix.

will reduce if we are interested in tracking groups of students, such as in a class or in a school. However, the error is still relatively large for a group/class of 30 students. Table 4 shows estimates of standard errors for mean growth scores for groups of various sizes, based on measurement error alone. The derivation of these figures is given in the Technical Appendix. Based on measurement error alone, the standard error of the class mean is 0.09 logit for a typical class of size 30. This means that the 95% confidence interval of the class mean has a width of around 0.36 logit, over half a year's growth, on average. This uncertainty is quite large. Any suggestion that teacher performance should be determined by such measurement of student growth is of serious concern.

Table 4 Estimated Standard Error of Mean Growth for a Group

Number of students in a school	Estimated standard error of mean growth for a school (in logit)
30 (one class)	0.09
60 (two classes)	0.07
100	0.05

In Section 6, I will show that systematic errors from equating two tests will further add uncertainty to the estimates, and these systematic errors will not diminish with increasing sample size. This source of systematic error is by far the largest in computing trend estimates.

5. Sampling error

If a survey is sample-based, then errors due to sampling variation need to be taken into account. Even if sampling is not involved and a full cohort of students is tested, one can still regard that a specific cohort is a sample from all possible cohorts, particularly in comparing trends.

In large surveys of students, typically two-stage sampling is conducted: schools are selected and then students within schools are selected (e.g., IEA, 2004, Adams & Wu, 2000). Such two-stage sampling is more convenient logistically, but a great deal of sampling efficiency (the accuracy of estimates) is lost due to intra-class correlation[10] (students within schools are more similar than students across schools). In many cases, we need to increase the sample size by five times or more to achieve an accuracy equivalent to using a simple random sample. An often adopted standard for large-scale surveys is that the 95% confidence interval for the mean score should be within ±10% of the standard deviation of the variable (e.g., IEA, 2004, p114). This is achieved by using a sample size of 400 if a simple random sample is selected. Whether this degree of accuracy is sufficient depends on the goal of the assessment. For example, in the Australian context, a comparison of state means is of interest. On a scale where the national average is set to 500 and the standard deviation is set to 100 (where each state also has a standard deviation close to 100), the state means will have a confidence interval of ±10 (where 10=0.1×standard deviation) if an effective sample size of 400 is achieved in each state. Based on this degree of accuracy, statistically significant difference at 95% confidence level can only be found if the observed mean difference is greater than 14 score points (effect size of 0.14). Comparing this magnitude to the NAEP trend results shown in Section 3 (Table 2 in particular), it would seem that this degree of accuracy is on the borderline of providing adequate trend measures.

[10] Typically, students within schools are more similar than students across schools. If 35 students within a school are selected, they do not provide as much 'information' as 35 students randomly selected from all schools. Consequently, a sampling procedure involving the selection of schools first and then students within schools generally results in larger standard errors than for simple random samples of the same size.

The following is an example, showing an excerpt of results from the 2004 Australian National Assessment Program – Civics and Citizenship survey for Year 10 (MCEETYA, 2006, Table 4.2).

			NSW	ACT	VIC	NT	TAS	WA	QLD	SA
		Mean	521	518	494	490	489	486	469	465
	Mean	95% CI	10.6	21.5	19.0	33.2	16.6	17.5	17.6	16.2
NSW	521	10.6		●	●	●	●	●	▲	▲
ACT	518	21.5	●		●	●	●	●	●	●
VIC	494	19.0	●	●		●	●	●	●	●
NT	490	33.2	●	●	●		●	●	●	●
TAS	489	16.6	●	●	●	●		●	●	●
WA	486	17.5	●	●	●	●	●		●	●
QLD	469	17.6	▼	●	●	●	●	●		●
SA	465	16.2	▼	●	●	●	●	●	●	

Figure 3 2004 NAP-Civics state mean score comparisons

The circles in Figure 3 indicate that the difference is not statistically significant. Note that this does not mean that there is no difference. It means that the sample size does not have the power to detect the difference. The difference in means between NSW and WA is 35 score points (about 0.3 in effect size, given that the standard deviation is 121), yet the statistical test is still not significant. The lack of power[11] in detecting differences in state means is due to the small sample size. Further, the large confidence interval also indicates that the mean estimates are not very 'accurate'. For example, while the reported mean score for SA is 465, we are only 95% confident that the mean lies in the range 449 and 481. In 2007, a repeat survey was conducted. The sample size was further reduced in this survey. A comparison of 2007 state means (MCEETYA, 2009) with 2004 state means is shown in Table 5.

[11] Strictly speaking, we should be examining Type II error to assess the (lack of) power of detecting differences. The statistical significance tests carried out in Figure 3 were based on Type I error.

Table 5 Comparison of 2004 and 2007 State Means in the Australian NAP-Civics Study

State	2004 state mean with 95% CI	2007 state mean with 95% CI
NSW	521 (±10.6)	529 (±17.0)
ACT	518 (±21.5)	523 (±19.6)
VIC	494 (±19.0)	494 (±17.1)
NT	490 (±33.2)	464 (±38.1)
TAS	489 (±16.6)	485 (±16.0)
WA	486 (±17.5)	478 (±22.6)
QLD	469 (±17.6)	481 (±13.9)
SA	465 (±16.2)	505 (±23.4)

From Table 5 it can be seen that the SA mean score has increased by 40 score points (about 2 years of growth in this domain), and the NT mean score has dropped by 25 score points. Such fluctuations in estimated mean scores hardly seem credible (when people ignore the large uncertainty and try to interpret the mean scores). This is typically the consequence of insufficient sample size, leading to inaccurate and thus fluctuating estimates.

In summary, sample size needs to be determined based on the purposes of the survey. For example, while we do not know the exact differences between state means (and, indeed, this is one purpose of the survey), we *do* have an idea of the order of magnitude of the differences. If we choose an effective sample size (such as 150 students, as in the example above) that is not sufficient to separate the state means, then it is a waste of time to carry out the survey because the survey result of finding no significant differences between the state means can be largely predicted without carrying out the survey.

The computation of the standard error presented in this section is based on sampling error alone. Given that there are systematic errors in equating, the confidence interval of means is much larger than the figures quoted in this section.

6. Equating error

Many large-scale surveys are conducted for the purpose of *monitoring student progress*. Consequently, tracking student performance over time is often an objective in many large-scale surveys. To be able to estimate trends in student performance, *equating* needs to be carried out between surveys at different time points. Equating has been an important area of psychometric research, and many methodological advances have been made over the past two decades (e.g., see Kolen & Brennan, 2004). Nevertheless, the specific issue of the impact of (IRT) model violation on equating has not received as much attention as it should have as, typically, equating error is not included in the computations of trend estimates, while this kind of error is by far the largest in the estimation of cohort means (Michaelides & Haertel, 2004). Model violation includes multi-dimensionality of items in a test, differential item functioning for subgroups, effect of item position in a test booklet (fatigue effect and context effect), and marker inconsistencies. The result of model violation is that item difficulty parameters are no longer invariant across different administrations of the same items, leading to serious errors in equating tests (Monseur & Berezner, 2007; Monseur, Sibberns & Hastedt, 2008). In summary, the problem arises when common items in two tests used for equating are not *working* in the same way. There are a number of reasons for the same items to have different difficulties in different tests. Some key reasons are discussed below.

Multiple constructs in one test

Typically, a domain for assessment (e.g., mathematics or reading) is multidimensional (Lissitz & Huynh, 2003), and it is difficult to capture all facets of the domain in a single test unless there are many items in the test (e.g., a few hundred items). For example, a single test of 40 mathematics items will not likely represent all facets of mathematics. Consequently, two 40-

item tests will likely be tapping into different constructs. This means that individuals and groups are likely to get different results on the two tests, over and above measurement error. Of course, the notion of different constructs is by degree: every item measures something unique as compared to other items. It is a matter of assessing whether differences in the constructs of two tests are sufficiently large to threaten the usefulness of the results. Wu (2009a) showed that different compositions of mathematics content areas led to different rankings of countries in PISA and TIMSS. That is, if one purpose of the survey is to compare country mean performance, then one must interpret the results in terms of the content balance in the respective test.

I would also like to note that there is a common, and incorrect, perception that there is fairness when everyone takes the exact same test. Actually, each individual person also has something unique (in ability), and one test may favour some but disadvantage others.

Differential Item functioning

Differential item functioning (DIF) refers to items that are relatively easier or more difficult for different groups of people. Clearly, DIF threatens the validity and reliability of survey results as it is open to manipulation in selecting items favouring certain groups. In international surveys, it is becoming clear the extent to which items function differently across different countries and cultural groups. Wu (2009b) identified significant differential item functioning between Asian and Western countries for items on formal mathematics, and items with real-life applications. Monseur and Berezner (2007) also showed the large impact of the selection of reading items on country performance. For example, for PISA 2003, Japan's mean score would increase by 10 score points if one particular reading unit was removed from the set of 8 anchoring units. A difference of 10 score points on the PISA scale is about 10% of the standard deviation (effect size of 0.1 (see Section 3)). This is large enough to change country

ranking orders, and large enough to make trend estimates entirely meaningless. Given that the inclusion of particular reading units and items is largely by chance (one could have developed a completely different set of reading units and items if a different test development contractor was involved), it is alarming to assess how confident we are regarding the results.

An example is given below. For PISA 2000 Reading, item difficulty parameters for Japan are calibrated using only Japanese data. These item parameters are then compared with international item parameters. If there is no differential item functioning, then a scatter plot of the two sets of item parameters should fall reasonably around a straight line. Figure 4 shows the scatter plot (where each set of item difficulty parameters is centred at zero).

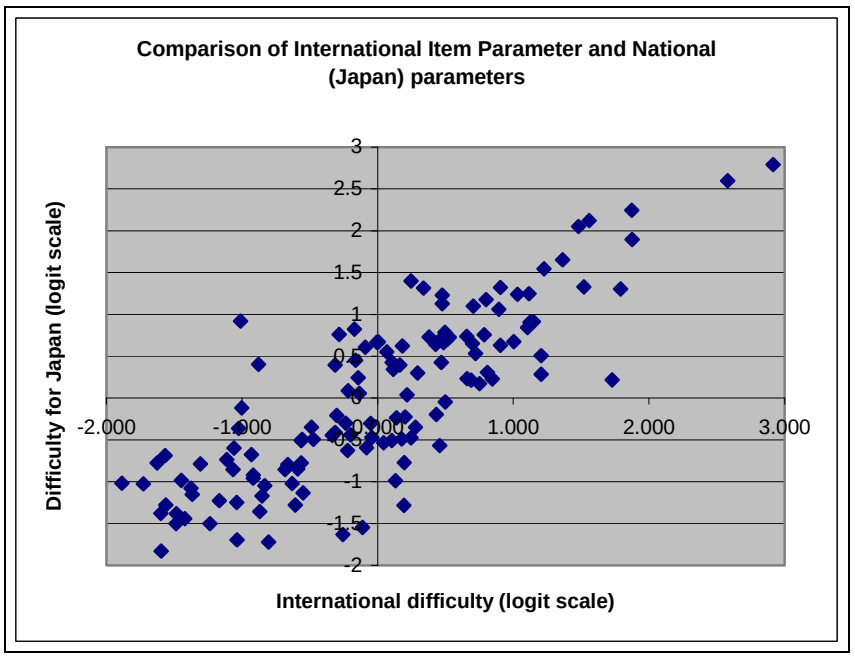

Figure 4 Comparison of international item difficulty and item difficulty for Japan

Figure 4 shows that there is a trend that an item that is difficult internationally is typically more difficult for Japan. However there are considerable variations at individual item level. For example, there are items where the item difficulties between international calibration and calibration for Japan differ by more than 1 logit! It is clear that the extent of differential item

functioning is substantial. For equating, if link items are chosen that happen to be more difficult for Japan than for other countries, then Japanese results for PISA 2003 can drop considerably. In fact, in 2003, Japanese mean score decreased by 24 PISA points. The set of link items chosen in PISA 2003 Reading are, on average, more difficult for Japan. The average difference between Japanese item parameters and international parameters is 0.08 logit for the link items. This difference is equivalent to about 8 PISA score points.

Differential item functioning is often present for other subgroups of students as well. For example, Wu and Les (2008) showed that boys performed equally well as girls on reading texts that are scientific or factual, but boys fell behind girls on narrative texts and those dealing with social relationships. Consequently, the proportion of different text types in a test will determine how large the gap is between girls' and boys' performance in reading.

In summary, the stability of test results across time points will greatly depend on the compositions of the tests. In many large-scale surveys, there has not been sufficient investigation into differential item functioning, and into the stability of the results should a different set of items be included.

Effect of item position in a test booklet

It has been observed that items placed in different *positions* in a test (i.e., at the beginning, in the middle, or at the end of a test) typically have markedly different item difficulties. Monseur (see Appendix 6, p 282 in Adams & Wu, 2000) showed that the item difficulty parameter for an item could differ by about 0.5 logit depending on whether the item was placed at the beginning or at the end of a test booklet. Wu (2002) also showed that the percentages correct dropped when an item was placed in a latter part of a test booklet in PISA 2003 field trial (see Table 6).

Table 6 Comparison of Percentages Correct of Items Placed at Different Positions in a Test Booklet

Item Code	Item Format	Item Position							
		Block 1		Block 2		Block 3		Block 4	
		%Cor	Om	%Cor	Om	%Cor	Om	%Cor	Om
M402Q01	OE	60 (61)	2.1					49 (54)	10.1
M402Q02	OE	30 (36)	17.5					23 (32)	28.0
M434Q01	OE	16 (16)	3.0					12 (13)	6.3
M453Q01	MC	56 (57)	1.9					49 (53)	8.4
M543Q01	MC	58 (60)	3.4					50 (54)	7.3
M520Q02	MC			52 (54)	3.2	49 (51)	3.4		
M520Q03	OE			49 (52)	5.4	48 (52)	6.5		
M500Q01	OE			68 (80)	15.6	63 (76)	18.0		
M455Q01	OE			13 (19)	33.4	11 (16)	32.3		

In Table 6, the ‘Item Format’ column shows whether the item is open-ended (OE) or multiple-choice (MC). The labels ‘Block 1’, ‘Block 2’, ‘Block 3’ and ‘Block 4’ indicate the first quarter, second quarter, third quarter and fourth quarter of an assessment booklet where the item was placed. The column headed ‘%Cor’ shows the percentage correct for the item in that booklet position. There are two numbers under this column. The first number is the percentage correct of the total number of students who were administered the item (i.e., including omissions but excluding not-reached). The second number (in brackets) is the percentage correct of the total number of students who gave some kind of response (i.e., excluding omissions and not-reached). As the students who omitted an item were mostly lower ability students, the percentage correct in the brackets is expected to be higher compared to the percentage correct of all students. This indeed was found to be the case. The column with heading “Om” shows the percentage of omissions.

Table 6 clearly shows that there was a fatigue effect as the testing went on. The percentage correct decreases when an item is placed in the latter part of a test booklet, even after we adjust for omission. For example, for item M402Q01, the raw percentages correct were 60

and 49 as the item appeared in Block 1 and Block 4 respectively. The corresponding percentages correct when we exclude students who omitted the item were 61 and 54, even though the group of students who attempted the item in Block 4 were of higher average ability than those who attempted Block 1 (having more able students who reached the end of the booklet). This is confirmation that performance deteriorated towards the end of the test. A similar pattern can be found with Blocks 2 and 3, although the differences between these two blocks are not as great as the differences between Blocks 1 and 4.

The implication of item position effect is that, for equating, link items must be placed at the same position in a test booklet, otherwise the item will not have the same item difficulty, and there will be a systematic error in equating. If, by chance, all link items appear at the end of a test booklet (e.g., Grade 3 test), and they appear at the beginning of another test booklet for equating (e.g., Grade 5 test), as is typically the case for vertical equating, the systematic equating error can be considerably large. Wu (2009c) showed that *booklet effect* (i.e., increase in item difficulty due to item placements in a booklet) could be as high as 0.4 logits in an IRT equating study (effect size of 0.4), making any estimation of mean scores, and of trends, entirely off the track.

The second issue often raised in relation to item position is the context effect that the item difficulty of an item may change depending on which items precede the item. Mazzeo and von Davier (2009) discussed the so-called NAEP 1986 Reading Anomaly (Beaton, 1988; Beaton and Zwick, 1990) where anomalous results were found due to changes in booklet design. As a consequence, NAEP now carries out a separate Long Term Trend (LTT) survey by using the same assessment booklets. Consequently, two separate streams of NAEP testing are carried out – the Main NAEP uses tests that reflect changes in content framework and assessment methodologies while LTT NAEP uses the original framework and test.

Equating across grade levels (vertical equating)

In relation to building scales across several age/grade levels, vertical scaling requires sophisticated statistical methodology to ensure that the results are reliable and valid. Haertel (1991) compared building a single, cross-age, scale with building separate scales within each age group. Haertel reported that the NAEP technical review panel found that, while NAEP vertical equating did not appear to be flawed when checks were made, it recommended that within-age scales should be used whenever feasible after evaluating the relative merits of within-age and across-age scales, such as the utility of scale scores and misuses of these scores. Psychometricians have typically expressed reservations about the validity of vertical equating (e.g., Lissitz & Huynh, 2003). In general, the further the grades are apart the less reliable the vertical equating across grades is found to be. Lissitz and Huynh recommended a combination of within-age and across-age scales, where within-age scales are constructed but the separate scales for different grades have a set of common policy definitions, such as the meaning of being *proficient* within the grade level.

In my experience, equating involving multiple-grade levels often produced different results depending on the equating methodology used. This is an indication that the item response data do not fit the underlying model and model assumptions are violated. The only situation where I found stability in vertical equating was when the tests were multi-level where a test booklet contained a large number of items arranged in difficulty order, and each student began the test from an entry point as advised by his/her teacher. In other words, this is a form of adaptive testing. In this way, items were not always in the same position as it depended on the entry point, and the linking of all items was extremely strong.

However, in general, one should remain cautious of the results obtained from vertical equating.

Magnitude of equating error

The problems as discussed above in relation to the violation of the measurement model can lead to substantial equating error. Michaelides and Haertel (2004) found that, with 44 common items in equating two tests, the standard error due to the sampling of common items (i.e., should the set of common items be different) is by far the largest in the estimation of the mean score of a cohort (effect size of about 0.03). In general, the error sources associated with the estimated mean score have three components: error due to the sampling of common items, error due to the sampling of students, and error in measuring individual students. Michaelides and Haertel computed the relative magnitude of these three sources of errors in terms of the percentages of the total error. They were 83%, 11%, 6% for errors due to sampling of items, sampling of students, and measurement error respectively. Monseur and Berezner (2007) have shown that the standard errors due to the selection of common items for PISA 2003 reading assessment range between 6 to 20 PISA scale points (effect size of about .06 to 0.20) at the country level. This means that the 95% confidence interval for the mean score will be between 24 and 80 PISA score points wide from this source of error alone. Recall that expected changes between cycles are less than an effect size of 0.1 (less than 10 PISA score points); this makes any trend estimates totally meaningless!

Typically, in reporting equated mean scores the error due to the sampling of common items has often been ignored. In PISA, equating error is reported and it is around 5 PISA score points (which, according to Monseur and Berezner, 2007, is an under-estimate). Even using this

conservative estimate, it means that the 95% confidence interval of the mean score is around 20 score points wide (about half a year's growth) just based on equating error.

There may be an argument that equating error only needs to be reported when trend is reported. However, I would argue that a large equating error due to the sampling of common items suggests that country results (or any cohort result) are dependent on the particular set of items included in a test. Consequently, large equating error means that country mean scores can vary according to the set of items included in a test, so that in one survey the relative standings of countries according to their mean scores are subject to a larger source of error than reported. That is, for any survey (and not just in the context of estimating trends), one should consider the effect of changing the item set, and include this source of variation in the computation of the standard error. This is particularly important for surveys that contain a small number of items, such as PISA minor domains and national tests with only 40 items per domain.

7. Summary of issues in relation to large-scale assessments

Table 7 summarises the issues associated with particular types of assessment.

Table 7 Summary of Characteristics and Issues in Relation to Large-scale Assessments

Type of assessment	Example	Assessment Characteristics and Issues
Test consisting of one instrument (e.g., 40 items) where every student takes the same instrument.	• NAPLAN	• Large measurement error at individual student level. Large error even at class level. No high-stakes decision should be made based on these results (e.g., teacher performance should not be judged by student scores in these tests). • No power to measure growth between two time points for individual students or at class level. • Equating is difficult because of fixed item positions and differential item functioning. • Test validity is not well-established due to the lack of coverage of content areas.
Test consisting of a number of rotated test forms, where each student only takes a small number of items.	• PISA • TIMSS	• Typically, results at individual student level are not of interest. • Item position effect can be moderated through a balanced test booklet design. • For international surveys, item by country interaction is very large. This threatens the validity of the results. If the number of common items for equating is small, equating error could be very large.
Sample-based assessment.	• NAP-Civics	• If two-stage sampling is used (selecting schools and then selecting students within schools), sampling efficiency could be reduced by a large factor. The result is that the sample may not be sufficiently large to provide any useful information.
Assessment across a number of grade levels.	• NAPLAN	• Vertical equating is difficult due to item position effect, curriculum differences across grades, and the lack of coverage of content areas.

8. What are the solutions?

In this paper, I considered three main sources of error: measurement error at individual student level, error due to the sampling of students (for cohort results), and item sampling error due to the selection of common items for equating. Increasing the test length will reduce measurement error at individual student level. Increasing student sample size will reduce sampling error. Increasing the number of common items, and maintaining the same

administration conditions, will reduce the equating error. However, there are difficulties in implementing all these solutions at the same time. For example, it will be difficult to increase the test length because of practical considerations such as fatigue. To overcome this difficulty, we can administer tests on multiple occasions. However, increasing the number of test administrations and student sample size will increase the cost and resources. So we must find cost-effective ways of testing students. The use of technology is part of the solution. Computer-delivered tests can reduce the cost considerably, as objective items can be marked automatically. Further, instant feedback can be provided to students and teachers. The current paper-and-pen format such as in NAPLAN has many drawbacks, in addition to the unreliability of the test results. To improve assessments, it is essential to utilise technology to build electronic item banks, and move towards computer-adaptive testing of students. Under such a system, students can be tested on multiple occasions throughout the year, the tests can be calibrated centrally, feedback can be provided immediately[12], and, above all, teachers can be relieved of many hours of marking. Computer-adaptive tests can reduce measurement error. Once such a system is built, a large number of students can be included in the assessment without too much additional cost so that sampling error will be reduced. Student progress can be monitored regularly under this system because there is longitudinal data at multiple time points throughout the year. Finally, with a wide coverage of content areas and moving away from fixed item positions in a test, equating error will also be reduced.

In terms of assessment design, one recommendation is to avoid constructing an assessment system that can "solve it all". Focused studies may well be more suited to establish the effectiveness of a particular intervention, or a particular policy change. Smaller, purposeful

[12] Currently, NAPLAN provides results four months after test administration.

and targeted assessment programs may achieve a narrow but well-defined set of objectives rather than a large-scale assessment system that does not providing any useful data.

9. Summary and Conclusions

Assessment has received more attention than ever as we move into the 21st century, particularly with the No Child Left Behind (NCLB) Act in the United States. A great deal of public money is spent on assessing students. There are international surveys, national testing and state-wide testing programs. These programs are costly. A close examination of many of the surveys and testing programs has revealed that the survey methodologies are often not meeting the objectives of the programs due to various factors. In particular, there is a lack of critical statistical validation of the survey results. Claims and conclusions are made without statistical rigour. Instead of proliferating testing programs, we should stop and reflect on whether objectives are met in the programs, whether and how improvements can be made, before millions of dollars are spent to produce invalid results or, often, useless results. There is an urgent need for stakeholders to be well informed of current assessment programs, and for those carrying out assessment programs to be responsible in designing assessment studies and reporting results. Often, statistical complexity prevents non-technical stakeholders fully appreciating the caveats in the results, leading to misinterpretation, over-interpretation and, even worse, making inappropriate policy decisions. It is the responsibility of assessment designers to ensure technical soundness of the surveys.

In some areas of educational research, specialists are carrying out sophisticated research in developing measurement models to improve educational measurement. The development of plausible values, latent regression, replication methods for computing standard errors, and sampling design, all employ complex quantitative modelling to improve the measurement of

student achievement. Yet all the improvements in the accuracy achieved by these methods are quickly thrown out by equating errors, by differential item functioning (between countries and between gender groups, for example), by fatigue in students during testing, by inconsistent marker performance, and by insufficient sample. There only needs to be one weak link in the chain of the assessment process to make results invalid. In my opinion, the current weakest link is in instrument design (sampling of items) and test design (the delivery of the test), and a lack of sampling rigour (e.g., sampling coverage and sample size). When the test instrument does not capture the full domain, and when items exhibit differential item functioning or, even worse, when poor items are included in a test or when insufficient number of students are sampled, all other efforts to improve measurement are carried out in vain. The mere implementation of an assessment program and production of a report will not serve the community any good unless valid results are produced. There are still many assessment programs that do not meet their objectives. A common scenario is that a government sets aside a budget for conducting an assessment program, and a contractor is appointed to conduct the program. The contractor designs a program based on the budget, and not on sound statistical basis. The program is completed and a report is produced. The usefulness of the results is not critically examined, and the validity of the results is not critically established. Nevertheless, accountability seems to have been achieved, with the money spent.

I hope this paper provides sufficient rationale to call on those involved in assessments to stop implementing large-scale assessment programs for the sake of implementing them. Rather, the benefits of implementing an assessment program need to be thoroughly reviewed, before so much public money is invested.

References

Adams, R.J., & Wu, M.L. (2000). *PISA 2000 technical report*, OECD, Paris.

Beaton, A.E. (1988). *The NAEP 1985-86 Reading Anomaly: A Technical Report*, ETS, Princeton, New Jersey, USA.

Beaton, A.E., & Zwick, R. (1990). *Disentangling the NAEP 1985-86 Reading Anomaly*, Princeton, New Jersey, USA.

Haertel, E.H. (1991). *Report on Technical Review Panel analyses of issues concerning within-age versus cross-age scales for NAEP*. (ERIC Document Number ED404367): Washington, DC: National Center for Education Statistics.

IEA (2004). *TIMSS 2003 Technical Report*. Chestnut Hill, M.A: TIMSS International Study Centre.

Kolen, M.J., & Brennan, R.L. (2004). *Testing equating, scaling, and linking – Methods and practices.* Springer.

Lissitz, R.W., & Huynh, H. (2003). Vertical equating for state assessments: issues and solutions in determination of adequate yearly progress and school accountability. *Practical Assessment, Research & Evaluation, 8(10).* Retrieved July 19, 2009 from http://PAREonline.net/getvn.asp?v=8&n=10.

Mazzeo, J., & von Davier, M. (2009). Review of the Programme for International Student Assessment (PISA) test design: Recommendations for fostering stability in assessment results. Retrieved July, 2009, from http://edsurveys.rti.org/PISA.

MCEETYA (2006). National Assessment Program – Civics and Citizenship, Year 6 and Year 10 Report 2004. Ministerial Council on Education, Employment, Training and Youth Affairs.

MCEETYA (2008). National Assessment Program – Literacy and Numeracy. Retrieved July, 2009, from http://www.naplan.edu.au/verve/_resources/2ndStageNationalReport_18Dec_v2.pdf

MCEETYA (2009). National Assessment Program – Civics and Citizenship, Year 6 and Year 10 Report 2007. Ministerial Council on Education, Employment, Training and Youth Affairs.

Michaelides, M.P., & Haertel, E.H. (2004). *Sampling of common items: An unrecognized source of error in test equating.* Center for the Study of Evaluation (CSE), CRESST, University of California, LA.

Monseur, C., & Berezner, A. (2007). The computation of equating errors in international surveys in education. *Journal of Applied Measurement*, 8(3), 323-335.

Monseur, C.H., Sibberns, H., & Hastedt, D. (2008). Linking errors in trend estimation for international surveys in education. In von Davier, M and Hastedt, D. (Eds.), *Issues and methodologies in large-scale assessments*, IEA-ETS Research Institute. Hamburg, Vol. 1, p113-122.

Mullis, I.V.S., Martin, M.O., Gonzales, E.J., & Chrostowski, S.J. (2003). TIMSS 2003 International Mathematics Report. Chestnut Hill, M.A: TIMSS International Study Centre.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy*, 14, 58-94.

Wu, M.L. (2002). Timing issues in PISA 2003 field trial – PEG, MEG, SEG discussion paper, September 2002, PISA.

Wu, M.L. (2009a). A Comparison of PISA and TIMSS 2003 Achievement Results in Mathematics and Science. *Prospects*, UNESCO. In Press.

Wu, M.L. (2009b). *A Critical comparison of the contents of PISA and TIMSS mathematics assessments.* Paper presented at the NCES "What we can learn from PISA" research conference held on June 2, 2009, Washington D.C.

Wu, M.L. (2009c). Facets – Analysing booklet effect. Topics by Example on ConQuest website. Retrieved August 2009 from http://www.conquestwebsite.com/resources.html.

Wu, M.L., & Les, M. (2008). The difference between boys and girls. *EQ. Summer 2008, p35-36*. Curriculum Corporation. Melbourne.

Technical Appendix

1. Computation of Measurement Error

Suppose a test has I dichotomous items with item difficulties δ_1, δ_2,…, δ_I. Let θ_n denote the ability of a person, and $\hat{\theta}_n$ denote the maximum likelihood estimate of θ_n. Then, the measurement error is $\sqrt{\text{var}\left(\hat{\theta}_n\right)}$, where

$$\begin{aligned}\text{var}\left(\hat{\theta}_n\right) &= -\left[\frac{\partial^2 \lambda\left(\Theta|\mathbf{X}\right)}{\partial \theta_n^2}\right]^{-1} \\ &= \left[\sum_{i=1}^{I} \Pr\left(X_{ni}=1;\hat{\theta}_n,\hat{\delta}_i\right)\left(1-\Pr\left(X_{ni}=1;\hat{\theta}_n,\hat{\delta}_i\right)\right)\right]^{-1} \\ &= T^{-1}\end{aligned}$$

where T is the test information function.

In Table 3, the item difficulties for a 20-item tests were taken to be -2, -1.8, -1.6, …, and the ability was taken to be zero. That is, the test is well-targeted in the sense that the ability is at the middle of the item difficulty range.

For a 30-item test, the item difficulties were taken to be -2.15, -2.0, -1.85, …

For a 40-item test, the item difficulties were taken to be -2.0, -1.9, -1.8, …

2. Estimation of Ability Difference in Relation to Raw Score Changes

To investigate the sensitivity of changes in ability estimate as a function of changes in raw score, one can use the following approximate procedure.

Let I be the total number of items in a test. For computational simplification, assume that the item difficulties of all I items are equal (δ). Further, assume that the raw score for this person is $I/2$. That is, the person obtained the correct answers for exactly half of the items in the test. Then, the estimated ability of the person is equal to the item difficulty, δ. That is,

$$
\begin{aligned}
Raw\ score &= \frac{I}{2} \\
&= \sum_{i=1}^{I} \frac{\exp(\theta-\delta)}{1+\exp(\theta-\delta)} \\
&= \sum_{i=1}^{I} \frac{\exp(\delta-\delta)}{1+\exp(\delta-\delta)} \\
&= \sum_{i=1}^{I} \frac{1}{1+1} = \frac{I}{2}
\end{aligned}
$$

If the ability is now 0.5 logit less than δ, what is the expected raw score difference from $I/2$? Let d be the raw score difference, then we need to solve the following equation

$$
\begin{aligned}
&\frac{I}{2} - d \\
&= \sum_{i=1}^{I} \frac{\exp(\theta-0.5-\delta)}{1+\exp(\theta-0.5-\delta)} \\
&= \sum_{i=1}^{I} \frac{\exp(\delta-0.5-\delta)}{1+\exp(\delta-0.5-\delta)} \\
&= \sum_{i=1}^{I} \frac{\exp(-0.5)}{1+\exp(-0.5)} \\
&= I \times 0.377541
\end{aligned}
$$

So

$$d = \frac{I}{2} - I \times 0.377541$$

The following table shows various values of the test length, I, and the raw score difference, d.

Test length, I	**Raw score difference, d, corresponding to 0.5 logit change in ability**
20	**2.45**
30	**3.67**
40	**4.90**

A simple approximation to work out the number of score point difference that is equivalent to 1 logit difference is $\frac{I}{4}$, where I is the total number of items. This is because the test information is $\frac{I}{4}$ if all items matched the person ability. That is, the derivative of the test characteristic curve is $\frac{I}{4}$ at θ where the expected test score is $\frac{I}{2}$.

3. Estimation of the Percentage of Students Regressing Backwards

Assume that a student was measured at two time points, t_1 and t_2, and that the student's true ability has increased by 0.5 logit between t_1 and t_2. Owing to measurement error, it is possible that the estimated ability at time point t_2 is less than the estimated ability at time point t_1. An estimation of the probability that $\hat{\theta}_{t_1} > \hat{\theta}_{t_2}$ is given below.

Assume $\hat{\theta}_{t_1}$ is sampled from a Normal distribution with mean 0 and standard deviation 0.36 (measurement error for a 40-item test), and $\hat{\theta}_{t_2}$ is sampled from a Normal distribution with mean 0.5 and standard deviation 0.36. Then,

$$\begin{aligned}\Pr\left(\hat{\theta}_{t_1} > \hat{\theta}_{t_2}\right) &= \int_{-\infty}^{\infty}\left[\int_{y}^{\infty} f\left(x\right)dx\right]g\left(y\right)dy \\ &= \int_{-\infty}^{\infty}\left[1-F\left(y\right)\right]g\left(y\right)dy\end{aligned}$$

where $f\left(x\right)$ is the pdf of N(0,(0.36)^2), $g\left(y\right)$ is the pdf of N(0.5,(0.36)^2), and $F\left(y\right)$ is the cdf of N(0,(0.36)^2). The above integral is computed numerically.

4. Standard Error of Mean Ability Estimate for Groups of Students

If the population variance of ability estimates is 1.0 logit, then the standard error of the mean ability of any n randomly selected students can be taken as

$$\frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{n}}$$

However, groups such as schools or regions that are of interest typically consist of students who are more similar in ability. In a number of surveys conducted in Australia, the intraclass (more precisely, intraschool) correlation in Australia is around 0.2. That is, the between school variance of ability is around 20% of the total variance. And the within school variance is around 80% of the total variance. If the total variance is 1 logit, then the between school variance is around 0.2 logit, and the within school variance is around 0.8 logit. Consequently, the standard

deviation of ability distribution within a school could be around 0.9 logit. The standard error of mean ability for a school could be around

$$\frac{0.9}{\sqrt{n}}$$

where n is the number of students in the school.

5. Standard Error of Growth Measure for a School

If $\hat{\theta}_1$ is the estimated mean ability for a grade level, and $\hat{\theta}_2$ is the estimated mean ability for the same students measured one year later, then $\hat{\theta}_2 - \hat{\theta}_1$ provides a measure of growth for this cohort of students. The standard error of $\hat{\theta}_2 - \hat{\theta}_1$ can be estimated as follows:

$$\begin{aligned}
\text{var}\left(\hat{\theta}_2 - \hat{\theta}_1\right) &= \text{var}\left(\hat{\theta}_2\right) + \text{var}\left(\hat{\theta}_1\right) \\
&= \frac{1}{n}\text{var}\left(\hat{\theta}\right) + \frac{1}{n}\text{var}\left(\hat{\theta}\right) \\
&= \frac{1}{n} \times 0.36^2 + \frac{1}{n} \times 0.36^2 \\
&= \frac{1}{n} \times 0.259
\end{aligned}$$

where 0.36 is the measurement error for a 40-item test, and n is the number of students.