A report to the Curriculum Council of Western Australia regarding assessment for tertiary selection

David Andrich
Murdoch University

Acknowledgements

Case Study 1 has been provided by Sandy Heldsinger and Stephen Humphry with permission from the Department of Education and Training. This study has arisen from their sustained and comprehensive studies in assessment in general and in writing in particular. The data for Case Study 2 were provided by the Curriculum Council.

A report to the Curriculum Council of Western Australia regarding assessment

The terms of reference

To prepare a report and advice on the comparability of standards for the new courses.

The aim will be to ensure that:

- the assessment process of each course has sufficient rigour to enable the highest academic standards to be maintained;
- assessment is such that the fine grained measurement of student achievement is valid and reliable particularly where university entrance is involved;
- the measurement processes being developed will enable comparability of standards between courses and enable statistical adjustments to be made if necessary.

These terms of reference are taken from correspondence with the then CEO of the Curriculum Council, Mrs Norma Jeffery (Appendix).

Summary and abstract

The key recommendation in this report is that for both school based and external assessments, analytic marking of the traditional kind using marking keys that arise directly out of the assessment tasks, be used for student assessment for each unit of a course, and for each course as a whole at the end of Year 12. A related recommendation is that, simultaneously, a rating of student performance into one of eight generic levels of achievement that arises out of the outcome statements be used as part of the assessment. The former provides marks for the assessment and measurement of students at a relatively micro level suitable for feedback to students and for use in tertiary selection according to the policies of the Curriculum Council. The latter provides ratings for classification at a relatively macro level suitable for monitoring the general progress of students and the operation of a course and is commensurate with the generic nature of the level and outcome statements. The two assessment processes, distinguished by their level of precision and relevance, are compatible and can be combined and integrated. By taking advantage of this complementarity, the Curriculum Council can genuinely advance the communication of educational achievement in Western Australia.

A report to the Curriculum Council of Western Australia regarding assessment

Overview of approach to the report

Three particular features of the approach taken to address the terms of reference are presented at the outset. First, this report is concerned with *principles* of assessment, and recommendations which follow from these principles, rather than detail in the assessment of each particular course. Two case studies in assessment are, however, presented in different amounts of detail to illustrate these principles and to show the rationale for some of the recommendations. Second, it is concerned with assessments that would meet the requirements of being sufficiently rigorous and sufficiently *fine grained* that they can be used for equitable selection into tertiary programs of study. Third, it is indicated and assumed in the report that if two major policies of the Curriculum Council, first combining school based and external assessments, and second combining different courses to form a Tertiary Entrance Rank (TER) are to be implemented correctly, then the following constraints are imposed on the assessments: (a) the school based and external assessments must have the same order of precision and (b) the measurements derived from them must be on the same scale.

The analysis of assessment is considered in the context of *outcomes based education* (OBE) and its reforms for post compulsory education as articulated by the Curriculum Council (1998). The report is neither a criticism nor an endorsement of OBE in general – instead it provides recommendations with a rationale for the kinds of assessments that should meet the requirements of rigour and precision for use in tertiary selection. In addition, because assessment affects teaching and learning, these recommendations are intended to be compatible with sound teaching and learning practices.

It is assumed that OBE is a much wider set of educational principles than a set of assessment practices, and in particular, that it is not characterised by the nature of assessments that need to be used for competitive tertiary selection. Such assessments need only be compatible with the OBE principles in meeting the requirements of rigour and precision required for tertiary selection.

Some terminology

In this report, the terms *assessment* and *measurement* will both be used in a way compatible with that implied in the terms of reference.

Specifically, the term *assessment* will emphasise the stage of design, administration and marking of performances elicited by the tasks. These performances are marked against criteria made explicit through marking keys, sometimes referred to as *rubrics*. Designing assessments and marking keys is present in all assessments. It is the common form of assessment. It is assumed that they arise directly from the teaching and learning that it is expected to have taken place in a Course, which in turn should characterise the required educational outcomes. It is this connection between the assessment tasks and the outcomes that gives the former validity.

The term *measurement* will emphasise the stage of scoring of assessments into a numerical form and *transforming* them into a quantitative scale using statistical models. This second stage is used in exactly those situations where it is necessary to ensure that the assessments across components of courses or across courses are on a *commensurate* scale. The concept of commensurate scales and their application is elaborated in a later section.

The terms of reference refer to assessments which are sufficiently *fine grained*. As implied above, in this report, the term *precision* will refer to the level to which the measurements are *fine grained*.

University entrance in Western Australia involves forming a Tertiary Entrance Score (TES) based on combining marks on school and external assessments for each Course, and then combining the marks across courses. Students are ranked on the basis of this second set of marks into a Tertiary Entrance Rank (TER). Complementary requirements for particular tertiary programs of study, such as prerequisites may be imposed. General reference in the report is made to the TER on the understanding that to create a TER, a TES has to be formed first. It is the TES that needs to have the required precision in forming the TER.

Competitive entry and its implications

The terms of reference refer specifically to university entrance. The implied concern is not simply with university entrance per se, but with selection that is competitive. I have written elsewhere on the impact that competitive entry has on education and the constraints it imposes on possible selection processes (Andrich, Rowley, & van Schoubroeck, Andrich and Mercer, 1997). Effects of competition, generally distorting some ideal aspects of education, are inevitable when more students apply for entry to various programs of study than there are positions available. Transparent and publicly defensible processes that minimise these distortions are then particularly necessary in a public system. Four aspects of this competition are highlighted as providing a backdrop for the need for measurement to be precise enough for competitive selection.

First, the references to *university* and *tertiary entry* need to be understood to imply entry into particular tertiary programs of study, and not to tertiary institutions as a whole. For example, in universities, the entry is referenced to Commerce, Arts, Physiotherapy, Education, Medicine, Law, Veterinary Studies, and so on. In the technical and further education (TAFE) sector, entrance is likewise according to programs of study, not to TAFE in general.

Second, although some programs have a very high competitive profile, competition is present for individual students in many other programs in both the university and the TAFE sectors. Thus students compete for offers in commerce, in science, in education, and in other courses, and for individual students on the margin of these offers, the competition for them can be as fierce as it is for students on the margin for selection into the high profile courses such as medicine and law. Therefore, as expressed in Curriculum Council documents, the selection process must be consistent

and fair across the spectrum of achievement, and not cater only for the high profile competitive courses.

Third, the selection process is sufficiently significant that if it is not accounted for credibly within the schooling framework, and tertiary institutions decide to initiate an independent selection process, then that selection process will inevitably impact even more than the present process does on the teaching and learning in post compulsory education.

Fourth, it is required that different courses continue to be equivalent in difficulty for obtaining scores for competitive tertiary entry. This has implications for assessment and measurement of student performance.

Finally, this report is written from the perspective that the credibility in the following two Curriculum Council policies already mentioned, (i) the combining of school based and external assessment within a Course, and (ii) the combining of marks across courses to form a single TER, is paramount.

Credibility and transparency are important in this process. As will be explained in more detail in the report, this involves measurement principles and scaling to ensure that the measurements to be combined are on the same scale. This is required no matter which educational principles are used to organise the teaching and learning.

The understanding of these principles needs to be placed in context, and one of the best ways of understanding this context is to consider alternatives, for example to consider how other systems, similar and different in various ways, attempt to solve the same key problem, that of competitive entry. For example, in England, all assessments are external; in the US where the school based assessment is different in different schools with little if any standardisation, the Scholastic Achievement Test (SAT)[1] dominates selection; and in Continental Europe, there may be specific assessments for entry into specific programs of study in specific institutions which are uncoordinated.

> ***Recommendation 1*** *That professional development be provided by the Curriculum Council to relevant education personnel and to Principals regarding the broader context of location of Year 11 and 12 Study, the constraints imposed by competitive tertiary selection into particular courses, the advantages and disadvantages of the process implemented in responding to these constraints, and potential alternatives with their advantages and disadvantage as exemplified in other countries or other states.*

[1] The SAT is three hours and 45 minutes long and measures skills in three areas: critical reading, math, and writing. Although most questions are multiple choice, students are also required to write a 25-minute essay.

Summary of key principles of OBE relevant for considering assessment

The context of OBE from which this report written is summarised in this section. Greater detail of the interpretation of the OBE in a historical context and with broader implications for assessment are reported in Andrich (2002a, 2002b). Four essential features summarised here are (i) the structuring of the School Curriculum into Eight learning areas, (ii) the articulation of outcomes within each of these areas, (iii) the identification of aspects of these outcomes, and (iv) the specification of 8 levels of achievement for each outcome.

Because of its level of generality, it is argued that OBE can provide a broad frame of reference for organising teaching and learning but that it is too broad for making direct assessments precise enough for tertiary selection. Clearly, the assessment will need to be compatible with the teaching and learning that takes place in schools and arise from the OBE framework, but need not be determined by it in any one particular way.

The 8 learning areas are shown in Table 1.

Table 1
The eight learning areas

The Arts;	Mathematics;
English;	Science;
Health and Physical Education;	Studies of Society and Environment;
Languages other than English;	Technology and Enterprise.

Most courses have only four outcomes, though some have as few as two. This inevitably makes them general and abstract. Illustratively, the English course has four outcomes (listening and speaking; viewing; reading; and writing). These are clearly generic and overlap in practice considerably. Again, each of these outcomes is elaborated. For example, the outcome Writing is documented and elaborated as shown in Table 2.

The Writing outcome of English

Table 2
The writing outcome

Outcome 4: Writing

Students write for a range of purposes and in a range of forms using conventions appropriate to audience, purpose and context.

In achieving this outcome, students:

- use the conventions of written texts with increasing understanding and critical awareness;
- demonstrate increasing critical awareness of the ways language varies according to context and how language affects the ways students view themselves and their world; and
- select with increasing effectiveness from a repertoire of processes and strategies when writing by reflecting on their understanding of the way language works.

This outcome is then divided explicitly into 8 levels on an achievement continuum. Levels 5 to 7 of the Writing Outcome are reproduced in Table 3. This outcome has been elaborated because it provides the basis for the first case study which illustrates some important issues in assessment in the context of specifying levels a priori.

Table 3
Levels 5, 6 and 7 of the Writing outcome on the scale of achievement

Level 7	Students write sustained, complex texts, controlling conventions to engage with readers in different contexts: critically appraise and review their own writing and that of others, reflecting on the processes and strategies for improving their own writing.
Level 6	Students write with a clear sense of purpose and structure, exploring different perspectives, experimenting with language conventions and varying their expression to enhance effect and to meet the expectations of different audiences; and use appropriate strategies to evaluate and review their writing.
Level 5	Student explore challenging ideas and issues in a variety of text types; select language to suit specific audiences, purposes and contexts and adapt language structures and conventions necessary for clear communication; and apply a range of planning and reviewing strategies to shape writing.

Clearly, the levels, as the outcomes, are abstract and general, and given that they span the 12 years of schooling, each can cover a broad range of achievement which can take a considerable time in school to learn. The implication of the generality of the outcomes and of the levels statements is central to the key points of the report and to the recommendations.

Giving a further structure to the organisation of the courses, for Years 11 and 12 each has 6 units, termed 1A, 1B, 2A, 2B, and 3A and 3B which are targeted at sequentially higher levels. These are semester units. This structure clearly has implications for assessment, both conceptually and practically. Students requiring a TER need to complete at least two units, though most will complete four, one each of a semester in each of Years 11 and 12. In particular, mainstream students in a course will tend to complete units 2A, 2B, 3A and 3B. This flexibility for students has major resource implications for assessment.

The Arts Learning Area

The second case study arises from assessment in The Arts, and in particular within the content of Drama. It is not presented in the same detail as the first one; it is presented primarily to illustrate points beyond those that could be illustrated with the first case study.

The Arts learning area also has four outcomes (arts ideas; arts skills and processes; arts responses, arts in society), each of which is elaborated and also divided into eight levels of increasing levels of achievement. It is once again evident that these outcomes, covering such a wide range of content, must be abstract and general.

There is a further organisational structure – first, each of these outcomes is referenced to the five content areas of Dance, Drama, Media, Music and Visual Arts, and second, each outcome is referenced to Creating, Interpreting, Exploring, Developing, and Presenting. Further, there is a level description for each of these combinations for each of the outcomes.

Further structure to the courses

As indicated already, a common feature of the structure of courses and outcomes is that there are 8 levels of achievement for each outcome.

From an assessment perspective, teachers and examiners will have evidence in a performance (e.g. a written essay, a solution to a mathematics problem, a response to an arts stimulus) on more than one outcome.

The general task is to assess such performances with respect to the outcomes in a reliable, valid, and sufficiently precise way across Years 11 and 12 that the resulting measurements can be used for selecting students in a range of competitive programs of study at tertiary institutions. Within this structure, and with some elaborations from the Curriculum Council, this will involve four stages of aggregation:

(i) within each unit, school assessments across outcomes;
(ii) school assessments across units within each course;
(iii) school and external assessments for each course; and

(iv) assessments across courses that will produce a TES and a TER.

These aggregations are reconsidered in more detail throughout the report.

Regarding (i) above, it is proposed that this be obtained from two formal assessments within the school. This particular proposal is also addressed in the report.

Some expectations across learning areas

The above summary pertains to the structure of each particular course. Further, the standards of levels *across* outcomes within a particular course are expected to be the same, so for example Level 4 in Writing of English is to be of the same standard as Level 4 Reading. A standard refers to intellectual, and where relevant, practical demands of the students. *Although reasonable to propose at the level of course design and teaching, it is extremely unlikely to be able to define equal standards a priori in the assessment stage from very generic descriptors at the level of precision required for tertiary selection.* Indeed it is an empirical matter whether or not any particular set of assessments show specific equivalences.

An even more demanding expectation from the Curriculum Framework is that the levels *across* courses are to be of the same standard. Thus the standard in Level 4 English should be the same as the standard in Level 4 mathematics, and so on. *Again, although a reasonable framework to have at the level of course design, which is indeed already present as evidenced by the capacity to form a TES and a TER, it cannot be assumed that this equivalence can hold with any particular assessments without empirical checking.*

The intention of this expectation is that the courses meet the requirement that they be equally difficult for students to obtain similar scores for tertiary entry. Because standards are difficult to decree a priori at the level of precision of assessment required for tertiary entry, this requirement of equal standards has implications for the operational stages of assessment and measurement. Substantial reference to this demand is made later in the report.

The context of the requirement that the same standard holds across the same levels of different outcomes within a course and across different courses, needs to be appreciated. The context is, it will be recalled, that there is a *similar* number of outcomes in all of the courses, all of which are divided into *exactly* the same number of levels of achievement. This degree of similarity in the number of outcomes and the identity of the number of levels confirms that they must be at a substantially general and abstract level that do not arise out of the content and processes of any of the courses. Therefore, it follows that they can only be useful as organising principles *governing* the teaching, learning and assessment at the same general level, and not for *determining* assessment at a much finer level of precision necessary for tertiary selection, and even for student feedback.

The learning areas and courses do not in themselves have any particular reasons for being in exactly 8 levels, nor to have outcomes that are from 2 to 4 in number, with a majority having 4. The reasons for imposing these restraints must be administrative and organisational and they can be useful in these terms. However, they are not

inherent to the courses. Furthermore, because they are organisational and administrative structures at a very general and abstract level, they cannot determine every form of assessment for every purpose.

Fragmentation

In Andrich (2002a), a paper published based on work commissioned by the Curriculum Council, I make connections, comparisons and contrasts between Bloom's Taxonomy of Educational Objectives (Bloom, et al, 1956) and Outcome Based Education as then to be implemented by the Curriculum Council. In that paper, I argue that the former was in many ways a forerunner of the latter. I make the point that the Taxonomy became ineffective, after excellent initial work and despite a warning from the original authors, because of *fragmentation*. One symptom of this fragmentation was the popular change in terminology from the original as The Taxonomy of *Educational* Objectives to one of concerns with *Behavioural* objectives. The fragmentation led to endless and unmanageable checklists for assessment of students rather than for the organising of teaching, learning and assessment. I also warned that because the OBE movement has many similarities to the development of the Taxonomy, that it was similarly prone to fragmentation.

At present it is proposed that there be six discrete units in a course, of which a typical student would take four, and must take at least two before being able to sit the external examination and obtain a TER. Within each unit it is proposed that four outcomes be assessed from two formal assessments and that these be averaged on an outcome basis and be submitted to the Curriculum Council for each unit.

For four units of a typical student within one course, this indicates that 16 scores of school based assessment, derived from 32 formal assessments, will contribute towards a TER. At present, there is only one such score required.

Moreover, by the end of first semester of Year 11, a student studying six courses will require a school to submit 24 scores. I consider this amount of formal assessments to be carried out and submitted to the Curriculum Council reflects the original potential danger I warned against – that of fragmentation leading to over assessment. The number of scores submitted and the prescription for the assessment are excessive, with major potential negative impacts on teaching and learning and on teachers and students, and on the positive aspects of the reforms of the Curriculum Council including the introduction of the OBE framework.

This is not to say that teachers will not carry out many assessments of the work of their students and perhaps even more than the above number; however, formalising these assessments in order to provide them to the Council, and having these scores immediately have a high stakes element, will add considerably to the administrative burden of assessment for the Council, for schools, for teachers and for the students.

I understand that it is proposed that the these assessments be used in part for feedback to schools regarding the standards of the assessments and that only the last two units studied by a student will be used in the TER. However, for the feedback to be relevant, the assessments at Year 11 need to be as good as they can be, and for any student, the mark might be the one that is used in the TER – this makes the four marks

per unit, based on a strongly prescriptive mode of assessment, very high stakes and therefore inevitably time consuming. Thus for a course, each of the 16 scores will need to be as reliable, valid and precise, as the one score that is provided now. This amount of assessment from the Council, with the commensurate resources that it will consume, has the potential to distort the teaching and learning far more than any external assessment can distort it. One analytic mark per unit and one level mark should suffice for all the purposes required.

> ***Recommendation 2*** *The number of marks submitted by the school for each unit of a course be a maximum of 2, one an analytic mark out of 100 for each unit, and one in one of 8 levels to describe the generic student achievement.*

Analytic marking will be elaborated as part of Case Study 1, but for the present it is simply a traditional mark which can be scaled within a school to be out of 100.

Although the accumulation of marks is an important principle to motivate learning, it is incompatible with the principle that a student's standing at the end of Year 12 be used as the basis for selection for further study and that a student's achievement should be recognised irrespective of the way in which it has been learned. It is not tenable to consider that the level of student achievement in Semester 1 of Year 12, and maybe even in Semester 1 of Year 11, should be the same as the final semester, generally Semester 2 of Year 12. Analytic marking should account for this potential anomaly, and the marks in both units used for a TER for any student will tend to be similar. However, this will not always be the case and averaging the marks, or even weighting in a particular way a priori, will not always be just. The schools are in the best position to know the final achievement of a student in a course.

> ***Recommendation 3*** *That for those students eligible for a TER, and who have therefore completed at least 2 units of study, the school provides a final analytic mark and a level to the Council for the course as a whole.*

In most cases this mark will be the same as the mark that arises from the two units, but the recommendation permits schools to vary this mark in the case that such a variation is warranted and can be justified. This is the mark that should be used for the TER.

Scaling and equating

Because the terms *scaling* and *equating* of measurements need to be used, the basic principles that these terms imply are now also briefly summarised.

Common elementary measurements in every day experience and in the physical sciences have a well defined origin and an arbitrary but well defined unit. Each has a long and rich history in being established and agreed upon in different jurisdictions, a history that is generally overlooked. This arbitrariness of the origin and unit is understood by children in primary schools and is part of the mathematics curriculum. There may be more than one well defined unit for the measurement of some property of objects. This is the case for the measurement of mass with the pound and the kilogram or the division of the pound into ounces and the kilogram into grams. With

the typical measurement of temperature in either Fahrenheit or Centigrade, both the arbitrary origins and the arbitrary units are different. Often measurements in one scale need to be converted to measurements in another scale as in the case of the measurement of mass and temperature where conventional uses of one or the other in different jurisdictions are different.

In education and the social sciences, measurements are used in a way which approximates the use of measurements in the physical sciences. However, the unit and the origin of most assessments are unique to those assessments - there is no natural origin of zero knowledge for example, and no well defined unit such as a pound or a kilogram for mass for measuring the amount of knowledge in any course. Ironically, although measurement in education and the social sciences has even more arbitrariness and certainly less conventional agreement on the unit and origin of scale, social measurement seems not to be a topic in any school curriculum. This deficiency tends to persist in university curricula and only in some units within some degrees are they broached. A second element to the irony is that there is a tendency for greater belief in the consistency of origin and unit in social measurement than there is in physical measurement where their arbitrariness is made explicit. Thus numbers assigned to characterise degrees of a construct are used inappropriately as if they had the properties of physical measurements. A third element to the irony is that there are abundant examples of quantification in the social sciences, including of course assessment and measurement of student achievement which lend themselves to this study.

In the case where different measurements need to be compared, reconciled, or summed, it is necessary to convert them onto the same scale. The processes of *equating* and *scaling* are used for such a purpose. The principles behind these will now be described briefly. To make them concrete, the present process of reconciling marks between school based assessments and external assessments for each subject, and for reconciling measurements among subjects to provide a tertiary entrance score (from which a tertiary entrance rank is then obtained) will be used.

The term *scaling* is used generally when there are two or more assessments of the same construct and they need to be placed on the same scale, that is, to have the same origin and unit. This is done, for example, with school and external assessments in order to obtain a single score. The term *equating* is used when two measurements which do not refer to the same construct need to be brought to the same scale before summing them. This is done, for example, when the marks from different subjects are placed on the same scale. The basic principle that is used in both cases is the same and the terms *scaling* and *equating* are often used interchangeably. Therefore, this principle is summarised and only the term *scaling* will be used. The implications of adding measurements in different circumstances, for example school based and external assessments, performance and written assessments, and across different courses, are discussed further in the report.

The principle behind equating of social measurements

Consider two sets of measurements which are available for each person on a single construct where it is not known whether the measurements are on the same scale, that is, whether they have the same unit and origin. Often the need for having the

measurements on the same scale, and illustrated above, is that they will be summed and averaged.

For example, if two measurements of the same temperature were taken, and one measurement was in the Fahrenheit scale (say 99.5^{o}F) and the other in the centigrade scale (say 37.5^{o}C which is substantively identical to 99.5^{o}F), averaging them without recognising scale differences would obviously provide a number (68.5) that was not meaningful in its context and that would be grossly misleading.

To test whether two measurements are on the same scale in the social sciences, the following principle is used. *If the same group of people are measured twice on the same scale, then the two sets of measurements should have the same average and the same spread.* This can be elaborated to expect that they should have the same distributions, not just the same average and spread, but for the present exposition, the average and spread will be considered. The technical term for the idea of spread is the *standard deviation*. This indicates how far the scores are from the average and reflects on the unit of the scale. The two measurements of particular individuals may be different reflecting possible real relative differences and as well as error, but it is considered that for the group as whole, the two measurements should have the same average and the same standard deviation if they are on the same scale.

Conversely, if two sets of scores on the same group of persons do not have the same averages and standard deviations, it is taken to reflect that the origin and unit are not the same in the two measurements. They can readily be made the same by transforming the scores of one or the other, or both, so that they have the same average and standard deviation. Sometimes more advanced processes are required to deal with the whole distribution of measurements when it is evident that the unit is not consistent across the scale, but, as indicated above, these are not considered here for reasons of simplicity.

In the present process, the school based assessments and external assessments are both brought to the same scale. The external assessments themselves are transformed to ensure that the unit is the same at different points of the scale and to fix it to a conventional distribution which is the same from year to year. The unit and origin are inevitably arbitrary and the choice is made for convenience. The school based assessments are equated to the external assessments. The reason for this is not that the external assessment is more valid, but that it is taken by all students. Thus it provides a measurement that has the same origin and unit for all students.

When two measurements of the same construct are averaged, the average is considered to reflect more reliably and validly the measurement of the construct. In general this follows if the measurements, when placed on the same scale, are similar, or in technical terms, that they are homogeneous.

> ***Recommendation 4*** *That professional development be provided by Curriculum Council Officers to Principals, teachers and students regarding the arbitrariness of measurement units in educational assessment and the implications this has for placing the assessments on the same scale and ensuring that other policies of the Curriculum Council are applied correctly.*

The order and breadth of constructs such as subjects and courses

When two measurements are averaged, they define a construct which is more or less broad – the breadth is a matter of degree. Thus school based assessments and external assessments in the same subject in part assess the same content and in part complement each other and permit the assessment of content not assessed by both. They do not assess the same detail, but different aspects, of the same defined subject. They assess a *higher orde*r and *broader* construct of the subject than either the school based or the external assessment alone.

It is nevertheless expected that, after transformation to the same scale, individual student's scores on the different measurements, will be relatively similar. Extremely different measurements of a student generally need, and have, an explanation. The sum or average of similar measurements is a better summary of the broader construct than is just one measurement provided the measurements are relatively similar when placed on the same scale.

When two measurements of relatively different constructs are summed or averaged, they imply an even higher order and broader construct than in the case of an external and school based assessment of the same subject. For example, in summing the measurements from different subjects to obtain a TES, the TES becomes a summary of academic performance at a higher order and of a broader construct than characeterised by a score in just one subject.

Again it is expected that, when placed on the same scale, the measurements for a particular student will be relatively similar in their values. This is the principle applied in the summing of measurements across different subjects. The measurements of students in different subjects, after transformation, are generally relatively similar because students choose the subjects they study, and they choose those subjects which suit them most and on which they will perform the best. This justifies in general summing the scores to obtain a single TES which has been studied in Tognolini and Andrich (1996). However, to bring the measurements of the different subjects to the same scale, the assessments behind them need to be relatively similar in their precision.

To make this point concrete, suppose a broader concept of the size of people was to be characterised, rather than just height and weight, and that therefore the height and weight measures were to be summed. For this purpose, the height and weight measurements need to be of the same order of precision and on the same scale even though it might be difficult to say that weight and height are measured with equal precision. For example, suppose the original assessment of people for weight was in grams but the height was in feet (30.5cm). The measurement of virtually all adults in feet would be perhaps 4, 5, 6 and 7 feet, a range of 3 feet while the measurements in grams of weight would have a much greater range, perhaps 6000 grams. Thus even if the measurements in feet and the measurements in grams were brought to the same scale (same average and standard deviation) before being summed, the measurements in grams would dominate in distinguishing among the sizes of the people. By

analogy, it is important that assessments, for example school based and external, have a commensurate degree of precision before they are transformed to the same scale.

The scaling of subjects or courses is most important in order to assure all concerned that a higher measurement is not obtained in one subject compared to another just because the arbitrary units and origin of that scale are different. It is important for student equity *post hoc*, and for course selection a priori, that the measurements from the different subjects are on the same scale before being summed to give a higher order and broader summary of a student's relative standing. This is the reason behind the policy, mentioned earlier, that the different courses will not have different difficulties in obtaining the same score or measurement towards the TER. It is also behind the intention to have levels across courses of equivalent standard.

In summary, first it is considered that if the same group of students have measurements on each of two or more constructs, that for the group as a whole, the average and standard deviation of the these measurements should be the same. If they are not the same, then this is taken as a reflection of the measurements of the different constructs being on a different scale. Second, for such scaling of different measurements to be effective, it is necessary that the original assessments behind the measurements are of the same order of precision.

Further requirements for scaling and equating

.
Two further scaling requirements need elaboration.

First, I referred above to the idea of the *order* and *breadth* of a construct and its measurement. When two measurements of the same construct which are essentially of the same kind qualitatively are averaged, then the main point of averaging is to increase reliability and precision. When two measurements of different constructs are averaged, then the main point of the averaging is to summarise the performance on a broader construct. Thus a tertiary entrance score characterises a broader construct than does a score on a particular subject. The cases are not always this extreme. For example, somewhere in between these extremes is the case where a single construct is composed of qualitatively distinct components, for example where practical performance and written aspects are both relevant. These, too, need to be placed on the same scale using the same principles before being summed or averaged. However, to average the measurements in order to provide a summary for a broader construct is a policy issue, not inherently a measurement issue. It is a policy issue to decide that a course will be sufficiently broad to include both a practical and a written component, and that for example these two components will be given equal weight. This issue is picked up in Case Study 2 when data from the assessment of Drama are considered.

Second, the TER that is produced is a generic score that is used by a range of tertiary programs which have no specific subject prerequisites. Minimising prerequisites is another policy issue with major educational implications, including postponing at least to some degree premature specialisation by students. Some tertiary programs do specify additional prerequisites in conjunction with a minimum TER, or some may make selections based on individual subjects independent of the TER or in

conjunction with it, and some may even choose to use disaggregated components, for example only the performance components. At this stage, however, the generic TER signals a degree of academic achievement that includes a written component in each subject. It must be consistent with this signal.

> ***Recommendation 5*** *That professional development be provided by the Curriculum Council to Principals, teachers and students articulating the rationale for major policy decisions, their implications, and the mechanisms necessary to implement them.*

Further relationships between assessment and measurement

Some further principles of the theme of the report concerning the relationship between assessment and measurement, already broached to some degree, need some elaboration in leading to further recommendations.

First, as stressed already, data in the stage of assessment needs to be reliable, valid, and have sufficiently fine levels of precision that they are useful for the purpose for which they have been obtained. In particular, finer precision of measurement cannot be generated from assessments than the precision inherent in the assessments.

Second, analysing assessment data statistically with models to convert them into measurements can be used to diagnose and locate problems with the assessment stage and point to where the assessment stage can be improved. However, although problems can be diagnosed and used to improve further assessment, most cannot be overcome *post hoc* – only differences in origin and units can be transformed to a common scale *post hoc*.

Third, before being transformed to be on the same *measurement* scale, assessments are assumed to be on the same *assessment* scale when provided by a teacher for a particular class. It is necessary that the teacher carry out the assessments according to general policies for the course as an integral part of the teaching of the course and be consistent in assessing all of his or her students. With two or more teachers in a school teaching a course, it is expected that they will work together closely to ensure that the assessments across classes within the school will be on the same assessment scale.

However, for the principles of equating of averages and standard deviations outlined above, it is necessary that the number of students is not too small. Therefore, in addition to working together to provide the same assessment scale within a school, it is necessary to have a large enough numbers of students on the same assessment scale for the measurement transformations to be justified. The Curriculum Council has in place the requirement that teachers in schools with numbers of students 8 or less must work with a teacher in another school in order to ensure that they are part of a large enough group that will permit scaling of measurements.

The process used among schools with small numbers to ensure that their assessments are on the same assessment scale needs to be similar to the processes used within schools which have two or more classes taught by different teachers. That is, they

need to work together through the organisation of the teaching and learning of the course. *It might even be appropriate that where there is only one class in a school on a particular course, even if it is a large class, that the teacher has opportunities to interact with teachers in other schools.*

If resources are to be increased for school based assessment, then one obvious and potent contribution would be to fund some release time for teachers who have to engage with teachers from other schools. This kind of support has many other tangible benefits for staff and students when the number of students in a class is relatively small. If teachers from some schools cannot get together with teachers in other schools as required, then Curriculum Council officers may need to provide the required support. However, it is not considered here that the main challenge is to have every teacher provide the *correct* assessment on a specified scale for each student – the main challenge is for students in designated groups that are sufficiently large to have their fine grained assessments consistent with each other and on the same scale ready for further analysis and scaling for measurement purposes.

> ***Recommendation 6*** *That resources be made available in terms of some time release for teachers who have to work with teachers in other schools to ensure that they have a large enough group of students for equating and scaling to be effective.*

The current practice of identifying and reconsidering profiles of students that are not homogeneous enough to justify inclusion in equating and scaling should continue. This is assumed and not considered a special recommendation.

Consistency of classification and precision

The essential concern is to have reliable and valid assessments which can be used to form relatively precise measurements. In leading to further recommendations, the key theme is the *relationship* between the consistency and precision of assessments and precision of measurement in the context of the OBE course structure summarised above. One aspect of this theme has both a conceptual and practical component which in the first instance can appear counterintuitive. It is therefore important to have some of the features of this aspect clarified.

Clearly, in any assessment and measurement, there is a need to have both consistency of classification and a high enough level of precision for the task at hand. In particular, in assessments associated with OBE, there is a premium placed on teachers classifying students into levels. *However, and perhaps counter intuitively in the first instance, consistency of classification can be achieved readily at the expense of precision of measurement.* Therefore, these two simultaneous demands, consistency and precision, need to be reconciled with each other to ensure that consistency does not result in lack of required precision.

However, before considering issues in such reconciliation, an example of the uncertainty in classifying tasks into levels is provided. This case involves the classification of items constructed to reflect levels of outcomes. Case Study 1 involves the outcome of Writing. These two examples, together with Case Study 2

which involves the assessment of drama, provide information across the spectrum of the learning areas and courses.

An example of uncertainty of classification of mathematics items into outcome levels

Analysing data statistically can reveal systematic variation among the relative difficulties of tasks and should be used to account for this variation in order to place performances on the same scale. *Evidence from statistical analysis of performance against stated outcome levels demonstrates that a priori classification of levels are difficult to make unequivocally even in a single course across a wide range of the full 8 levels.* The reason for this is that tasks themselves provide different opportunities for students to reveal different levels of performance. These differences among tasks, ostensibly as the same level, arise because of ancillary features of tasks that make some easier than others. They are inherent to the situation and do not seem to arise simply from a lack of training of the markers, the item constructors, and so on.

To illustrate this point, Figure 1 shows the relative empirical locations of scored mathematics items against their original classification into levels by the test constructors who worked against an outcomes framework. The items were developed in order to monitor the progress of the students in the West Australian Education System (Van Wyke, 1998) specified in terms of the 8 levels.

A student's performance on each item was scored simply 0 or 1 for failure or success respectively. It was deemed that if a student completed an item successfully, then the student showed achievement at a particular level of the Curriculum Framework characeterised by the item. It was expected that the higher the level of an item, the more difficult the item.

The horizontal axis of Figure 1 shows the difficulties of the items on the same scale, and the vertical axis the designated level of each item. The pattern of relative difficulties is as required, with the *average* of the difficulties at any particular designated level greater than the average of the previous level. However, at the same time, it is clear that the locations of the items designated to be in particular levels *overlap* with each other considerably and that there is no clear demarcation of levels. This overlap is inevitable and is not evidence of poor test construction, lack of teacher professional development, and so on. However, it does indicate that the final analysis and placement of students on a common scale needs to take account of such differences. This can be done statistically as part of transforming the assessments into measurements. This example is discussed in more detail in Andrich (2002b) to illustrate the uncertainty in classifying students according to levels and further inferences that can be gleaned from the example are considered again in this report.

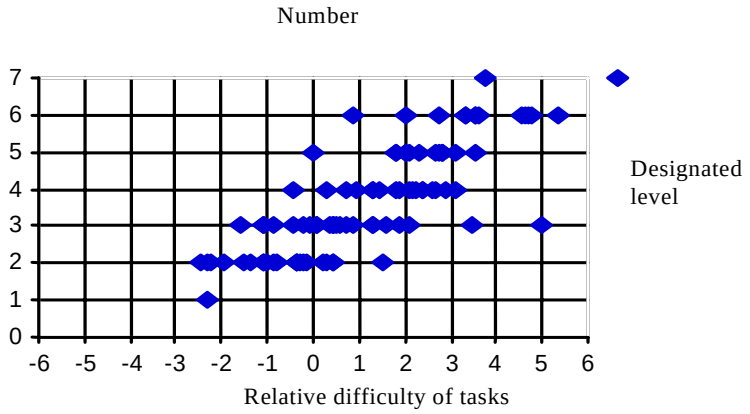


Figure 1 Relative empirical difficulties of tasks designated a priori at particular levels.

Potential sources of artificial consistency and imprecision

There are three main sources of consistency of classification which can work against achieving precision of measurement that are discussed in some detail here because of their pertinence. First, the classification system produces agreement because it is relatively crude in its classification[2]. Second, the classification levels for aspects of performances does not arise from the features of the assessment task; rather, it is an arbitrary classification system[3]. Third, there might be an undue effect of the classification of one aspect of a performance on the classification of other aspects[4].

Consistency arising in these circumstances is referred to as *artificial* consistency. Artificial consistency works against obtaining rigour and precision of measurement.

Each of these sources of artificial consistency is elaborated further below. They are elaborated because they are readily present in assessment according to levels of performances across outcomes or aspects of outcomes. Case Study 1 illustrates each of these sources of artificial consistency and in fact draws attention to them. They are summarised first for expository reasons.

---

[2] For example, if a person were to be classified into a level of height by judges, there would be much more agreement if the classification was made to the nearest 1m than if the classification was made to the nearest 1cm. However, despite the greater consistency in the response in the former, the estimates of height would be much less precise than in the latter.

[3] For example, all of the courses for Years 11 and 12 in are divided into 8 major levels, as are each of the aspects, and then proposed to be in 3 further levels. It is unlikely that every piece of assessment can be characterised naturally or usefully into the same number of levels. If there are A levels specified, but the performance of the tasks does not fall naturally into these levels, the classifications may all be in two or three levels and appear consistent.

[4] For example, if an assessor classifies the first aspect of an assessment into a level, he or she tends to classify the second or other aspects into the same level at a much greater rate than if a second assessor was asked to classify the second or other aspects. This suggests consistency, but is the well known, but often difficult to control, halo effect of over consistency.

Elaboration of potential sources of artificial consistency and imprecision in assessments arising from an OBE framework

The three sources of artificial consistency referred to above, that of a relatively crude level of classification, arbitrary levels of classification, and the impact of assessing one aspect on assessing other aspects, are now elaborated.

Relatively crude level of classification

As indicated earlier, the outcome statements in the Curriculum Framework and elaborated for different learning areas as illustrated above, begin with each outcome of a learning area summarised into a progression of 8 levels. Table 3 showed levels 5, 6 and 7 for the Writing Outcome.

Such descriptors as those in Table 3 are referred to in this report as *nodal* descriptors. They describe nodes of achievement with more or less distinguishable features of increasing demand. However, in general, students do not leap in knowledge in achieving at the level of one of these levels to another – progress is continuous in multiple related directions captured by the various aspects of the outcomes, which are brought together at different levels of scale in learning and in assessment. In addition, they are unlikely to be clearly at a particular level as evidenced by performance on any task or set of tasks. They will be somewhere in a range between these nodal descriptors. Sometimes they will at the boundary between the two. To understand these operationally, it is necessary to have a more precise assessment than provided by the nodal descriptors themselves. This is necessary for teaching and learning, not only for competitive tertiary selection. Further discussion of this point is articulated in Andrich (2002a) and Andrich (2002b), and is not repeated here. However, it is illustrated in detail in Case Study 1.

Despite their operational ambiguity as illustrated in the example of Figure 1, the nodal descriptors imply substantial growth from one to the other. The 8 nodal levels of achievement for each outcome in each learning area span Years 1 to 12 of schooling. Taken very broadly, 8 distinct levels cover learning in 12 years of schooling with each level capturing growth over approximately 1.5 years. This suggests that it is conceivable that all self selected students aspiring for entry into a particular competitive programs of study in a university or in TAFE, could be at the same nodal outcome level in a particular course at secondary level. Again, more precise assessment is needed than can be provided by the outcome levels.

A suggestion for increasing the precision is to have classifications within levels into three further sublevels, for example, 6-, 6, 6+, which might be weighted 6.2, 6.5 and 6.8 numerically. Within level 5 the three levels would be scored 5.2, 5.5, and 5.8. It also begins to presuppose, without further consideration of unit of scale, that the numbers represent measurements already on the same scale. This cannot be taken for granted. In addition, this proposition starts to encroach on the second potential cause of artificial consistency, that of having an arbitrary level of classification which is elaborated in the next subsection.

Arbitrary levels of classification

The Curriculum Framework itself has some aspects of an arbitrary classification, though of course they are not totally arbitrary. They arise from previous conventions and understandings. Thus the classification into 8 different learning areas has some sense of arbitrariness as the number of learning areas might have been specified to be 7 or 9 or 20. The courses, likewise, have a sense of coherence and a sense of arbitrariness. They also inevitably overlap in content. However, the learning areas and courses are well understood in the learning community and there is no problem with the arbitrary element. The classification is a useful convention for planning teaching, learning and assessment even though in various education learning contexts particular tasks may span even the 8 learning areas, as, for example, in writing reports in some aspects of scientific investigations.

The eight levels of classification of each outcome, however, seem totally arbitrary. There seems not to be general conventions in the history of the areas of learning that has these in any particular number of levels, let alone every area into exactly 8 levels. Again, however, having levels and their nodal descriptors can provide an organising framework for teaching and learning which can help teachers monitor the progression of students in a learning area at a general level. Therefore, the above observation of arbitrariness of the number of levels does not imply they are not useful – they are useful in organising the teaching and learning. However, the choice of 8 arbitrary levels for all areas suggests they cannot be useful for precise assessment.

The problem of the arbitrariness of levels becomes greater when considering the aspects of outcomes. As the level of assessment becomes finer and incorporates more aspects, the idea that the same number of arbitrary levels can be applied to every to every aspect of an assessment task becomes less and less tenable. If the levels are arbitrary and do not match aspects of a task, then markers will have to distinguish amongst other aspects of the task and make an artificial classification, which in turn will lead to artificial consistency.

Impact of assessments on one another – the halo effect

As indicated above, it is expected that assessors will scrutinise performances on assessment tasks and classify aspects of outcomes into levels. In the circumstances of the description of the outcome into levels, and then aspects into the same number of levels, it is readily possible for a classification at the outcome to dominate the classification at the level of aspects, and for some outcome to dominate others. The assessments across aspects and outcomes will then be much more the same than if they were not dominated by some overall aspect or outcome. This is the well known *halo* effect.

It is possible for this dominance to be exaggerated or precipitated by lack of clear distinctions among the outcomes or aspects of outcomes which have overlapping criteria.

Of course, it is intrinsically likely that a piece of work which is at one level in one aspect might be at the same level on other aspects and be at the same level overall –

that is the nature of an overall performance. However, the concern here is when the assessment across aspects is more consistent than it should be, that is, is artificially consistent, given the quality of the performance assessed.

Relationship between the assessment and precision of measurement

The above are sources of artificial consistency that lead to imprecision. They are subtle because under typical circumstances consistency is a desirable property. It is important to stress, however, that this issue is not only, or even primarily, a measurement issue. The qualitative features of the assessment, the quantitative features of measurement, and the relationship between consistency and precision, are intimately related and it is the analysis of measurements that can expose problems with the original assessments.

In particular, and in principle, if the classification system is (a) arbitrary and incommensurate with the task, (b) crude relative to distinctions that markers can perceive, and (c) prone to interference of different aspects or outcomes on each other, then the task of assessment is also potentially difficult for the assessors to carry out. That is, these properties which reflect problems in obtaining precision of measurement through artificial consistency are symptoms and manifestations of inherent problems in the assessment itself. This point cannot be over stressed. Thus although the initial concern regarding artificial consistency might arise because of the need for precision of measurement, the concern is also relevant for the stage of assessment: assessment which generates artificial consistency is a problem in the assessment, even if it is not subsequently used in any application of measurement. Artificial consistency is an inevitable refuge for assessors who have difficulty marking the performances and the different aspect of the performances on their merits. It also cannot be overlooked, that justifying a mark to a student when the mark is generated artificially, becomes extremely difficult.

All of these features, (a) crudeness of classification, (b) arbitrary levels of classification, and (c) the interference of assessing one aspect on other aspects, can be interrelated. They are all reflected in Case Study 1. However, importantly from the work of Heldsinger and Humphry, the Case Study also presents a process for overcoming these problems.

Case Study 1 – the assessment of writing

In presenting the recommendation to the approach to assessment that is most likely to overcome the problems of artificial consistency, and provide assessments which are precise enough for tertiary selection, a summary of the steps taken in Case Study 1 are presented to make concrete the basis for these recommendations.

The context of Case Study 1 is the assessment of the Writing Outcome within the English learning area at Years 3, 5, 7 as part of the Monitoring Standards in Education (MSE) program in the Department of Education and Training in Western Australia. The MSE assessment is carried out externally to the immediate school based

assessment carried out internally by teachers, though it is administered by teachers according to specified procedures. The MSE program not only has the task of monitoring the standards in literacy and numeracy, but in contributing to the professional development of teachers and to feedback to schools on results of the assessment.

As an illustrative case study, it is important to note that the assessment was only of one outcome and across a wide range of levels of achievement. Therefore, in principle, the writing outcome should have been easier to assess in terms of levels than if many outcomes were involved. Therefore, that there were problems with such an assessment indicates the problems might be even greater when more than one assessment is to be considered.

The construction of the task and its initial assessment

In assessing Writing, it is possible to ask students to write in a range of different genres. The genre chosen will be governed by trying to ensure that the students can show their capacities to write and to meet the outcomes to be assessed. This point is returned to later in the report. In this particular example, and for a combination of reasons, narrative writing was chosen as the writing task which was required of the students. The one narrative written by each student is assessed on a number of related aspects or criteria.

In this regard, the assessment of writing has one of the key principles of assessment that is used in OBE. That is, one performance on a task can be assessed with respect to more than one aspect of an outcome. The more general case is the same performance may be assessed with respect to more than one outcome. It is therefore considered relevant in gleaning an approach to assessment within an OBE framework involving designated levels. Requiring performance of tasks of this kind is potentially beneficial in the education process as it requires an integrated performance reflecting more than one aspect of an outcome as in the example cited, or more than one outcome.

In the initial stage of the assessment of narrative writing produced by children, the marking guide was closely aligned to the student outcomes statements (SOS) of the Curriculum Framework. In particular, the different aspects of writing were classified into the levels that arise from the Curriculum Framework outlined above, with in principle, their being 8 levels for each aspect. Specifically, the guide was developed such that criteria were aligned with outcomes, and *categories* within each criterion were aligned *directly* with the *levels* of the outcome statements. That is, for most assessment criteria, a score of 2 represents a level 2 outcome, a score of 3 a level 3 outcome, and so on. One general criterion, termed an on balance judgement (OBJ) of the writing as a whole, also aligned levels, was obtained from the markers.

The aspects of writing assessed, taken from the English learning area, are shown in Table 4.

Table 4 Original classification scheme for the assessment of writing

Aspect	Levels	Score Range	Aspect	Levels	Score Range
On balance judgement (OBJ)	9[1]	0-8	Form of Writing (FW)	8	0-7
			Subject Matter (SM)	8	0-7
Spelling (Sp)	6	0-5	Text Organisation (Paragraphing) (TO)	8	0-7
Vocabulary (Voc)	8	0-7			
Sentence Control (SC)	8	0-7	Purpose and Audience (PA)	8	0-7
Punctuation (P)	7	0-6			
			Total score range		0-61

[1] Includes levels 1 to 8 and 0.

Although the classification of all aspects arose out of the 8 levels, some modifications were made because students in the higher years of schooling were only in Year 7, and so some of the highest levels were not evident and were not in the classification system. It is clear that the minimum number of levels was 6 for each aspect, one being 7 and all others 8, except the OBJ which was 9 and includes category 0 for no response. If an aspect had 6 levels, then the scores were from 0 to 5, that is, one less than the number of levels. From Table 4 it follows that the possible scores when the aspects were summed ranged from 0 to 61.

The operationalised form of each of the levels for the markers arose from the levels in the writing outcome of the Curriculum Framework. There was substantial professional development and training of the markers, many of whom were teachers, and multiple marking of the essays[5]. Indeed, this professional training and the marking guide were considered very helpful by the teachers involved in understanding and assessing the outcomes. In the first instance, each marker marked each piece of writing on all aspects.

However, in the analysis of the data for the stage of measurement in locating students on a writing outcome continuum, a number of symptoms showed that there was substantial artificial consistency in the data[6]. It will be recalled that artificial consistency produces redundant data rather than independent relevant information for each aspect. Redundant data reduces the precision of measurement.

One of the key symptoms of the observed artificial consistency was that that it seemed that too many students had a relative small number of scores that were the same. Thus although the possible range of scores was from 0 to 61, some score points appeared much more often than could reasonably be expected. This meant that the information did not provide differentiation among students among whom differentiation should have been possible.

[5] Overall, approximately 100 teachers and 15000 scripts were involved in different aspects of the studies.

[6] Among reference points for identifying this artificial consistency was the availability of results in reading and mathematics which were not marked in accordance with the SOS. It might of course be argued that it was these other two areas that had the problem. However, the other important reference point was the model of analysis of the data. The model builds into it the requirement that the data do not show artificial consistency, and if they do, then there are diagnostic features which reveal this.

The halo effect in assessing across aspects

In studying these symptoms, one hypothesis was that the marking of all the aspects of a piece of writing by a single marker would generate a halo effect, especially in the context of having an OBJ criterion. That is, that the same judge would tend to give the same level on all aspects more than would different judges if they each marked a different aspect.

This hypothesis was investigated by having each marker mark only one aspect. It indeed proved to be the case that the symptoms of artificial consistency were reduced. However, although reduced, there was still strong evidence of artificial consistency remaining in the marks.

Constructing an analytic marking system

Having removed the halo effect, yet still observing artificial consistency, Heldsinger, Humphry, and the MSE program, concentrated their studies on analysing the marking keys or rubrics of the aspects. It became evident to them that some of the marking keys across aspects overlapped logically and semantically, and that it was not surprising that the classification of levels across aspects tended to be the same. This overlap seemed to arise from the use of generic descriptors. Therefore these were reexamined logically and semantically as well as empirically by studying the writing that was produced.

In the process, and in summary, it became evident (i) that not all relevant aspects of the narrative writing which bear on the quality of writing were covered by the outcomes, and (ii) that because all aspects were placed in the same arbitrary levels, most of the aspects were marked in a way which was not relevant to narrative writing that the markers were assessing. For example, *characterisation and setting*, aspects central to narrative writing, were not present as aspects, and the aspect of paragraphing only worked operationally in one of 2 or 3 levels, and not in 8.

As a result, the marking guide was rewritten substantially and into the form of an analytic marking key. First, aspects not covered by the writing outcome originally that were relevant to narrative writing were included, and second, each of the outcomes was classified into levels which arose from the aspect and which markers could distinguish. The challenge was to have the number of levels with which the markers could work successfully, neither too many which would generate many redundant and confusing ordered categories, nor too few which would create frustrations for markers who could see differences in performance but could not reflect these in their marking. *In summary, the classification system arose from the task.*

It is important to note that this is an example of assessing Writing, which often is considered to lend itself to holistic marking and not in mathematics or science which generally are considered to lend themselves less to holistic marking. This point is returned to later in the report when inferences from this example are elaborated.

This does not mean, of course, that the assessment was not related to the levels. It was related because it arose from the outcome of writing. *However, it meant that in assessing the performance on the particular task, that the generic framework of similar levels on all aspects was deficient in omission of some features and commission of others*. The OBJ assessment, which is most relevant at the general level of the writing outcome in the recommendations, was retained. Because students only up to Year 7 were involved, the highest level retained was level 6. However, the categories of the OBJ were aligned with the level nodal descriptors.

The classification system which resulted from this work is shown in Table 5. The full new marking key for narrative writing in Years 3, 5 and 7, based on further research, is described in detail by the Department of Education and Training (2005).

Table 5 Revised classification scheme for the assessment of writing

Aspect	Categories	Score Range		Levels	Score Range
On balance judgement (OBJ)	7	0-6	Character/Setting 7****	4 6	0-3 0-5
Register 3 5**	3	0-2	Spelling	8	0-7
Narrative Structure 3*	4	0-3	Vocabulary	7	0-6
Narrative Structure 5 7***	5	0-4	Sentence Structure	5	0-4
Ideas	6	0-5	Punctuation	2	0-1
Character/Setting 3 5	3	0-2	Paragraph 3 Paragraph 5 7	3	0-2
Total Score Range			Year 3		0-42
Total Score Range			Year 5		0-43
Total Score Range			Year 7		0-42

*Pertains to Year 3, ** pertains to Years 3 and 5; *** pertains to Years 5 and 7, ****pertains to Year 7

It is evident from Table 5, in comparison to Table 4, that some, not all, aspects were changed, and that the number of categories for the classification system was different across aspects. Furthermore, some criteria were separated in application to different year levels and some were not applied to some year levels. This variation exemplifies making the criteria relevant to the task and to the performances of the students engaged in the tasks. Both the chosen aspects and the number of categories reflected the evidence that could be obtained from the writing. The method of analysis shown in this report permits placing the analytic marks of the Year 3, 5 and 7 students, and the OBJ level assessment, on the same scale even if not all the students were assessed on exactly the same criteria. This method of analysis is the application of the Rasch model.

Particularly relevant is that the possible scores that could be obtained from the classification system of Table 5 ranges from 0 to 42 or 43, which is a smaller range of possible scores than 0 to 61 which can be obtained from the scheme in Table 4. However, in fact the precision of measurement obtained from the scheme of Table 4 was greater than that which could be obtained from that of Table 5. This is because the redundancies in the assessments across aspects were virtually removed.

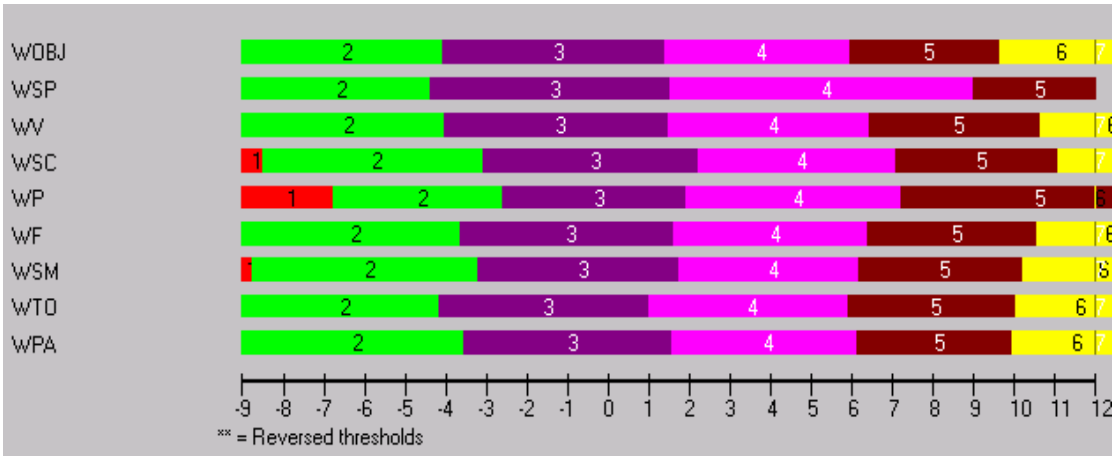
Crudeness of classification

The third source of artificial consistency, that of relative crudeness of classification, was also removed in the process.

Conceptualisation of the levels as marks on a ruler

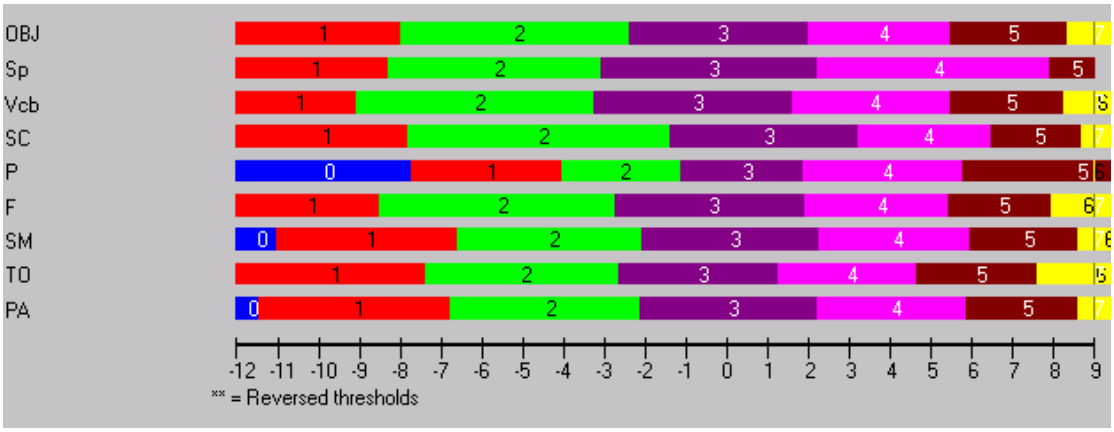
In the original marking guide, each of the aspects of the writing outcome was divided essentially into the same number of levels. Theoretically, these levels were the same in number because it is intended that they also be the same in intellectual and conceptual demands. The statistical analysis of the data provides opportunities to examine the way that these levels relate to each other. This is done by having a threshold between each level characterise each aspect on the same scale. These thresholds are essentially like markings on a ruler which designate different numbers of units from the origin. Just as units on different rulers, the thresholds from different aspects may be different. The analysis permits identifying the empirical distances between thresholds – that is, the effective sizes of the units of each aspect in the data at hand.

Figures 2a, 2b and 2c show the threshold alignment characterising the levels for each of the three analyses referred to above, (a) the original data when all levels were expected to be the same (and were the same in theory), (b) when the halo effect was removed by having a different marker mark each aspect of a single narrative, and (c) when the marking key was rewritten to remove overlapping conceptualisations of levels and aspects and where the number of categories for each aspect arose from a closer semantic and empirical study of the task.



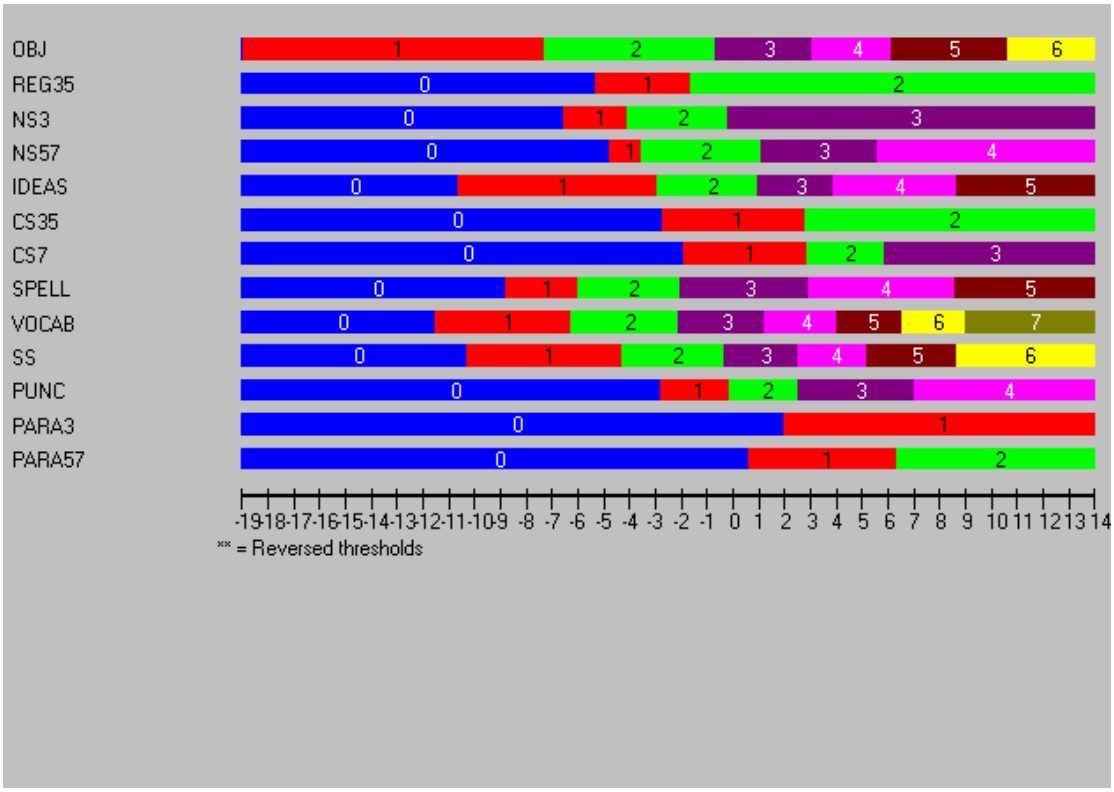
No reversed thresholds

Figure 2a. Original threshold map across aspects showing they are closely aligned: numbers between thresholds correspond to levels.



No reversed thresholds

Figure 2b. Threshold map when different markers marked different aspects showing less alignment than in Figure 2a: numbers between thresholds correspond to levels



No reversed thresholds

Figure 2c. Threshold map when different aspects had different criteria showing that the thresholds are not aligned. Only for the OBJ do numbers between thresholds correspond to levels.

The first criterion in the Figures is the OBJ classification into levels and is returned to later in the report.

It is evident from these Figures that in the first of these cases (2a), the thresholds were very much aligned, that in the second case (2b) they were fairly well aligned, but not as well as in the first case, and that in the third case (2c) they are very much not aligned.

The first case gives the impression of consistency and successful classification of persons across levels and a demonstration that the levels do work as expected in that each level is quite close across aspects in intellectual demand.

The second case, which shows less systematic aligning of thresholds, takes account of the halo effect, and so is clearly the more rigorous data set of the two sets. Thus better aligning of levels through thresholds cannot be equated with rigour of assessment.

Finally, the third case, where the categories are not expected to be aligned, indeed shows that the thresholds are not aligned. Again, the last of these assessments involved aspects of characterisation and setting, clearly relevant to narrative writing, and the category descriptors arose from the conceptual and empirical analysis of such writing rather than of a general set of levels. Therefore, it should be the case that such an assessment is more rigorous than an assessment which has generic category descriptors on aspects that may or may not be manifested in the writing task. *Thus the more rigorous the assessment, the less the categories of the assessments of the aspects are aligned.*

Figure 2c shows that the different aspects can be equated, and in addition, how they can be equated to the overall general OBJ classification at the levels. This point is again returned to when considering further inferences from this Case Study.

The relationship between precision and alignment of thresholds can be confirmed by considering the distribution of persons relative to the distribution of thresholds for each of the three analyses referred to above. Figures 3a, 3b and 3c show the distribution of persons in relation to the location of thresholds for each of these cases. The horizontal axis shows the distribution of persons (above) and the distribution of thresholds (below) on the same scale.

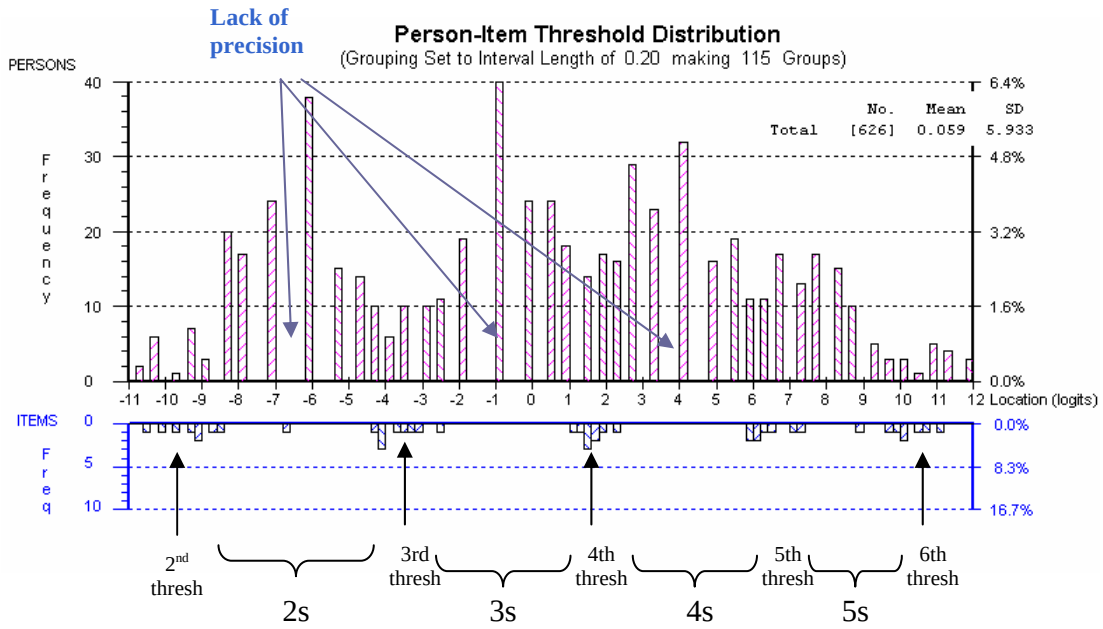
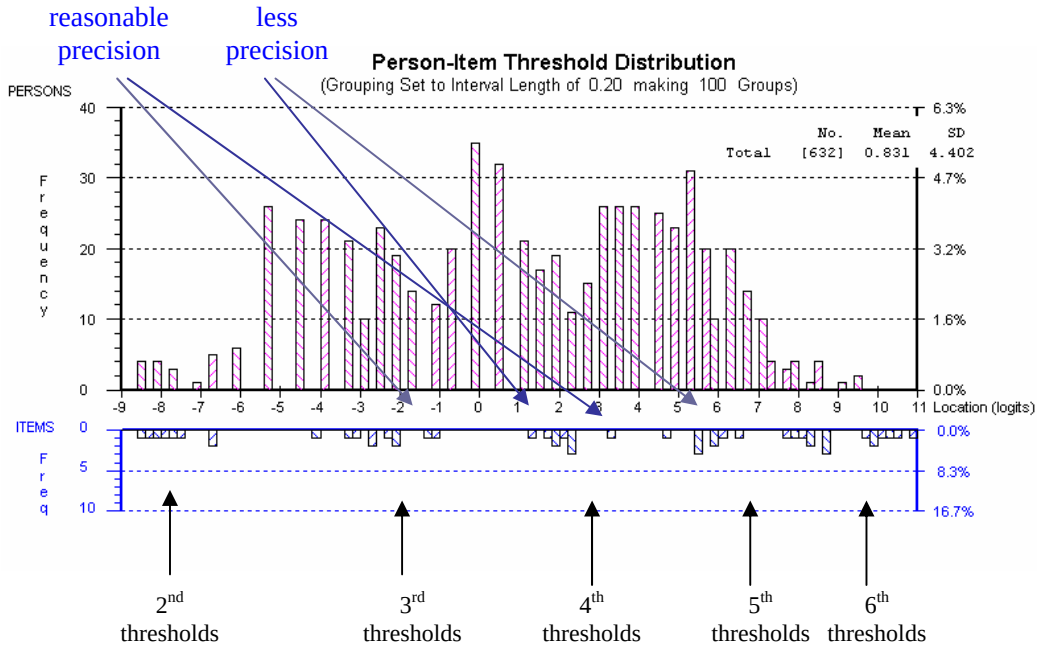


Figure 3a Person distribution and summary threshold map for original analysis when there is explicit alignment among categories. Thresholds are below the horizontal axis.



Clumping of thresholds across most or all criteria with different precision in different regions

Figure 3b Person distribution and summary threshold map for the analysis when different markers marked different aspects of the same narrative.

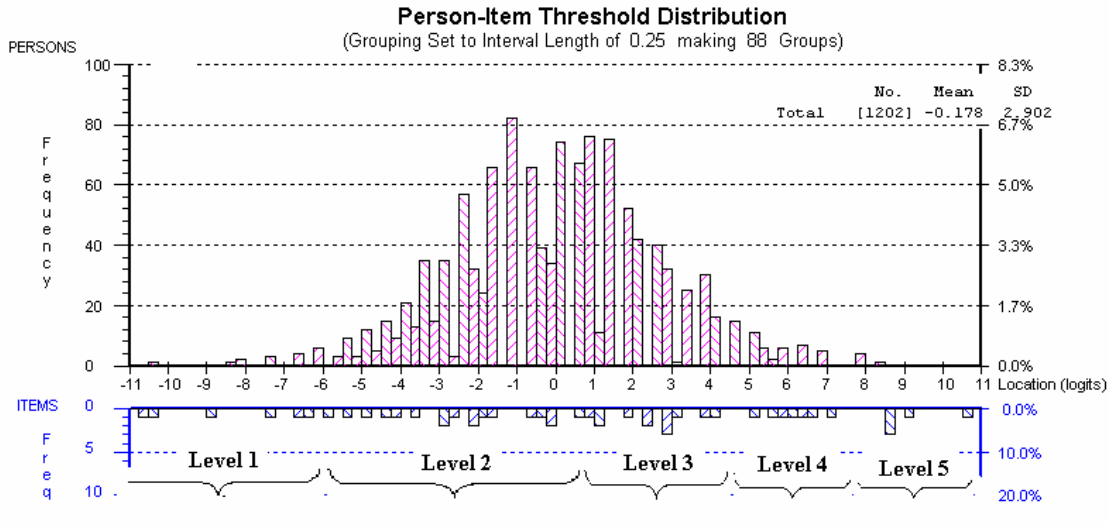


Figure 3c Person distribution and summary threshold map for the analysis when different aspects had different criteria – thresholds are not aligned but distributed throughout the continuum.

One of the key observations from Figures 3a, 3b and 3c, stressed by Heldsinger and Humphry (2005), is that in Figure 3a, the thresholds are aligned close to each other

from the different aspects on the continuum, and the person distribution shows peaks between where the thresholds are located. Where no thresholds are located is a region of *relatively crude* assessment, and writing performances, which might be distinguishable using a more refined marking key, are not distinguished by this assessment.

The assessments reflected in Figure 3c has thresholds distributed more evenly across the continuum and as a result it has the person distribution more continuously distributed without the same kind of peaks as in the first assessment. The more even distribution of thresholds from the different aspects, which arose from different aspects having different numbers of relevant categories in the analytic marking guide, provides precision of assessment across the continuum, whereas in the first assessment, there is precision only in narrow regions of the continuum.

Heldsinger and Humphry also make the point that having a marking key in which the categories on different aspects are not aligned to levels also removes powerfully the halo effect because markers cannot simplify their task when there is doubt in a mark by giving the same level as that of another aspect. It should be evident that the halo effect can be very powerful when an assessment is being made in a general way from generic descriptors for a performance including more than one aspect, or more than one outcome. It is important that it be circumvented in any assessment.

Observations on the assessment process

The above discussion focussed on the quantitative analysis. Heldsinger and Humphry report that the original marking scheme of Table 4 was considered useful by the markers in giving guide to the outcomes, the levels and their interpretation; accordingly, there was some concern to shifting to a new system. However, they report that following some training, the new marking scheme was found to be much easier to use than the original one. In addition to the analytic marking on the relevant aspects, the OBJ classification into an outcome level was carried out simultaneously. Further, the analysis can formalise the relative locations of the levels with the locations of the categories of the marking scheme.

In Figure 3c, the OBJ classification in terms of successive integers is with respect to the original outcome statement level, that is, the integers correspond to the levels, with a score of 3 corresponding to a level 3 outcome, and so on. The classification in the other criteria into ordered categories is also scored with successive integers even though these do not correspond to levels, only to the relative ordering of performance. It is also useful to further explicate the interpretation of the thresholds that mark off the region of these integers. It is relevant to note that as a first step in the formal analysis, the integer scores on the marking guide are simply summed, as is done in conventional marking.

First, it is evident from Figure 3c that with the analytic marking key, any particular student's performance can be located at a point throughout the continuum, and not just at the distinct levels of 1, 2, 3 etc, described by the nodal level descriptors. *The precision with which this point is located depends on the assessment task and the precision of the marking guide.* As can be seen in Figure 3c, the categories corresponding to the level descriptors obtained from the OBJ are mapped jointly with

the analytic marking results onto the same continuum efficiently and simultaneously, illustrating the second point.

Second, the scale of the continuum is shown at the bottom of Figure 3c – it ranges from -9 to +9, but this is a metric that arises from the data and can be transformed to other convenient values. It permits the marks from the different criteria to be mapped against each other.

Third, a pair of successive thresholds marks off the region where the particular level *is most likely to be given.* Thus the OBJ level of 2 is most likely to be given to a student who is located throughout the region between approximately –6.0 to 0.5 on the scale and an OBJ level of 3 is most likely to be given to a student who is in the region between 0.5 and 4.3 on the scale. It is evident that the region for level 2 is longer than for level 3, which in turn is longer than for level 4, and so on.

In the case of the Year 5 marking key of Table 5, these correspond to analytic scores between 6 to 19, and 19 to 27 respectively. That is, a raw score of 19 is at the threshold between level 2 and level 3. It is emphasised that the level classification is *not* made with certainty, but probabilistically relative to the analytic marking. *Thus in theory, a Year 5 student with an analytic score of 19 is just as likely to be placed in Level 2 as Level 3. In fact, there were 24 students in Year 5 with a score of 19, and of these 10 were placed in Level 2 and 14 in Level 3.* This is well within chance levels, as 10 or less in one level with a probability of 0.5 from 24 replications would arise around 27% of the time by chance. *This powerful analysis is possible because the analytic marking is at a finer level of scale than the classifications into levels.*

Fourth, it can be seen that the range of a score of 1 on *ideas* maps onto a level ranging across 0 and 1 on the OBJ. It can also be seen that ranges reflecting scores of 3, 4, and 5 in spelling overlap with level 2 of the OBJ. Figure 3c further emphasises the crudeness of the OBJ levels relative to the precision that can be obtained from the analytic marking scheme shown in Table 5.

Fifth, the Figure places into perspective the use of the nodal descriptors for the levels – they characterise a region of the achievement continuum and not a point, and a student in the region is most likely to be classified in a level, and is not going to be classified with absolute certainty. Of particular importance is the interpretation that a student can be at the margin between two levels, in which case there is 50% chance that the student will be classified in one level or the other. This is not a reflection of incompetence on the part of the markers, it is an inevitable consequence of the assessment, measurement, and the vagaries of student progress – they do not involve huge jumps and they involve uncertainties in the classification.

Finally, as indicated already, the level statements are generic, and they should be only used at the same general level to guide the teaching, learning and assessment, and not to make precise assessments at a finer level of scale necessary for other purposes. In this report the focus is on precision for tertiary selection, but the point can be extrapolated for precision of assessment for student feedback and other communication.

Further principles that can be derived from Case Study 1

This section abstracts the general principles to be learned from this detailed research study into the assessment of the outcome of writing which bears on the issue of rigour of assessment and precision of measurement

First, having many aspects and aligning the levels of all of these aspects of an outcome does not necessarily lead to rigorous assessment and precise measurement. If this is the case with the one outcome of Writing, it is unlikely that an alignment of levels across different aspects and across different outcomes from the same course can be sustained and certainly needs detailed checking.

Second, as indicated in the introduction, it is assumed that within the context of the Curriculum Framework, articulated student outcomes, and the development of courses, the teaching and learning is governed by the OBE principles. In addition, the setting of assessment tasks can be governed by the same principles in ensuring outcomes are covered and assessed. However, if rigour and precision adequate for distinguishing among students for tertiary selection is to be reached, then the marking keys and rubrics for assessment tasks which are generic and not referenced to the tasks, are most unlikely to be adequate. *Marking keys with the relevant number of categories need to be constructed in relation to actual tasks, not directly against the outcomes. These need to arise naturally from the task and not artificially from an a priori prescription.*

Third, it is noted that the above example involves writing, perhaps in some ways the most amenable to holistic or level based marking. The conclusion above, therefore, can be extrapolated to all areas of learning. It should be used in all areas of learning listed in Table 1, and not merely for some subset of areas.

Fourth, in the context of the finer level data summarised in point 2 above, the classification of the OBJ against 8 generic levels by course or by the performance on a task can be retained as a distinct criterion. This permits the scores on the other aspects which are assessed and which arise naturally from the task (which in turn arises from the Curriculum Framework and the course), to be mapped simultaneously onto these general levels. The mapping of students into 8 generic levels can be used for the purposes for which such a level of generality is sufficient. However, for purposes of selection into competitive courses in converting the assessments into measurements, the OBJ criterion is most likely too crude. Thus both a level classification for monitoring learning and much finer marks for competitive selection can be generated from the same assessment tasks, be analysed simultaneously and be mapped against each other. *However, the marks must arise from the analytic marking of a task, and these marks mapped on to the levels; the process cannot be reversed.*

Fifth, although these principles have been abstracted from a study which is most closely aligned to the format of an external assessment, there is nothing particular in the situation that indicates these principles are not relevant to all assessments, including school assessments. *The school based assessments need to have the same level of precision as the external assessments if they are to have the same contribution to the final selection. Therefore, school based assessments should follow the same principles.*

In summary, the tasks used to assess achievement of outcomes need to have analytic marking keys for aspects that arise naturally from the tasks, in kind, number and levels of categories, and a simultaneous OBJ for the less precise assessment of a level for each task.

In this case the marking key is for an essay, but the same principle can applied to other performances, such as drama or productions of other kinds or the solving of mathematics problems.

Case Study 2 – the assessment of drama

The second Case Study arises from data provided by the Curriculum Council involving the external assessment of Drama in 2004. This case study provides a number of important illustrative points for assessment. Two papers have been prepared by the Chief Examiner Robin Pascoe (2002, 2004) on the assessment of drama in Western Australia. Pascoe (2002) makes the important observation that

> It pioneered and developed a range of practices and procedures designed to marry more conventional assessment and statistical approaches with outcomes approaches to assessment (p.2).

How external assessment of drama was derived from the curriculum of the course is described in Pascoe (2002) and only some key elements of the assessment plan are summarised here.

For the data set analysed, the assessment scheme is summarised in Table 6.

Table 6 The structure and classification system for assessment of drama

Aspect	Categories	Score Range		Categories	Score Range
Performance			Written		
Solo Performance	21	0-20	Analysis, Interpretation Q1.	7	0-6
Improvisation	11	0-10	Analysis, Interpretation Q2	7	0-6
Monologue	16	0-15	Analysis, Interpretation Q3	9	0-8
Oral Interview	6	0-5	Australian Drama Q1 OR Australian Drama Q2 OR Australian Drama Q3	16	0-15
			20th Century Drama Q1 OR 20th Century Drama Q2 OR 20th Century Drama Q3	16	0-15
Total Performance		0-50	Total Written		50
			Total Score Range		0-100

There are clearly two major components in the assessment, *performance* and *written*. Within performance, there are four aspects of which all are compulsory; within the written, three questions of one aspect are compulsory and one of three questions in each of two other aspects are compulsory. The written and the performance are each supposed to be weighted equally; hence the maximum score of 50 for each. However, as will be shown shortly, this does not guarantee that the two components are weighted equally in the data.

Within each of the criteria, there are three ordered descriptors based on outcome levels and then there are further subdivisions. However, these subdivisions are not based directly on their own descriptors, but on qualifications by ratings within the descriptors. *It is particularly relevant that the different criteria do not all have the same number of categories - the number of ordered classifications ranges from 21 to 6.* The categories reflect the number of categories that the examiners considered markers could use profitably. In part because of the relative performance of the students against the criteria, and in part because of the actual use of the criteria by the markers, not all categories worked equally well.

Table 7 shows the frequencies of the response in each of the categories. In this study, the analyses of the assessments of 905 students who had complete data are reported. It shows that a number of categories at the extremes of each aspect had no responses, implying that they were not used. They can be retained, however, to provide a frame of reference for the marking key. In other years they may be used depending on details of the tasks set.

Table 7
Frequencies of responses in each of the categories

	Score									
Item	0	1	2	3	4	5	6	7	8	9
1	0	0	0	0	2	7	12	20	32	53
2	0	0	0	18	53	258	270	174	113	17
3	0	0	1	1	2	21	32	51	189	153
4	0	11	123	550	179	42				
5	0	5	61	361	404	69	5			
6	0	6	55	318	427	90	9			
7	0	3	11	45	174	403	216	48	5	
8	0	3	2	12	19	33	83	119	205	178
9	5	4	6	15	29	83	154	202	158	126

	Score										
Item	10	11	12	13	14	15	16	17	18	19	20
1	77	129	134	127	102	91	54	30	24	9	2
2	2										
3	222	95	63	50	22	3					
4											
5											
6											
7											
8	126	71	36	15	3	0					
9	80	25	8	10	0	0					

Figure 4 shows the distribution of persons and thresholds as in Figures 3a, 3b and 3c in Case Study 1. Relative to an arbitrary origin of 0, and on the natural metric of the data shown on the horizontal axis, the average is 0.516 and the standard deviation is 0.643. A traditional reliability index of internal consistency has a value of 0.827 relative to a maximum of 1.00. This is perhaps a somewhat low value, but results from the combination of the performance and the written components, which as will be shown and discussed shortly, are not overly highly correlated. Furthermore, the

performance component alone has a reliability index of 0.901, while that of the written component a reliability of only 0.682. This perhaps in part reflects the feature that the written marks were all in the range of 10 to 41, whereas the performance marks ranged from 10 to 48. The range of observed raw total scores for the combined components was 27 to 85. It is evident from the bottom distribution of Figure 4 that the thresholds span the whole continuum and from the top distribution that there is only minor evidence of persons falling in too few groups.

From the analysis, it is possible to show how the scores on the performance and written components equate with each other. Figure 5 shows this relationship graphically. Two equivalent scores are shown on the graph for the average location value of 0.52 on the common continuum. Thus the average score of 27.90 on the written component composed of Q5,Q6,Q7, and Q9, Q12 (recall that there was a choice in the written) equates to an average score of 31.00 on the performance. At the higher location of 1.5, a score of 34.30 equates to a score of 38.10 on the performance, a difference of almost 4 marks.

Given that these scores are derived from the same students, and using the principles for scaling and equating outlined earlier in the report, it would be concluded that the higher marks on the performance are somewhat easier to obtain than on the written, rather than that the students themselves are in some sense differentially able on the two. This again results in part because the performance marks range from 10 to 48, while those in the written only range from 10 to 41, a difference in range which is interpreted as a feature of the marking scheme rather than of the students' capacities in the different components of performance and written. This inference in the broader context of construct definition is picked up shortly. However, although the analysis used here takes account of this difference in the relative easiness of the marks in the two components, by simply adding the raw marks from the two components, and using these raw scores directly without further scaling, the weightings are not equal as intended – the performance plays a greater role in distinguishing among students than intended by the policy. It should be stressed that as far as equivalence is concerned, these data are perhaps as good as they get, and that these finer points of observation and requirement of scaling, are inevitable. Further, even if they were closer in this data set, there would be no guarantee that they will be as close in any other data set on any other year. It is necessary to check and routinely scale the marks to ensure that the stated policy is implemented.

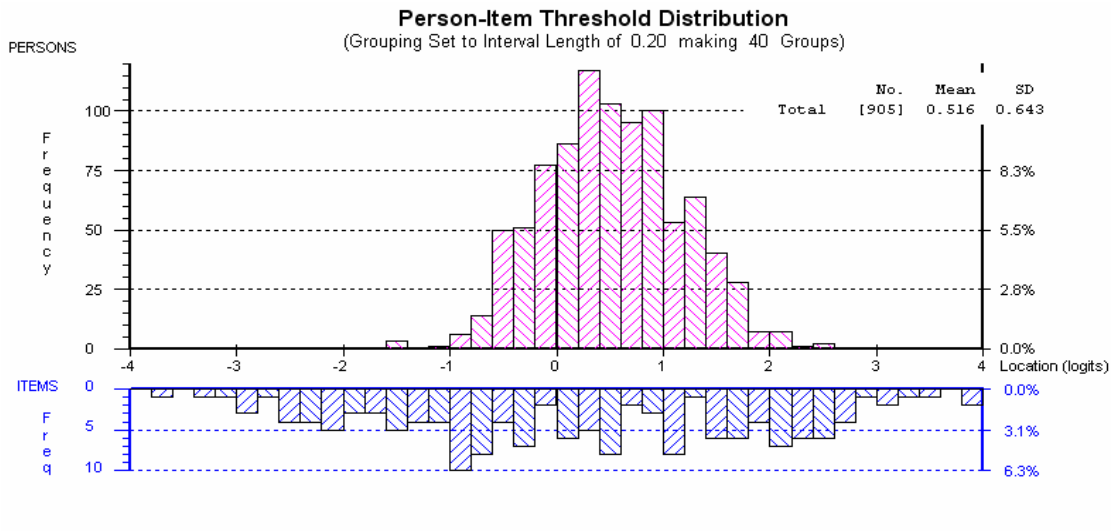


Figure 4. Distribution of the persons and thresholds for Drama

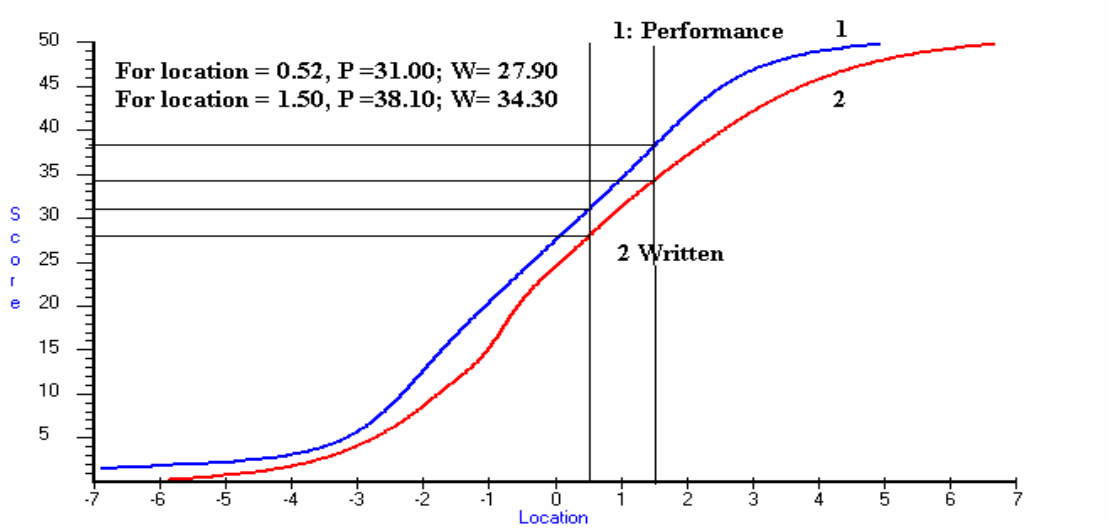


Figure 5. Score equivalence between the performance and written components

Table 8 summarises the raw scores of the students in terms of their means and standard deviations. It confirms the evidence in Figure 5 that the marks tend to be higher and have a greater spread in the performance component than they do in the written. Figure 5 shows in more detail how the marks are related over the entire continuum rather than just on the average.

Table 8 Mean and standard deviation for the main components of drama assessment

	Performance	Written	Correlation
Mean	31.08	27.87	
Standard Deviation	5.96	4.52	0.43
Range	10-48	10-41	

To summarise, given that the students are the same in the assessment of the performance and written components, the general conclusion is that the differences in scores shown in Figure 5 are a result of the different measurement units. To sum them, and to give them equal weighting according to the stated policy, it is necessary to transform them both to a common scale.

Variation in the performance and written assessments

The above features of the scores in the performance and written components of drama are consistent with another feature of the data - namely that the performance components are more highly correlated amongst each other than are the written components. The greater spread in the performance component is a manifestation of the greater correlation among the performance components than among the written components.

Table 9 shows this inter-correlation matrix with the correlations among the performance items and among the compulsory written items in bold. In addition, the correlations among the non compulsory written items are also in bold. It is clear that the correlations among the performance items are higher than the other correlations. In part this might be a result of a closer alignment with the components, or it might reflect some halo effect or other redundancy effects. Indeed, there is a suggestion of this effect from the distribution of the persons on the performance marks. Figure 6a shows the distribution of the students on the performance component only, and Figure 6b shows the distribution of the written component only. There is a suggestion of clumping in the performance distribution, which can signal artificial consistency as seen in Case Study 1. From these data, without studying the actual assessment processes, and perhaps without carrying out further data collection to assess these possible explanations, they cannot be completely resolved. However, an analysis of the marking key from the perspective of possible artificial consistency is indicated. For completeness, Figure 3b shows the distribution of the written component. It has less evidence of clumping than the performance component, but the scores are also spread less.

Table 9 Correlation matrix among the criteria in the assessment of Drama

	P1	P2	P3	P4	W1	W2	W3	W4	W5	W6	W7	W8	W9
	1	2	3	4	5	6	7	8	9	10	11	12	13
1	**1.00**	**0.66**	**0.73**	**0.56**	0.10	0.24	0.29	0.28	0.27	0.29	0.24	0.28	0.34
2	**0.66**	**1.00**	**0.69**	**0.59**	0.08	0.21	0.25	0.25	0.21	0.35	0.27	0.30	0.35
3	**0.73**	**0.69**	**1.00**	**0.67**	0.06	0.24	0.27	0.37	0.23	0.37	0.24	0.25	0.37
4	**0.56**	**0.59**	**0.67**	**1.00**	0.07	0.25	0.29	0.29	0.22	0.39	0.25	0.27	0.35
5	0.10	0.08	0.06	0.07	**1.00**	**0.48**	**0.40**	0.04	0.03	0.06	0.07	0.27	0.04
6	0.24	0.21	0.24	0.25	**0.48**	**1.00**	**0.63**	0.26	0.21	0.09	0.19	0.40	0.23
7	0.29	0.25	0.27	0.29	**0.40**	**0.63**	**1.00**	0.39	0.29	0.17	0.23	0.46	0.23
8	0.28	0.25	0.37	0.29	0.04	0.26	0.39	1.00	.(a)	.(a)	**0.30**	**0.41**	**0.47**
9	0.27	0.21	0.23	0.22	0.03	0.21	0.29	.(a)	1.00	.(a)	**0.31**	**0.49**	**0.38**
10	0.29	0.35	0.37	0.39	0.06	0.09	0.17	.(a)	.(a)	1.00	**0.25**	**0.33**	**0.48**
11	0.24	0.27	0.24	0.25	0.07	0.19	0.23	0.30	0.31	0.25	1.00	.(a)	.(a)
12	0.28	0.30	0.25	0.27	0.27	0.40	0.46	0.41	0.49	0.33	.(a)	1.00	.(a)
13	0.34	0.35	0.37	0.35	0.04	0.23	0.23	0.47	0.38	0.48	.(a)	.(a)	1.00

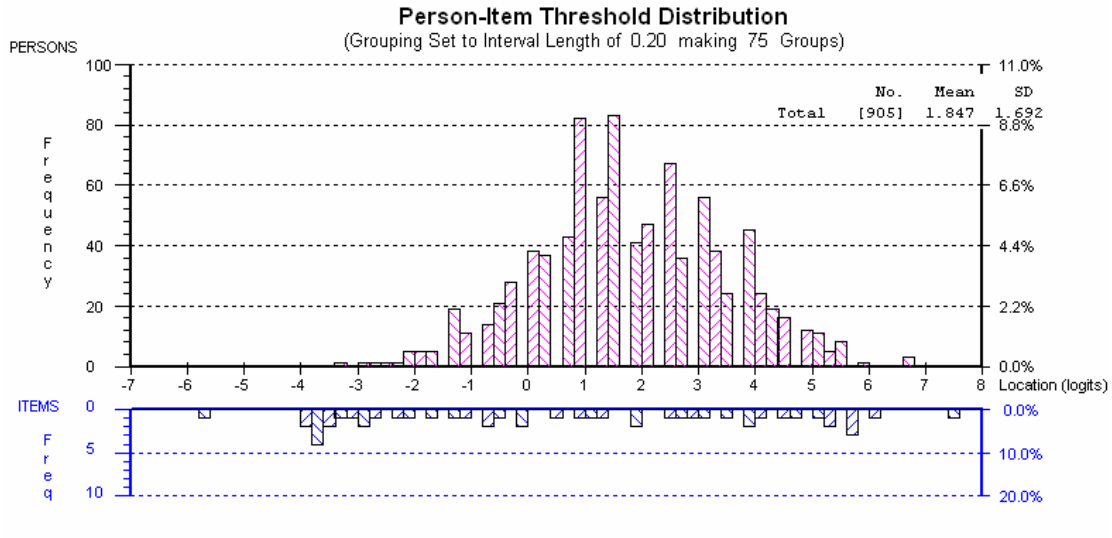


Figure 6a. Person distribution of the performance component of drama.

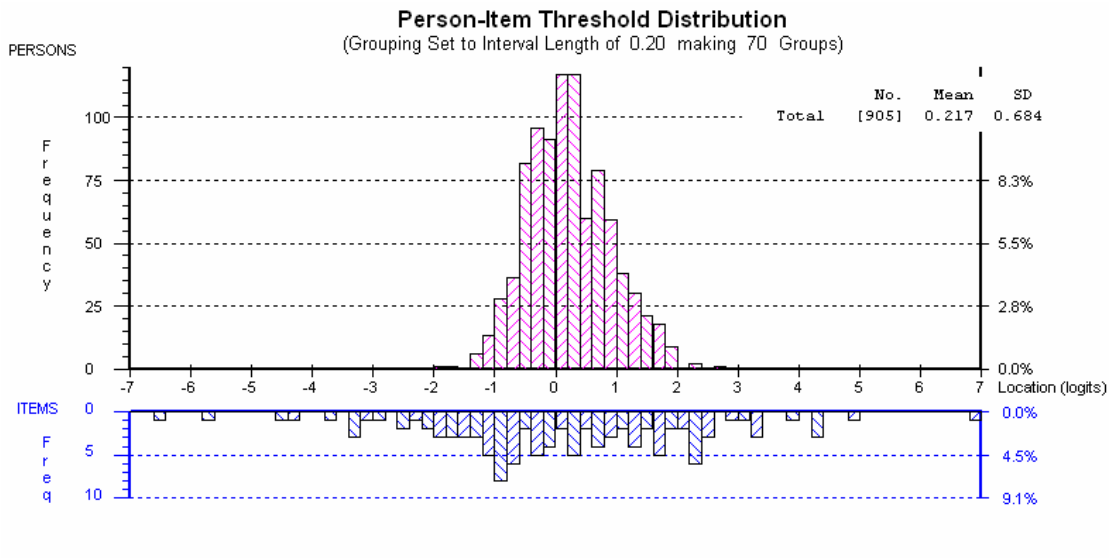


Figure 6b. Person distribution of the written component of drama.

Inferences from the examples for obtaining a TER

Below are the relevant inferences that are made from the mathematics example and the two case studies.

Inferences from the mathematics example in Figure 1

In drawing the inferences from the example of alignment of mathematics items classed a priori into levels that was shown in Figure 1, it is reiterated that 5 recognised mathematics education experts from Western Australia, well versed with the outcomes and in assessment, constructed the items and that the items were trialled before the final administration. The averages of the item difficulties at each level conform to the expected trend that the higher the level, the more difficult the item. Nevertheless, there is an item designated to be at level 2 which is more difficult than an item designated to be at level 6! This illustrates the following two points.

First, because the location of the *average* difficulties of items designated within each outcome followed the expected pattern, it shows that the experts understood the outcomes levels. It needs to be stressed that there was substantial debate and discussion in constructing the items and in designating their levels. It involved much more debate and discussion than could normally be given to the classification of an item.

Second, however, at the *finer degree of analysis*, specific items, and therefore the demonstrated performances of students, can vary very widely across the macro levels. The macro nature of the outcome statements and levels implies that it is difficult to use them for a micro assessment directly – they need further operationalisation Without such operationalisation, endless debate can ensure as to whether a

performance demonstrates one level or another for a precise analysis of performance. This can generate a great deal of anxiety in teachers and students. The features of Figure 1 are replicated in many studies.

> ***Recommendation 7*** *That in the literature from the Curriculum Council and in the professional development of staff, it is recognised that the classification of items and tasks into levels is inherently probabilistic and not deterministic.*

Inferences from Case Study 1 - the assessment of writing

A number of important inferences can be made from Case Study 1, including some that reinforce the point made above from the mathematics example.

First, consistency of classification into levels is not a sufficient condition for obtaining precision of assessment. Second, that assessing an outcome such as writing against generic level outcomes can lead to consistency at the expense of precision. Third, that it is difficult to construct criteria that follow the outcome levels and that are directly relevant to the task that has been set.

Fourth, therefore, the criteria should arise naturally from the task and that the scoring of performances into degrees of quality on the criteria (e.g. 0,1,2; 0,1,2,3,4) should arise naturally from the criteria rather than artificially from a generic outcome levels.

Constructing such criteria and scoring keys and marking according to them, is called in this report *analytic* marking. It is a common terminology that is contrasted with holistic ratings.

Such analytic marking is carried out routinely by teachers and therefore it should be relatively straightforward for them to continue to use it in constructing tasks that are guided by the outcomes statements. Some refinement and support might be given into constructing keys for tasks which themselves arise from outcomes, but the task should be substantially less demanding than the task of assigning levels, or sublevels, precise enough to be used for tertiary selection.

Fifth, on balance judgements (OBJs) of outcome levels and analytic marking can be carried out simultaneously with the cut-off points between outcome levels readily mapped on to the finer analytic marking scale. To use outcome levels in this way is compatible with the outcome statements being at a relatively macro level of specification. That is, the level or nodal descriptors should be used primarily at their level of specification, and not at finer levels than their original standing.

Sixth, the same joint analysis of the OBJ into levels and the finer analytic marking confirms the uncertainty in the classification of the levels, while at the time showing that each level characterises a region of development. This, too, is consistent with the generic nature of the level descriptors. It is important that these relationships, distinctions and complementarities in the levels of assessment in their purposes are appreciated. It is consistent with the analysis made in Andrich (2002b) and with the observation made with the mathematics example above.

Seventh, and while not present in this study, it can be inferred that analytic marking is more compatible with the kind of marking required for detailed student feedback on performance, rather than a statement about the level of achievement. The OBJ on levels can provide a general indication and can be reported to students, but this is not incompatible with a traditional analytic marking that arises from the assessments. As evidenced from both the above examples, level statements, even though relatively crude, are not without substantial debate regarding interpretation even by experts.

Eighth, therefore, it is further inferred that both school based assessment and the external assessment requires the usual fine grained analytic marking for purposes of assessment for tertiary selection.

> ***Recommendation 8*** *That for both external and school based assessments, analytic marking keys which arise from the tasks set be used in conjunction with classification into one of only 8 levels. That the former and relatively precise marks be scaled as required to meet the tertiary selection policies and be used for tertiary selection, and that the levels be used for other educational purposes at the generic level at which they are described, for example, for monitoring teaching and learning at a generic level.*

This recommendation is considered relevant for all areas of learning, from English to Mathematics. Indeed, as shown in the example of Writing, the case for analytic marking in the humanities type courses are no less necessary than in mathematics and science courses. Holistic marking is very susceptible to the halo effect.

Ratings relative to levels

One of the proposals for the school based assessment that is intended to relate the levels to the performances and attempt to give the required precision is to have a student performance assessed against four outcomes and then rated against levels into three further categories. Specifically, a student rated at level 6 for example, would be further rated into sublevels of such as 6^-, 6, 6^+ which would be given ratings of 6.2, 6.5 and 6.8 respectively. Similarly, a student might be rated at 5 and then rated further at 5.2, 5.5 or 5.8, and so on. Then the assessments on say two aspects from an outcome may be 6.2 and 5.8, and these would be averaged to 6.00 to two decimal places to provide an assessment for the outcome. This would be the basis for the school based assessments that would be provided to the Curriculum Council to be integrated with the external assessments.

Clearly, and consistently with *Recommendation 8*, I am recommending strongly against this process for all courses for the following specific and additional reasons. First, as indicated already in several places, the levels are generic, that is, general and abstract, and cover a wide range of achievement, both in breadth and in the range between levels. Therefore, the tasks that are provided for assessment will not fall naturally into the 3 sub-levels any more than they will fall naturally into levels on the various aspects that are assessed, or the outcomes to be assessed on different tasks. Further, even if two different tasks assess the same outcome, there is no guarantee that

the evidence of the achievement of the outcome, and the degree of its achievement, will be the same.

Second, and very importantly, this kind of approach gives the impression that the distance between levels is the same in some sense – that is, that the difference in achievement between levels 5 and 6 is the same as between levels 6 and 7. As revealed in Figures 2a, 2b and 2c, in a particular example, where the thresholds between levels are shown, this is not the case. It is not even the case in Figures 2a and 2b where the thresholds are relatively well aligned artificially. The raw marks that are obtained in analytic marking are summed *as counts*, not as equal distances, and then these counts are transformed through models to approximate measurements. It would be unfortunate if the impression is given that these *qualitative* distinctions between levels, which are ordinal, are somehow measurements which are equidistant. The impression that the numerical scores given to assessments do not need to be transformed to implement policies should not be promoted. *Teacher assessments need to be internally consistent and valid, but need not be in particular measurement units to the degree necessary for tertiary selection.*

Third, the issue is compounded by the application of the same generic values across outcomes within a course and between the courses. It has been noted already that there is an intention to make the levels across outcomes and across courses of the same order of intellectual demand. This is justifiable at the organisational level of courses as a general working framework for various purposes of teaching and learning, but it would be misleading to suggest that this can be achieved at the measurement level, especially at a precise enough level for tertiary selection and detailed student feedback. It would perpetuate a misunderstanding about the use of numbers in educational assessment and measurement.

Fourth, the generic descriptors seem to be for communication and understanding amongst teachers and experts in the field. They seem not to be the best descriptors for communicating with students, parents and the community. I believe this is the source of some of the unfortunate press – that the formal language used within the profession for its own communication, is considered the language for communicating with students, parents and the community. For example, the descriptors of levels 5, 6, and 7 shown in Table 3 seem not ideal for communicating with students without further operationalisation as would be carried out with analytic marking keys. Further, because of the uncertainties described earlier, attempts to assess directly against outcomes for purposes of precise assessment can be inordinately time consuming.

> ***Recommendation 9*** *That the language of outcomes and level descriptors be recognised explicitly as the technical language of the education profession for its own communication, and that analytic marking keys be used as the basis for providing feedback to students on their progress and for communicating to parents.*

Inferences from Case Study 2 - the assessment of drama

The assessment of drama involved both a performance and a written component, and therefore illustrates an issue associated with courses that have these components. The analysis summarised in this report only involved an external assessment.

The assessment combined ingredients of analytic marking and marking according to levels in that there were different criteria, 4 for performance with no choice and 5 further distinct criteria for the written component with some choice. In addition, the different criteria were scored with different numbers of ordered categories, reflecting the nature of the criteria. The marking guide was, therefore, substantially analytic. It confirms the points inferred from Case Study 1 but deserves further study from an analytic marking perspective. There was no OBJ for the level, but that could be readily added to the assessment, and perhaps should be included so that the fine grained analytic marking can be mapped to the generic level statements as in Case Study 1.

It can be inferred from this data set that for the same total score on the written and the performance, the performance component generally provided a higher score. This suggests that the performance component was easier. In this data set, then, if there was a policy that the two components should be weighted equally, a transformation of them would have been necessary to ensure that this policy was carried out.

This example can be extrapolated to the combination of assessments from external and internal assessments to give a course assessment and measurement. The school based assessment serves a number of functions in complementing the external assessment, including assessing content that is not amenable to being assessed externally, and also to provide more than one opportunity to demonstrate the knowledge, skills and understandings of a course. However, there can be no guarantee, a priori, that either the school based or the external assessments will be on a particular scale, that is that they will have a particular origin and particular unit, nor that they can be on the same scale. Therefore they both have to be transformed to the same scale.

The broader variable combining written and performance components

The summary score, including both the performance and written components in Drama reflects a *higher order* and *broader* variable than each of the components of assessment, the performance and the written. A student with a very high score on the total will have a high score on both, but there are students with high scores on one component and a low score on the other. Two extreme examples are a student who has 28/50 on performance and 41/50 on written and another who has 44/50 on performance and 22/50 on written. There must be some explanation for these extreme profiles, and they could be inherent differences in relative capacities of students on the two components, or there could be some other reason that needs to be taken into account.

With a low correlation of 0.43 between the written and performance components, they clearly are not just two assessments of exactly the same relatively narrow construct.

Instead, Drama is defined as a sum of two imperfectly related components to construct the broad variable.

An argument could be made that many of these students, or even most given that they have selected to study Drama, are genuinely excellent in performance, but that they are not as good at the written component. For this argument, other kinds of normative evidence would need to be produced with other selected groups and assessments that might provide the wider frame or reference. Within these data alone, such an argument cannot be confirmed. However, in considering this argument itself, the purpose of the summary score and its general frame of reference needs to be considered.

As summarised earlier in the report, the broader variable combining performance and written performances is to be used for the TER in tertiary selection. The TER is a ranking for tertiary selection and is used as the only criterion for many tertiary programs of study that do not set prerequisites, for example, Psychology, Law, and many others. Therefore, the policy that all courses used for obtaining a TER should have a written component that characterises the literacy skills of the student seems eminently reasonable. Making the written component and performance component scores each out of 50 indicates that they are intended to contribute equally. These are important and defensible policy decisions with implications for their implementation.

Finer analysis of the data indicates that to achieve the equal weighting, further adjustment of scores is required. This finer analysis cannot be generalised to assessments from year to year – in a subsequent year, the components might turn out to be somewhat different in different ways, perhaps not uniformly, or perhaps with the written component having higher scores than the performance.

If a host tertiary institution considered the performance component more important than the written component for selection into particular programs of study, it would need to make this explicit. This is analogous to providing prerequisites for particular tertiary courses. However, for the TER itself, the performance and the written need to be scaled against each other to implement the stated policy in weighting performance and the written component equally. Again, needing to do this is not a reflection of poor assessment or incompetence of the markers; it is simply the case that there are no well defined a priori measurement units in Drama relative to a single repeatable application of one instrument to various assessment tasks.

> ***Recommendation 10*** *That the Curriculum Council make explicit that its policy for courses that have both a written and a performance component, the written component will be weighted not less than 50% for contribution to a generic TER score. That it also negotiates with tertiary institutions to provide disaggregated scaled scores in the performance and written components in the cases that the performance based component is considered more important by a tertiary institution for a particular program of study.*

Further issues on aspects of assessments

Some further issues on the formalities of assessment have not been canvassed in the report, and are considered briefly here. They are concerned primarily with aspects of external assessments although a point is consolidated concerning the relative emphasis on the levels and analytic assessment and a possible implication for teachers' workload.

School based assessments

As already indicated, *for purposes of tertiary entrance selection,* analytic assessments should take precedence over outcome level classifications, and the levels assigned to students should be considered relevant at the general levels at which they are specified, that is, in general monitoring of the teaching and learning. They are too crude for refined assessments. Teachers are used to constructing analytic assessments so this should not be a major imposition on them. The effort should go into ensuring that they understand the content and principle of the new courses, and be given support, where required, in generating analytic marking keys based on the courses and content. They should not spend inordinate amounts of time assessing the level of a student, or degrees of a level, directly. This requirement of analytic marking is not seen as an extra task, rather as the main task as it is at present, and that in addition, general levels assigned to students are non-high stakes for them individually, and not overly time consuming for teachers or the Curriculum Council.

It is most important that the school-based assessment does not become so formalised and excessive that it overwhelms the teachers and students, and in the process, distorts the teaching and learning even more than does the external examination. The process recommended should not be much more demanding on the teachers than the present process, while achieving sufficient precision for tertiary selection. The extra demands are that the marks need to be submitted on a unit by unit basis as well as on the basis of the whole course.

External assessments

External assessments generally take the form of a written examination for various aspects of efficiency, but not necessarily so, as exemplified by the assessment of Drama. The comments to follow are concerned with a written external examination. In the various Curriculum Council documents, there are discussions that the written examination might be two hours rather than three hours. This may be appropriate in some courses but it needs to be considered carefully, because the shorter time may cause greater anxiety and be less valid than a longer time if the students feel rushed.

The examinations in the courses are not expected to be speeded as such. Often, tasks in examinations take too long just because there are many of them, and speed effectively is used to separate students more than is perhaps intended. Speed does not have to be invoked in such a way artificially, and the separation of students should arise from increasingly more challenging questions relevant to the content of the course. It is recommended that the examinations be set so that good students can

finish most of the paper in the time, even if they are given strong challenges in the questions that they would complete in the last half hour or so of the examination. Of course, some questions of this kind might take students longer, but the bulk of the paper should be completed by good students in good time. Perhaps there might be a designated section at the end of the paper that is explicitly more challenging. If an examination is to be two hours, then the same principle should apply. The above points are made simply as points for consideration.

The courses are proposed to have six units. Different students may take different combinations of these units. In order to facilitate the performance of students in the examination, and to not generate irrelevant decision making in the examination, the unit or units relevant to each question should be specified for the student. Reading time before beginning the examination should be retained for all the reasons that it is in place now.

> ***Recommendation 11*** *That the questions in the external examination show the unit or units of study for which the question is most relevant.*

Finally, the Curriculum Council should consider using open book examinations in all external assessments where these are possible. This would make the examinations more valid and less threatening on irrelevant details. Having closed book examinations is tantamount to permitting an examination to be no more than a memorisation of material from a text book or notes. Examinations that are open book need to be more creative than closed book ones, and wherever appropriate such examinations should be carried out. The effect would be fed back to the schools, where the teachers would need to prepare students for such examinations by conducting them. Open book examinations enhance learning by going beyond rote memorisation of details. This could be a major innovation in upgrading the validity of the external examinations and making them compatible with the outcomes in the courses. The Curriculum Council should not fear the press on this issue - universities now use open examinations and it is easy to demonstrate that the learning associated with an open book examination, where relevant, is enhanced compared to the use of a closed book examination. Indeed, given that the external examination is worth 50% of the final mark, it seems essential that the Council moves to such a form of external assessment.

> ***Recommendation 12*** *That the Curriculum Council has external examinations that are generally termed "open book" and that in consultation with assessment panels, course experts and teachers, it makes explicit the materials that can be taken into the examination. In principle, the restrictions should be a minimum consistent with sound learning and assessment practices. It is recommended that the Council embarks on demonstrating the case that in general, and in each course, such an assessment is more valid than a "closed book" examination.*

Consolidation of some policy implications for assessment

Some of the recommendations regarding analytic marking and its application for tertiary selection may appear very much like the status quo. However, this is governed in part by the outcome levels descriptors being very generic – 8 across the

whole spectrum of 12 years of schooling which would be difficult to make less general, and in part by policies that have also retained the status quo. As indicated elsewhere, it is considered that this is only one aspect of the move to OBE, and that the analytic and level assessment can and should complement each other, with the use of the generic level statements governing the course design, teaching and learning, and the analytic marking used for detailed student feedback, communication, tertiary selection and communication with parents and the community. These are not seen as incompatible, but complementary, each enhancing the other when used for the purposes for which it can be used.

The differences from the present organisation of Years 11 and 12 studies seem to arise from the differences at the organisational structure of the courses, and the teaching and learning, not at the level of assessment in general or for tertiary selection in particular. As indicated early in the report, it is not considered that OBE is characterised by the methods of selection that are relevant for tertiary selection in Western Australian.

Thus if particular policies for tertiary selection are the same, then it is inevitable that matters directed by the policies will remain the same unless they were implemented inadequately. I do not believe they have been, and indeed they would be the envy of many places around the world.

To be explicit, the first policy that has been retained is that students are competing directly for specific programs of study at the tertiary level. This is not directly under control of the Curriculum Council or the universities – it is a Federal and State cost issue. Further, the TER as the generic indicator for programs of study at the tertiary level which have no prerequisites, or as a baseline indicator for those with other prerequisites, has been retained. There are again many good reasons for this policy when compared to potential alternatives. The main change is that the number of courses eligible specifically for university selection within the broader framework of tertiary selection has been increased.

Any specific accounting of differences in requirements of different courses at Years 11 and 12 for different tertiary programs needs to be made by the tertiary institutions by making specific qualifications to the generic TER. For example, additional requirements may be imposed, or particular courses may be required as now for some tertiary programs. Similarly, with the advent of many new courses that can be taken to obtain the generic TER, tertiary institutions should also consider the disaggregated scores of various components that compose the TES, and hence the TER, for example the school based component or a performance component, for particular programs. This needs to be negotiated at a different stage from the development of the TER itself, perhaps with the Curriculum Council and other relevant groups, recognising that the general TER cannot cater for every possible contingency of this kind. However, for consistency and fairness, disaggregated scores in a course used for selection should be scores that are scaled to the same origin and same unit as those in other courses.

The second policy that has been retained is that different courses will not have different intellectual demands so that it is in principle easier for the same students to obtain higher scores in one course compared to another course for no other reason

than that one is in fact easier than another. This, of course, is a sound policy. However, it immediately has the implications that the measurements on the different courses have to be on the same scale, and therefore assessments need to be transformed to this scale. This too is the current policy. The alternative of ensuring equivalent marks only by moderation and using only ratings into levels or sublevels, I consider, would be far more draining on resources, and would not achieve the credibility of the present process. Indeed, by having data on both levels and on the analytic marking, it will be possible to verify and check the degree to which the intention that levels are equivalent across courses has been achieved.

The third policy that has been retained is that the school based assessment and the external assessment will be weighted equally. This has implications that the assessments from the two components must have the same order of precision, and that the measurements derived from them must also be on the same scale. Both assessments must be transformed to ensure that they are on the same scale before this policy of equal weighting (or any other weighting) can be effected. Because of time constraints towards the end of Year 12, and to avoid skills in negotiation on the part of schools as a factor in the scores for selection, it seems that scaling of the two against each other at this point must be done automatically. Thus even assuming that all the school based assessment scores are on the same scale, they and the external assessments need to be rescaled to the same origin and the same units to ensure that the policy of equal weighting has been applied. Teachers must not be given the impression that the marks they submit will not be rescaled and it must be explained to them why they must be scaled to implement policies. This is part of the professional development in *Recommendation 4*.

Given that the school based assessments and the external assessments need to be transformed to the same scale, the only question is whether this transformation is to be done on the basis of a school and clusters of schools which work together, or with the assessments across all schools assumed to be on the same scale. If the latter, for purposes of equity and efficiency, this would have to be checked – it could not just be posited or assumed. I believe, however, this approach would be unwieldy. As now, the organisational level of checking, that is which schools and clusters of schools will be assumed to have the same scale, will need to be made explicit in advance. Once again, it points to retaining the present process on this issue. It can, as recommended in *Recommendation 6*, be enhanced.

Again because of time constraints at the end of Year 12, the policy should be that the school based assessments will be scaled automatically at whatever organisational grouping of schools is considered to have their assessments on the same scale. Further, if the external examination is an open book examination, it will be the most valid kind of assessment for this purpose of scaling.

As now, with experienced teachers and with those who give support to less experienced teachers, the school based assessments should not vary from external marks by large amounts – that is, minimising changes in marks through scaling is a matter of teaching experience and opportunities for teachers to work with each other. This feature of professional assessment is not something that arises out of OBE or any other approach to organising teaching and learning, but something necessary to place assessments on the same, relatively arbitrary, scale. If there were no scaling, and

differences in marks were only of the order of 5 marks out of 100 for each course but they happened to be in the same direction for different courses within a school, the effect would accumulate to have a very tangible effect which could be exploited in different ways. I am sure that teachers would not sanction such circumstances.

> ***Recommendation 13*** *That final school based assessment mark and the external assessment mark in a course both be scaled routinely at the end of Year 12 studies to ensure that they are on the same scale before being combined.*

If a general achievement test is considered for purpose of scaling using the principles outlined above, then it will be imperative that the score on the test contributes to the student's final TES. If it does not, then students will not take it as seriously as it is required for the purpose of scaling. On the other hand, if it is used only for the purposes of monitoring, then maybe it would not need to be part of the TES.

The automatic scaling of assessments at the end of Year 12 and the requirement for a general test, if used, to contribute to the marks, do not have a separate recommendation. The former is implied in *Recommendations 4 and 5*. The latter is not given a separate recommendation because it is considered that it can be used only for general information and not for actually scaling students' scores.

Conclusion

This report is written in a style of critique of assessment practices and a strongly argued case for marks arising from analytic assessments to be used for tertiary selection. It is stressed that it is not a criticism of the introduction of OBE in principle, or of the other reforms instituted by the Curriculum Council. Instead, it is considered that the type of assessment recommended is compatible with the OBE organising framework for teaching and learning, and elaborates it further at the micro level of assessment. The analysis is an extension of the analysis and my commentary on the OBE framework that was commissioned earlier by the Curriculum Council.

References

Andrich, D. (2002a) A framework relating Outcomes Based Education and the Taxonomy of Educational Objectives. *Journal of Studies in Educational Evaluation.* 28, 35-59.

Andrich, D. (2002b) Implications and applications of modern test theory in the context of outcomes based education. *Journal of Studies in Educational Evaluation*, 28, 103 – 121.

Andrich, D., Rowley, G. & L. van Schoubroeck, (1989). *Upper Secondary School Certification and Tertiary Entrance Research Project.* Report commissioned by the Minister for Education in Western Australia, Dr Carmen Lawrence.

Andrich, D. & Mercer, M. A. (1997) *International Perspectives on Selection Methods of Entry into Higher Education.* Higher Education Council, Commissioned Report No. 57. National Board of Employment Education and Training.

Department of Education and Training (2005) *Assessing Students' Writing Years 3, 5 and 7* Western Australian Literacy and Assessment Program, Department of Education and Training, Perth, Western Australia

Bloom, B.S. (Ed) (1956). *Taxonomy of Educational Objectives.* New York: David McKay Co. Inc.

Curriculum Council (1998) Curriculum Framework. Curriculum Council. Osborne Park, Western Australia.

Heldsinger, S. and Humphry, S. (2005). *Assessment and evaluation in the context of an outcomes based framework.* Paper presentation, School of Education, Murdoch University, May, 2005.

Pascoe, R. (2002) Critically examining drama: What does the examination of Year 12 drama students in Western Australia tell us about learning and teaching drama? Drama Australia conference, Brisbane (unpublished).

Pascoe, R. (2004) Finding common ground: Drama, Theatre and Assessment. (Unpublished paper).

Tognolini, J. & Andrich, D. (1996). Analysis of profiles of students applying for entrance to universities. *Applied Measurement in Education, 9*(4), 323-353.

Van Wyke, J. (1998) Personal communication.

Appendix : Correspondence regarding the terms of reference for the report

Our Ref: CS/0006

Professor David Andrich
Dean of Education
Murdoch University
South Street
MURDOCH WA 6150

Dear David

As discussed with you early in June, I would like to extend the work that you are doing for the Curriculum Council as part of the measurement expert group to prepare a formal report on the comparability of the standards for the new courses.

The Minister is very determined that the highest possible standards will be maintained in the new post compulsory education system as well as providing choice and flexibility for all of the students who will be staying on in years 11 and 12. To ensure that this happens, I would like to formalise our verbal agreement for you to prepare a report and advice on the standards in all of the courses.

The aim will be to ensure that:

- the assessment process of each course has sufficient rigour to enable the highest academic standards to be maintained;
- assessment is such that the fine grained measurement of student achievement is valid and reliable particularly where university entrance is involved;
- the measurement processes being developed will enable comparability of standards between courses and enable statistical adjustments to be made if necessary.

At the same time, the Council secretariat is investigating the issues associated with school-based assessment and is putting into place measures to ensure that teachers are well supported in this area.

My understanding is that you will have time available to undertake this task from early August. To assist with the process, the secretariat has been continuing exploratory work on external assessment and on testing the various models with the assistance of members of the measurement expert group, to ensure that the information you will need for your task will be readily available when needed.

As agreed, payment will be at the recommended hourly rate for an academic with your level of expertise and experience. The timeline for the report is the end of August. I look forward to working with you on this project

I would like you to report to me by the end of September.

Yours sincerely

NORMA JEFFERY
CHIEF EXECUTIVE OFFICER

6 July 2005