Using the Rasch Model in the Design of a New Curriculum Framework and to Moderate Teacher Assessments Within It

Andrew Smith[1]
Office for Educational Review, Department of Education (Tasmania)

Abstract

In 2003, the Tasmanian Department of Education published *Essential Learnings Framework 2* which described outcomes and standards for a new Tasmanian curriculum framework. The framework described five standards for 18 'key elements', grouped into five 'essential learnings'. The standards were described as covering the period from 'birth to sixteen years'. The Office for Educational Review calibrated six of the key elements: *Maintaining Wellbeing*, *Being Literate*, *Being Numerate*, *Acting Democratically*, *Being Information Literate*, and *Investigating the Natural and Constructed World*. In each case, calibration involved designing and trialling items that assessed the key-element domain, administering the items to samples of students from Years 2 to 10, and calibrating these items using the Rasch Model (RM). In order to moderate teacher assessments in *Being Literate* and *Being Numerate*, the calibrated scales for these key elements were equated using the RM with the Department's Literacy and Numeracy scales used in its statewide testing program in Years 3, 5, 7 and 9. A process of moderating teacher assessments was implemented by comparing school results on the statewide tests with teacher assessments that had been entered into a centralised reporting system.

1.0 Introduction

1.1 The *Essential Learnings* Framework

In 2000, The Tasmanian Department of Education published *Essential Learnings Framework 1*, which described a curriculum for 'learners from birth to age sixteen' to be implemented in all grades from Prep to Year 10 in Tasmanian Government schools.[2] It listed five 'essential learnings': *Thinking*, *Communicating*, *Personal Futures*, *Social Responsibility*, and *World Futures*. Each essential learning consisted of a number of 'key elements' (18 in all), described as 'organisers for significant ideas within the essential learning'. The five *Essential Learnings* and their respective key elements are listed in Figure 1.

Figure 1: Organisation of 'Essential Learnings' and 'Key Elements' in Tasmanian Curriculum Framework (2000)

Essential Learning	Key Element
Thinking	1. Inquiry 2. Reflective thinking
Communicating	3. Being literate 4. Being numerate 5. Being information literate 6. Being arts literate
Personal Futures	7. Building and maintaining identity and relationships 8. Maintaining wellbeing 9. Being ethical 10. Creating and pursuing goals
Social Responsibility	11. Building social capital 12. Valuing diversity 13. Acting democratically 14. Understanding the past and creating preferred futures
World Futures	15. Investigating the natural and constructed world 16. Understanding systems 17. Designing and evaluating technological solutions 18. Creating sustainable futures

In 2003, the Tasmanian Department of Education published *Essential Learnings Framework 2*. This document contained *Outcomes and Standards for the Key Elements of the Essential Learnings*, described as 'expectations for student achievement from approximately four years of age to sixteen years of age'. The document stated that there were five standards for each of the 18 key elements, and listed, for each standard, some 'illustrative examples of performance', described as 'behaviour that illustrates aspects of learning, which lead to the achievement of the standard'.

As an example, Standard 2 in *Being Numerate* was summarised as 'understands how to purposefully use and explain informal ways of thinking and acting mathematically in familiar situations'. The following illustrative examples of performance were listed:

> *Students demonstrate aspects of this learning when they:*

- *Make and extend patterns, conjecture in simple situations and share their thinking about the methods they used.*

- *Use number concepts and counting strategies (eg count on, count back) to solve number problems.*

- *Use informal measures to describe or compare objects and answer questions such as 'How long?' 'How tall?'*

- *Represent, recognise, group and name common shapes, and describe shapes used in construction activities.*

- *Use personally meaningful ways to represent data.*

In addition, 'performance guidelines' for each key element were listed. The performance guidelines were intended to encompass the 'significant aspects of learning covered by the set of standards' for the key element. Briefly, for *Being Numerate* these consisted of 'understanding how to think, act and communicate mathematically', 'understand number', 'understand measurement', 'understand space' and 'understand data'.[3] In keeping with the other 17 key element outcomes, however, the connection between the performance guidelines and the key element outcomes at any particular standard was not made clear (by, for example, showing the key element outcomes in grid form, with a row for each performance guideline, a column for each standard, and a key element outcome within each row-column intersection).

The decision to set five standards of achievement for each key element was an arbitrary one, and was not based on any research that identified five meaningful standards for each of the proposed key elements. In fact, there was an underlying assumption—not based on any prior research—that there would be a relationship between the 'typical' standard achieved by children and their age and school-grade, shown in Figure 2, although *Essential Learnings Framework 2* (2003: 8) stated that:

> *Standards are not tied to any precise age or grade. Consistent with outcomes-based education, students are expected to achieve each standard fully, although in different ways and at different times.*

Figure 2: Intended Relationship Between Typical Standard Achieved by Students by Age and Grade Level.

Standard	Approximate Years	Approximate Grade
1	2 – 4	end of kindergarten
2	5 – 7	end of year 2
3	8 – 10	end of year 5
4	11 – 13	end of year 8
5	14 – 16	end of year 10

Furthermore, the designers of the curriculum framework did not seem to be aware of the notions of construct validity or latent-trait theory, because some key elements (eg *Being Arts Literate*) were intended to be assessed using a single award, despite the likelihood that they were not latent traits. The two 'thinking' key elements also posed difficulties, as there was much initial discussion as to whether 'thinking' could be assessed without reference to another key element (eg *Being Numerate*, to

effectively assess thinking skills within a numeracy context); and whether or not 'thinking' by itself was a latent trait or, instead, whether 'thinking in numeracy' was different from, say, 'thinking in the Arts'.

2.0 Calibrating The Key Elements of The *Essential Learnings*

2.1 Applicability of the Rasch Model to Calibrating the Curriculum

After the publication of *Essential Learnings Framework 1*, the Department calibrated some key elements of the new curriculum framework, in order to provide a meaningful framework by which students' understandings and achievements could be assessed, to obtain empirical evidence as to whether the stated outcomes were sequenced correctly, and to estimate the relative increment in understanding required by each standard, as described in the curriculum framework.

The assessment demands for the *Essential Learnings* were considerable: unlike previous curriculum reforms in Tasmania that had focussed on particular year levels, the *Essential Learnings* were intended as a framework for students in Prep (pre-Year 1) and earlier (Kindergarten) through to Year 10. In particular, there was a need to ensure that students' assessments would be reasonably consistent across the entire government-school system, so that, for example, a student in Year 3 was not assessed by her teacher at a standard greater than her assessment in Year 7 (assuming she made growth between Years 3 and 7). The limited descriptors of the key element outcomes was a major factor in this respect, as it became clear from initial meetings that teachers in one grade had different understandings of the same key element outcome from teachers of another grade.

The Rasch Model (RM) was used as the basis for calibrating the key elements, since the RM enables item difficulties to be located along a continuum representing the underlying latent trait of the respective key element.[4] More importantly, the RM yields person ability and item difficulty estimates that are invariant: in the RM, the person and item parameters are separated (Rasch, 1960). It is this property of invariant comparisons that distinguishes the RM as a true measurement model.

By asking students questions and rating their answers centrally using trained markers, then applying the RM to the scored results, a measurement scale could be constructed with item difficulties calibrated along the scale. Unlike the use of item facilities in Classical Test Theory (CTT) to estimate item difficulties, the difficulties estimated by the RM have interval-scale properties. Furthermore, the RM enables item difficulties and students' abilities to be estimated on the same scale, so that it is possible to estimate the probability of a student of a given ability being able to answer an item of a given difficulty, or to respond in a specified manner to an open-ended item.

Other advantages of the RM include its capacity to identify items that 'misfit' (and hence may not measure the underlying latent trait—essential in investigating the construct validity of the key element—the relative easiness of equating tests administered to different year groups, and the fact that the RM is used extensively in Australian education departments, particularly in the statewide literacy and numeracy tests that each jurisdiction administers in Years 3, 5, and 7 (and, in some jurisdictions, Year 9 as well). In fact, it was a Departmental requirement that students' results in the Department's Literacy and Numeracy Monitoring (testing) Program be

reported against the *Being Literate* and *Being Numerate Essential Learnings* standards.

To date, the Department has calibrated eight key elements:

- *Maintaining Wellbeing* (2003)
- *Being Literate* (2004)
- *Being Numerate* (2004)
- *Being Information Literate* (2005)
- *Acting Democratically* (2005)
- *Investigating the Natural and Constructed World* (2006)

In addition, there was an attempt to measure aspects of *Thinking Inquiry* with *Maintaining Wellbeing* and *Being Numerate*, but the approach was not applicable to the perceived need to assess *Thinking Inquiry* as a separate and independent latent trait, as suggested by the *Essential Learnings Framework 2*. By 2006, the idea of investigating if *Thinking Inquiry* represented a latent trait was shelved.

The first three calibrations were undertaken by the Office for Educational Review (OER); the last three were contracted to the Australian Council for Educational Research (ACER). Items were written either by practising teachers (*Maintaining Wellbeing*, *Being Literate* and *Being Numerate*) or by staff at ACER (*Being Information Literate*, *Acting Democratically* and *Investigating the Natural and Constructed World*). Nearly all items were 'open-ended', constructed-response questions, and were marked using criterion-based rating scales. Some items involved students navigating through an artificially constructed web site for *Being Information Literate*; others in *Investigating the Natural and Constructed World* were linked to an episode of the television series *Mythbusters*. Most feedback from teachers and students about the items was very positive. Rating scales were initially developed during item writing and were refined following trialling of the items or even when the final calibration tests were marked.

Two consultants—one a measurement expert, the other a curriculum expert—were appointed for the first three calibrations. Curriculum experts were selected for their recognised abilities within the key element to be calibrated, and commented on both the particular aspects of individual items and the overall 'curriculum balance' of the final calibration tests. Items were also panelled (reviewed) by Departmental curriculum and measurement personnel. The measurement expert was consulted on issues such as item fit, selection of 'link items' and equating design, and other aspects of measurement.

A series of tests was then constructed using the pool of items that were accepted by the review panels and the curriculum and measurement experts. Each test was targeted at a year group, although items in each test were selected so that they were able to assess the full range of abilities likely to be encountered in the students who sat them. Calibration tests contained 'link items'—items that appeared in pairs of tests: for example, the same item might have appeared in a test targeted at Year 4 students as well as in a different test, targeted at Year 6 students. In addition, the Year 6 test might contain a different item that also appeared in a test targeted at Year 8 students. Parallel tests, linked by common items, were also used in each year group so that information about a relatively large number of items at each year was available. Using the RM, it was possible to estimate the relative difficulties of all items. Because the RM enables students' abilities on the underlying latent trait to be

estimated as well, the relative abilities of students sitting the tests could also be estimated.

Although it is not a requirement of the RM that items need to be calibrated with random samples of students (the RM only requires that the items be targeted at students so that the students' abilities are roughly commensurate with the items' difficulties), the students who participated in the calibration tasks were selected as randomly as was feasible at the time. (Schools were asked to volunteer, and a pseudo-random sample from these was used for calibration.) This was done to obtain a tentative estimate of the distributions of abilities of students within the year groups tested. This information was used later, to some extent, to demarcate the standards of the key element being calibrated.

Items were marked centrally using check-markers, and results were analysed using *RUMM* [5] and (for the last three key elements to be calibrated) *Quest* [6] and *ConQuest*.[7] Initially, each test was analysed separately for the following:

- targeting of items and persons (ensuring that the item difficulties and the students' abilities, generally, were well matched);
- reliability of the assessment instruments;
- item fit (ensuring that the items fitted the model reasonably, so that each item measured an aspect of the underlying latent trait);
- spread of item difficulties (ensuring that the item difficulties varied in reasonably small increments from 'very easy' to 'very difficult' for the target group, thereby enabling students' abilities to be estimated on the resulting calibrated scale);
- differential item functioning (DIF, or 'item bias', in which students of the same ability had different probabilities of success on particular items according to the demographic group they belonged to – eg 'boy', 'girl', 'Indigenous');
- rating scale design (ensuring that the categories for each rating scale were ordered such that it was more difficult to obtain a 'higher' rating than an 'easier' rating, and that this order was reflected in the category probability curves);
- some additional tests of unidimensionality (eg principal components analysis of residuals) were performed to ensure that serious violations of unidimensionality had not occurred with the calibration items.

Poor-fitting items were discarded from the final scale, and some rating scales were collapsed and the corresponding rating-scale criteria ('rubrics') rewritten so as to reflect the modifications. It was found that many raters were considerably interested in item characteristic and category probability curves, and were, in many cases, able to provide qualitative reasons for the results of the analyses of rating scales.

Finally, all the data was compiled into a matrix so that results were organised by students (rows) and items (columns), with nulls for student-item intersections where students had not been required to attempt the item (eg an intersection for a Year 5 student and a Year 10 item). This was then analysed using the RM, and the items were then located on a common, vertically equated interval scale with an arbitrarily defined origin (the mean of the item difficulties). Students' abilities on the underlying latent trait were also estimated, giving a reasonable idea of the variation in students' abilities by year group for the government system (remembering that the students were initially selected on a pseudo-random basis). Finally, the same checks as mentioned previously were applied.

2.2 Locating the *Essential Learnings* Standards on the Calibrated Scales

After the items had been calibrated, it was necessary to locate the five 'standards' specified by the *Essential Learnings Framework 2* of key element on the calibrated scale. Since the standards had been described, qualitatively, in published documents and there was resistance to markedly changing the qualitative descriptors of the standards, curriculum experts and measurement staff met at a series of meetings in order to compare the calibrated items and the responses to those items with the qualitative descriptions of the standards. By this means, the standards could be identified on the relevant item map. A standard, in this sense, could be thought of as a 'zone' on the calibrated scale, encompassing a range of difficulties (or abilities), with the items and the responses to those items within the standard reflecting the qualitative characteristics of the description of the standard. Because the qualitative descriptors of the standards had not been defined rigorously by the authors of *Essential Learnings Framework 2*, there was some flexibility in identifying the boundaries of each standard. The issue was not, however, entirely governed by the items themselves: because there was data showing the abilities of students in different year groups, some reference to the distribution of person abilities was made in order to identify the standards. For example, the lower boundary of Standard 5 was set so as there would be at least some students who would be expected to achieve at that standard.

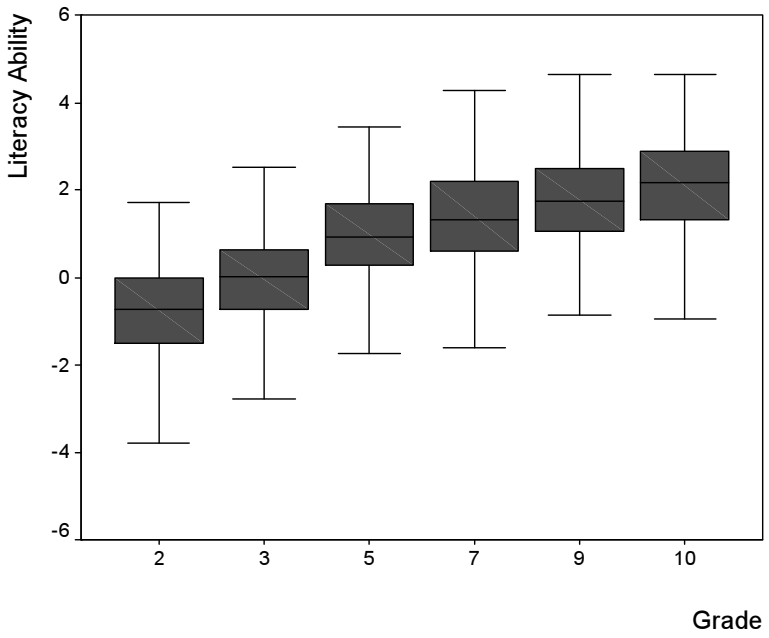
Another consideration was the 'width' of each standard. Standards that encompassed too small a range of difficulties (or abilities) would become confounded with the error associated with person abilities, so each standard initially was set to span approximately 1.5 logits. Given that previous experience with literacy and numeracy testing suggested that students, on average, gained approximately 0.5 logit in ability each year (at least in the primary years), the 1.5 logit width was roughly commensurate with the intended age range of three years 'growth' for each standard as specified in *Essential Learnings Framework 2*.

Some modifications to curriculum documents were made as a result of the calibrations. For example, the original key element outcomes for *Maintaining Wellbeing* suggested that children at Standard 1 were not capable of considering the wellbeing of others, and that their focus was entirely on their own wellbeing. Results from calibration, however, clearly demonstrated that children at this level were capable of considering the wellbeing of significant others (eg siblings and close friends). The key element outcomes for *Maintaining Wellbeing* were altered accordingly. These examples demonstrate the usefulness of empirical research and the RM in curriculum design, as it is essential that teachers match their teaching to the abilities of students, and not to preconceived ideas of the range of abilities of students at particular year levels.

Another result that confounded the locating of standards on the calibrated scales was the non-linear growth in ability that was found for all calibrated key elements. Figure 3 shows the estimated abilities of students by year group (or grade) from the calibration of *Being Literate*. It should be noted that the students who participated in the calibration study were not necessarily a random sample of their year group, and that the extreme abilities shown on each box and whisker were subject to greater error than those represented by the 'boxes'. Also, since all large quantities of content were supplied in items, the underlying latent trait reflected one of 'understanding' rather than 'knowledge and understanding', which might also partly explain the extensive overlap between year groups. Furthermore, the figure shows cross

sectional, not longitudinal, data. Nevertheless, the non-linear growth in estimated abilities from year group to year group is apparent, with a 'flattening' of growth of ability at about year seven and continuing through to year 10. A similar pattern was observed by Rowe and Hill (1996: 335) in studying students' progress on the *English Profiles Reading Strand*.

Figure 3: Estimated Ability By Year Group: *Being Literate* Calibration Sample



Note: the scale shown on the ordinate is in logits. The zero is arbitrary; the scale is interval.

The non-linear growth in *Being Literate* ability posed problems for setting the standards on the *Being Literate* calibration scale, as there seemed to be an expectation on the part of the curriculum designers that growth in student ability would be linear whereas in reality it isn't. This aspect was common to all six of the key elements calibrated so far. In fact, in some key elements (eg *Maintaining Wellbeing*) there was very little measured growth in ability between primary and secondary students, possibly because of the definition of the underlying latent trait.

Given that there was an expectation, from a system perspective, that students would at least appear to make progress through the standards as they proceeded through the education system—reporting that a student remained in Standard 4, say, in grades 8 and 9 would appear to show no progress in those years—each standard was then arbitrarily divided into three 'progression levels'. Each progression level was approximately 0.5 logit. Thus, for example, Standard 4 was divided into the following progression levels: '4 lower' (4L), '4 middle' (4M) and '4 upper' (4U). There was some tension between the need for extra qualitative descriptors (standards and progression levels) and the consequent narrowing of the width, or span, of each on the calibrated scale and associated problems with measurement error. The constructors of the calibrated scales, however, were forced to operate within these difficulties as it had already been decided by the Department that there would be five

standards, each divided into three progression levels. In addition, the standards and progression levels had to be achievable by students, so it was not an option to locate Standard 5, say, so 'high up' on the scale that few if any students would be expected to attain it.

In retrospect, it would probably have been preferable to calibrate each scale first, then divide the resulting scale into meaningful standards and progression levels, and finally describe each standard and level without the overarching constraint of there always being five standards and a total of fifteen progression levels.

Each progression level for each standard, for each key element calibrated, was described briefly and the descriptions ('progression statements') were made available on the Department's internet site.[8] These descriptions, however, are not sufficiently detailed for teachers and parents to assess a student's work at a particular progression level, or even standard, using the progression statements alone. Support materials were also written in order to overcome this difficulty.[9]

In recent months there has been a continuing drive for the progression statements and support materials to better reflect the findings from the calibration studies.

2.3 Equating Being Literate and Being Numerate with the Monitoring-Test Scales

The Department of Education (Tasmania) has tested students in government primary and secondary schools in aspects of literacy and numeracy since 1975.[10] From 2001, the Department used the Western Australian Literacy and Numeracy Assessment (WALNA) tests in Years 3, 5 and 7. These tests—like all other full-cohort literacy and numeracy tests in other Australian jurisdictions—are based on the RM and are equated horizontally from year to year and equated vertically for Years 3, 5 and 7. This has the effect that the resulting scale scores[11] are comparable (within error) from chronological year to chronological year and from grade (year group) to grade. Thus, a reading scale score of 500, say, for a student in Year 5 in 2005 would be comparable with the same scale score of 500 for a student in Year 3 in 2006, because the scale score of 500 for reading suggests similar reading ability, regardless of chronological year or grade.

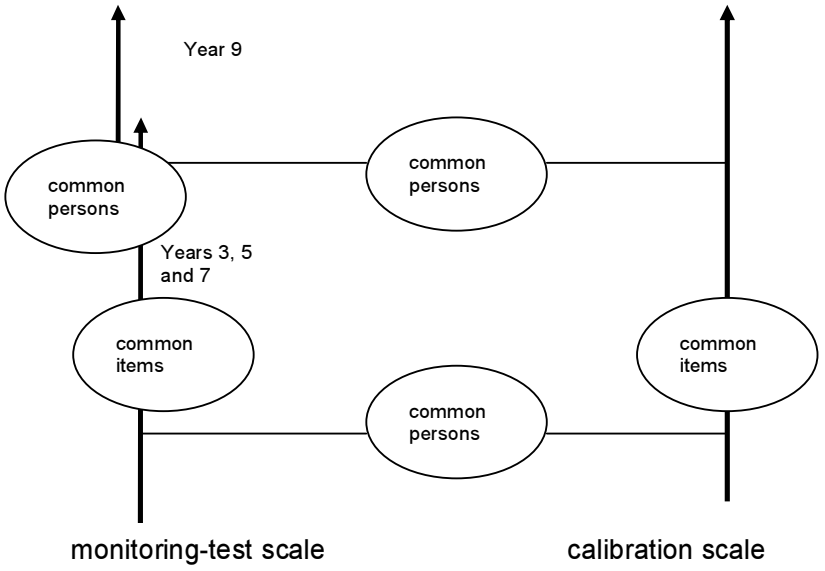
In 2004, about 300 Year 9 students in Tasmania sat the 2004 Year 7 tests in reading, writing and numeracy; and a similar number of Year 7 students sat the Year 9 tests. This enabled the two test scales to be equated using the RM (using 'common-person equating). Then, calibrated scales for aspects of literacy and numeracy were constructed, and new scale scores derived. Essentially, the resulting monitoring-test scale scores were similar to the calibrated scales for *Being Literate* and *Being Numerate* but their origins were different and they were each constructed from different items.

Next, the monitoring-test scales and the *Being Literate* and *Being Numerate* calibration scales were equated. This was possible because some 300 students in each of Years 3, 5, 7 and 9 sat both the 2004 monitoring test and the *Being Literate* and *Being Numerate* calibration tasks, and the scales could therefore be equated through common-person equating. One reason for this equating was the Departmental requirement that all monitoring-test results from 2005 onwards be reported against the standards and progression statements of the relevant key

elements in the *Essential Learnings*. An obvious key issue concerned whether or not the two tests (the monitoring test in numeracy, say, and the *Being Numerate* calibration tests) measured the same latent trait. Evidence suggested that, at least, one was a strong predictor of the other, as the results for the calibration tests and their respective monitoring-tests counterparts were strongly correlated[12]. Also, when the distributions of abilities of students in a particular year group for both the monitoring tests and the appropriate calibration key elements were compared, the distributions were similar, suggesting that the pseudo-random samples used in the calibration were, in fact, reasonably random.

The ease of equating tests using the RM therefore enabled vertically equated Year 3 to 9 monitoring results to be equated to Year 2 to 10 calibration results. This in turn enabled monitoring-test results to be reported against the *Essential Learnings* Framework. In fact, without the RM it is unlikely that this equating could have been achieved, since some other approaches (eg equipercentile equating) rely on distributions of abilities, and not all distributions (eg for the cohort for a calibrated key element) were known fully. Figure 4 is an overview of the equating procedures.

Figure 4: Equating the Monitoring Test and Calibration Scales



3.0 Moderating Teacher Assessments

3.1 Approaches to Moderation

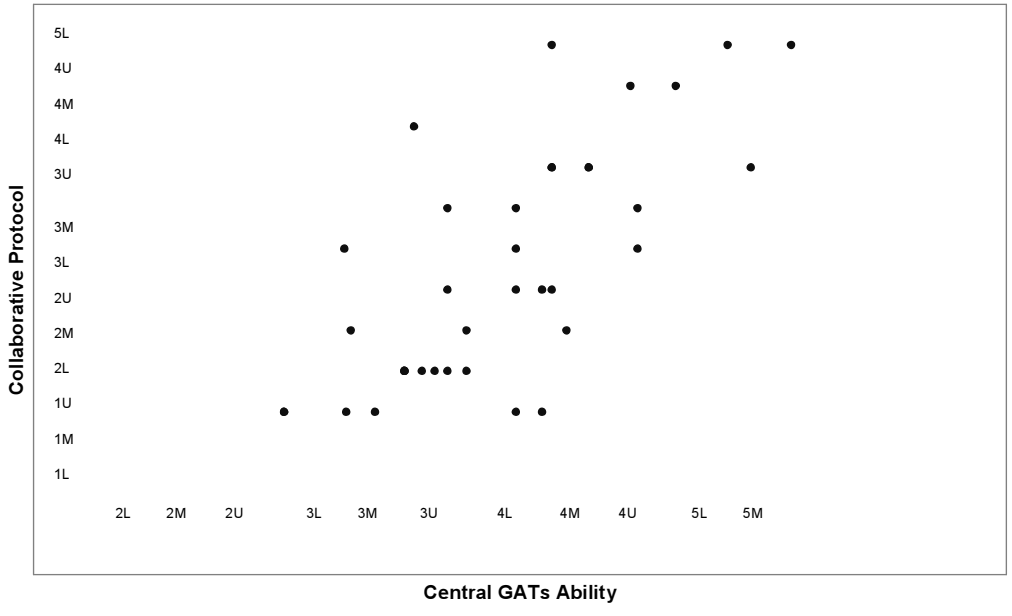
A requirement of the implementation of the *Essential Learnings* was that teachers would assess their students' performance against the standards and progression levels of (initially three) key elements of the *Essential Learnings*. Experience showed, though, that assessments made by teachers would need to be moderated in order to improve inter-teacher reliability. One moderation model that was proposed and

implemented was 'consensus moderation', in which teachers collectively discuss and assess samples of a student's work, in order to reach a common assessment award.

A major problem with this approach was that even with progression standards and work samples, there was frequently considerable variation between groups used in consensus moderation, because each essentially interpreted the progression statements and work samples slightly differently. An effect of this is shown in Figure 5, which is a scatterplot of assessments against *Essential Learnings* progression levels of a number of work samples made by expert consensus-moderation groups against assessments made by central raters. It can be seen that there was considerable variation in assessment awards and that the differences in awards between the two groups for the same work sample was often considerable. Other evidence suggested that consensus moderation would be unlikely to produce consistent and reasonable results across all schools for all year-group levels.

Because of these variations, a method of 'statistical moderation' was developed. This involved comparing the results of students, grouped by school, on a centrally administered and rated test with the group's teacher-awarded results. Essentially, the approach involved determining whether the 'ballpark' of teacher-awarded results was similar to the 'ballpark' of centrally awarded results that were based on a common, centrally administered task that was, in turn, equated to the relevant calibration scale.

Figure 5: Awards by Consensus Moderation Groups (Ordinate) Against 'Central' Awards (Abscissa)



3.2 'Central' Moderation Tasks

The Department's Literacy and Numeracy tests were used as moderating instruments in Years 3, 5, 7 and 9 for *Being Literate* and *Being Numerate*. This was possible because the results were reported against the *Essential Learnings* standards and progression levels—this was made possible because of the monitoring-test scales had been equated to their respective calibration scales.

In 2005, teachers had to report against the *Essential Learnings* standards for *Maintaining Wellbeing*. Since this key element had already been calibrated, the RM enabled new tests to be constructed and equated to the calibrated scales. Accordingly, two tests were developed, trialled, and administered to all Year 6 and Year 10 students in government schools. Each test was linked to the other with common items, and was equated to the calibrated *Maintaining Wellbeing* scale using common persons (drawn from students from the independent sector). The two *Maintaining Wellbeing* tests were called 'Guiding Assessment Tasks' (GATs).

3.3 Moderating Teacher Assessments

Because teacher-based assessments, from Prep to year 10, for all Government-school students, for the three key elements to be assessed in 2005 against the new curriculum framework, were entered (by teachers) into the Department's central Student Assessment and Reporting Information System (SARIS),[13] each student's results for the centralised assessment tests (the monitoring tests, for literacy and numeracy and the *Maintaining Wellbeing* GATs) could be matched centrally in a matter of a few minutes (by means of the Department's unique student identifier). Because the monitoring-test scales and the GATs had been equated with their respective calibration scales, centrally based assessments for *Maintaining Wellbeing*, literacy and numeracy could be compared directly with teacher-based assessments (because all were made against the standards and progression levels).

For year groups in which 'actual' results for both central and teacher-based assessments existed (Years 3, 5, 7 and 9 for *Being Literate* and *Being Numerate*, and Years 6 and 10 for *Maintaining Wellbeing*), it was possible to make a reasoned judgement as to whether the teacher-based assessments, for each school treated as a group, were comparable with the centrally based assessments, also treated as a group. Measures of central tendency and dispersion were used to make this judgement. Measures of dispersion were necessary, because it was found that teachers tended to 'clump' their assessment awards towards the average award, and the range of awards given was much less than the ranges of awards for the school as measured by the central assessments, even allowing for some regression-to-the-mean effects that would be expected from assessing students on multiple occasions. (The actual spread of centrally measured awards was reduced by a constant in order to model this effect.)

Although a variety of approaches were possible, those used in 2005 involved using F-tests and matched-pair t-tests[14] to determine if the two groups of results were significantly different from each other. An inevitable problem occurred with levels of significance and sample size, so an arbitrary difference was used to flag schools that might have over- or under-estimated their students' performances, in addition to carrying out the tests. It is likely that this process will be refined in future (perhaps by using non-parametric tests).

Schools that had assessed their students, as a group, markedly differently from the centralised assessments were contacted and given feedback. They were then invited

to re-assess their students and to re-enter the results into SARIS. The entire procedure was voluntary and no overt coercion was employed, although most schools noted the feedback.

In addition, results for some year groups were 'modelled' from actual data, by interpolating (and sometimes extrapolating). For example, a hypothetical distribution of awards for Year 4, say, could be generated for a school by interpolating the mean and standard deviation from data that existed for Years 3 and 5. Another modelling approach, used in 2006, was to obtain students' actual SARIS results for 2005 and to increment them by the average rate of growth obtained for that cohort and year group. Both approaches, however, depend on assumptions that may or may not be true. While these procedures are somewhat speculative and subject to cohort effects, it was found to be helpful to schools by sometimes making teachers reconsider their awards in the light of awards in their schools for adjacent year groups. Again, no coercion was used to force schools to change their awards.

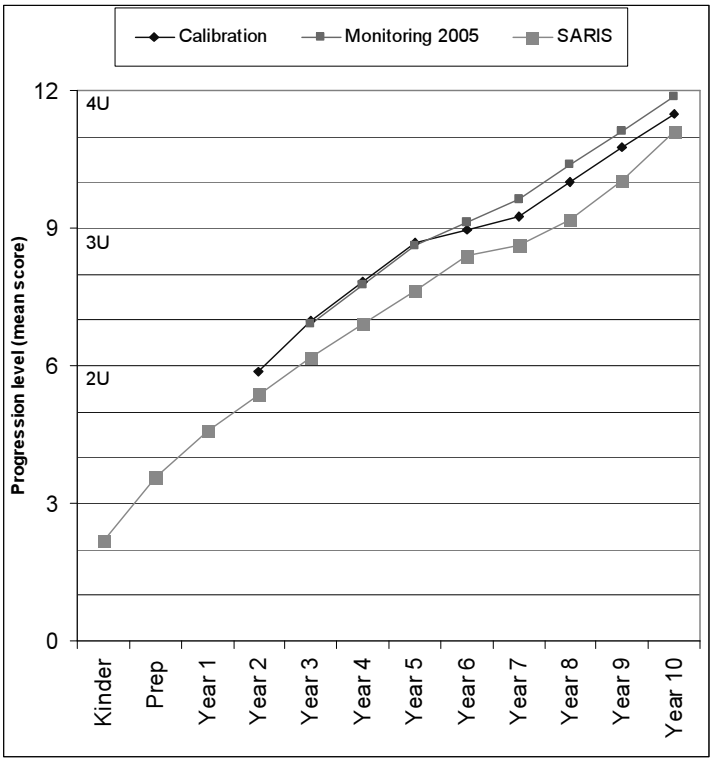
Figure 6 shows the mean award for students in government schools in 2005 for calibration (based on a pseudo-random sample), monitoring (based on full-cohort results for Years 3, 5, 7 and 9, and with modelled results for the other year groups), and teacher-awarded results entered in SARIS (some 60 000 in all), for *Being Numerate*. The dispersion of awards for each year is not shown. Results suggest that, generally, the teacher-based awards were approximately one progression level less than the awards based on monitoring, which were similar to the results from calibration. This can be explained partly by the general instructions given to teachers in professional development and consensus moderation meetings to 'rate on the hard side', and also for many high-school teachers to rate students in Years 7 and 8 similarly, perhaps to accentuate growth in achievement in high school, given that monitoring and calibration results suggested that growth was less, on average, in high school than in primary school. In fact, initial, un-moderated teacher assessments resulted in a 'dip' in performance between Years 6 and 7, possibly owing to this effect.

Another factor that probably resulted in the mean SARIS (teacher) award being less than expected was that some schools ignored advice about their awards and refused to alter them. Nevertheless, there was reasonable agreement between SARIS and centrally administered results, on a school-by-school, basis to argue that the moderation process had been generally effective, especially considering that the process was only in its first year of implementation.

Another effect noted was the tendency for teachers to 'bunch' their awards around the mean award given. This may have important educational implications in that many teachers may be unaware of the actual range of abilities of students within their classes, perhaps resulting in poor targeting of learning experiences to students.

Although the average SARIS award by year level is less than the average centrally-assessed award, the distribution of awards by year group for the whole Government-school population shows an overall consistency that was not present before moderation advice was given to schools.

Figure 6: *Being Numerate* Awards from Calibration, Monitoring and SARIS: 2005



4.0 Conclusion

The processes used by the Office for Educational Review in 2005 highlighted the need for properly calibrated measurement scales to underpin curriculum frameworks, particularly when teachers are expected to assess students' performances against these frameworks. Calibration (using the RM) not only highlighted errors with some aspects of the new framework, but also enabled measurement scales for whole-cohort assessments to be equated to calibration scales, and thereby facilitated moderation procedures. (Far from being seen as some desirable but esoteric criterion, moderation should be viewed as a component of equity in assessment: students have the right to be assessed on comparable criteria regardless of which school they attend or who assesses them.) Furthermore, results from calibration and centrally administered whole-cohort tasks highlighted the spread of student abilities within year groups, and the need for teachers to be aware of this so that they target learning experiences appropriate to their students' ranges of ability.

All these procedures ultimately hinged on the RM: without it, very little, if anything described, would have been achieved.

References

Department of Education Tasmania. (2003) *Essential Learnings Framework 2: Introduction to the Outcomes and Standards*. Hobart, Tasmania

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press.

Rowe, K.J. and Hill, P.W. (1996) 'Assessing, recording and reporting students' educational progress: The case for 'Subject Profiles'', *Assessment in Education* 3 (3): 309-352

End Notes

[1] Andrew Smith, Office for Educational Review, Department of Education, Floor 2, 99 Bathurst Street, Hobart, Tasmania 7000. Email: Andrew.Smith@education.tas.gov.au

[2] Information about the Essential Learnings can be found at: http://www.education.tas.gov.au/school/educators/curriculum/elscurriculum

In 2006 the Minister for Education announced a refinement of the Essential Learnings and referred to the curriculum framework as 'The Tasmanian Curriculum' (see: http://www.education.tas.gov.au/dept/about/minister_for_education/curriculumupdateparents).

Briefly, the key elements of the Tasmanian Curriculum were announced as:

- English/Literacy
- Mathematics/Numeracy
- Science and Technology
- Information and Communications Technology (ICT)
- Society and History
- Arts
- Personal Development

[3] The key element outcomes for all 18 key elements can be accessed at: http://www.education.tas.gov.au/school/educators/curriculum/elscurriculum/OutcomesandStandards.doc (accessed September 2006)

[4] A full discussion of the RM is not possible here. A general description by Stephen Humphry can be found at Wikpedia (http://en.wikipedia.org/wiki/Rasch_model). A comprehensive explanation of the dichotomous RM is:

Andrich, D. (1988). *Rasch Models for Measurement*. Sage Publications. Beverly Hills, CA.

[5] Andrich, D. Sheridan, B. and Luo, G. (1997-2005). *Rasch Unidimensional Measurement Models* (software), RUMM Laboratories. Perth, Western Australia.

[6] Adams, R. and Khoo S. (1993). *Quest: The Interactive Test Analysis System* (software). Australian Council for Educational Research (ACER). Melbourne, Victoria.

[7] Wu, M., Adams, R., and Wilson, M. (1998) *ConQuest: Generalised Item Response Modelling Software*. Australian Council for Educational Research (ACER). Melbourne, Victoria.

[8] http://www.ltag.education.tas.gov.au/assessment/outcomes/progstate.htm (accessed September 2006)

[9] http://www.ltag.education.tas.gov.au/references.htm#assessing (accessed September 2006)

[10] From 1975 to 1994, the Education Department in Tasmania tested 10- and 14-year old students in numeracy and reading. From 1996, tests were directed at year groups (Years 3, 5, 7 and 9). For some years, the Department also tested students in schools belonging to the Catholic Education Office.

[11] A scale score can be derived by multiplying the logit (of an item difficulty or a person ability) by a constant and adding another constant. Since the transformation is linear, the interval-scale property of the RM is preserved.

[12] Typically $r > 0.8$ after correcting for attenuation, but further research is needed to investigate this for literacy.

[13] See http://wwwfp.education.tas.gov.au/oer/SARIS/default.htm (accessed September 2006).

[14] This was achieved by 'scoring' 1L (standard 1, lower) as 1, 1M (standard 1, middle) as 2 and so on, remembering that each progression level represented approximately the same range of difficulty. The assumptions of normality and equal variances were not tested rigorously, though in the moderated results, schools responded to suggestions regarding 'spreading' their awards based on the spread of awards from monitoring or GATs.