

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

Executive Summary and Recommendations

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 research in assessment in general and in writing in particular. The data for Case Study 2 was 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; and
- 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 1).

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 course unit, 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 the five generic levels of achievement that arise out of the standards of the outcome statements of a course 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 constructing a tertiary entrance rank 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. The two assessment processes, distinguished by their level of precision and relevance, are complementary 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.

Overview of approach to the report

The approach taken to address the terms of reference has three features: it presents *principles* of assessment rather than detail in assessing each course; it considers assessments that would be sufficiently rigorous and *fine grained* that they can be used to form a tertiary entrance score and rank for selection into tertiary programs of study; its theme is that if the policies of combining (i) school based and external assessments and (ii) different courses to form a Tertiary Entrance Rank (TER) are to be implemented correctly, then the various assessments must have the same order of precision, and measurements arising from them must be on the same scale.

The report is set in the context of the structure of *outcomes based education* (OBE) as articulated by the Curriculum Council (1998) and its reforms for senior secondary education. The report provides recommendations with a rationale for the kinds of assessments that should meet the requirements of rigour and precision necessary for constructing a TER. It is assumed that OBE is a much wider set of educational principles that is not characterised only by the nature of assessments that need to be used for competitive selection in tertiary programs based on a TER.

Terminology

The term *assessment* emphasises the stage of design, administration and marking of student performances elicited by the assessment tasks. One kind of marking of performances involves what is termed *analytic marking* in the report. Analytic marking requires *marking keys* which have two properties: first, there is a marking key for each of the distinguishable criteria of the performance to be assessed; second each marking key has two or more ordered categories showing increasing qualities of achievement. The number of ordered categories varies from marking key to marking key depending on the number of meaningfully distinguishable ordered categories

possible for each criterion. Analytic marking is distinguished from the second kind of marking which involves *ratings* of performance into levels of achievement presented in the outcomes and aspects of outcomes, which may be made directly or indirectly using indicators of achievement or marking keys as a basis.

The term *measurement* emphasises the scoring of these assessments in a numerical form and their *transformation* into a quantitative scale using statistical models. In analytic marking, the scoring involves two steps: first assigning successive integers, beginning with 0, to the ordered categories reflecting increasing qualities of achievement, and second, summing these across marking keys to give a single number for the performance as a whole. In ratings into levels of achievement, scores are immediately taken as indicators of a specified a priori standard. The feature that distinguishes analytic marking from ratings is that the former has criteria and numbers of categories for each key that arise out of the task, whereas in the latter, the criteria and number of categories are determined a priori to be the same across tasks for all outcomes for all courses. The stage of measurement ensures that the assessments to be combined are commensurate in their precision and in the scale on which they are expressed. The concept of commensurate scales is elaborated in the report and briefly in this summary.

Competitive entry and its implications

The references to *university* and *tertiary entry* need to be understood to imply entry into particular programs of study, and not to tertiary institutions as a whole, which raises substantially the competitive edge in selection. For students on the margin of selection into any program of study, the competition can be as fierce as for those competing for places in the high profile courses such as medicine and law. Therefore, the selection process must be consistent and fair across the spectrum of achievement. In particular, it is required that different courses not be inherently different in difficulty for the same students to obtain scores for competitive tertiary entry.

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 on schooling than it does now. Therefore, this report is written from the perspective that the credibility in the following two Curriculum Council policies is paramount: (i) the combining of school based and external assessment, and (ii) the combining of scores across courses to form a single TER.

Recommendation 1 *That further 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, 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.*

It is most important that Curriculum Council officers and principals see this selection process in the broader context of these policies and that they can share these with their staff, students and the community.

Key principles of the structure of OBE relevant for assessment

Courses correspond to present subjects. At Years 11 and 12, each course has units 1A, 1B, 2A, 2B, and 3A and 3B. Students requiring a TER need to complete at least two units in a course, though most will complete four. Most courses generally have only four outcomes which therefore makes them general and abstract. However, they are elaborated through *aspects* which describe the skills, understandings and knowledge which underpin the outcomes.

The outcomes in all proposed courses for Years 11 and 12 are divided explicitly into exactly 5 levels on an achievement continuum, which are the top 5 levels of the 8 levels which cover the full range from Years 1 to 12. The aspects are also divided into the same number of levels. Inevitably, therefore, the descriptors of these levels, too, are general, abstract and span a broad range of performance. To support teaching, learning and assessment, annotated work samples and professional development are provided to help interpret these general levels of achievement.

Four stages of assessment require aggregation: (i) within each unit, school assessments across outcomes from at least two formal assessments; (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 single Tertiary Entrance Score (TES) and a TER.

The report is concerned with the assessments and processes of these aggregations.

Some expectations across learning areas

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 in Reading. Further, the corresponding levels across courses are expected to have the same standards. Thus ratings against levels summarised above are expected to be of the same intellectual standard across outcomes and across courses.

Although the equality of standards across courses is a reasonable goal in organising teaching and learning and making general assessment of student progress, it is unlikely to be precise enough, even if each level is further divided into three sublevels, for purposes of constructing a TER. The use of analytic marking keys can be used both for more precise assessment necessary for constructing a TER and to help monitor the ratings against levels of achievement. The degree of precision of assessment is directly relevant to ensuring that explicit policies of weighting school based and external assessments within a course and having equal difficulty in obtaining scores across courses are implemented correctly. From the perspective of this report, the analytic marking and the rating into levels are considered complementary, with the former more precise than the latter.

Amount of assessment

At present it is proposed that there be 6 discrete units in a course, within which a typical student would take 4 and must take at least 2 before being eligible 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 four outcomes within a course, this indicates that 16 marks of school based assessment will contribute towards 50% of the assessment towards a TER. At present, there is only one such mark required, and the external assessment will continue to also have just one mark. Moreover, by the end of first semester of Year 11 a school will have to submit 24 marks for a student enrolled in 6 courses. This seems to be a greater amount of assessment than is necessary to achieve the degree of precision required.

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 as many assessments as proposed, especially in terms of levels, in order to provide them to the Council, and having these immediately have a high stakes element, will add considerably to the administrative burden of assessment for the Council and for the schools.

The Curriculum Council needs to consider whether this amount of formal assessment for monitoring purposes is justified, or whether using just one analytic mark per unit which is precise enough for constructing a TER and one outcome level rating per unit is sufficient.

***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 5 levels to describe the generic student achievement.*

The accumulation of marks, which can be used with analytic marking and with ratings into levels, is used to motivate study and sustained learning. However, it is important for the Council to recognise that a student's standing in the last unit, usually after semester 2 Year 12, reflects a student's final location on the achievement continuum and should be the location used for selection for further study. The marks from both units used for constructing a TER will tend to be compatible, although it seems almost by definition that the levels will not be comparable. Therefore, averaging the marks, or even weighting in a particular way a priori, will not always be just. The decision of the final standing of a student seems best decided at the school level, where all details and contingencies of performance are available.

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 level to the Council for the course as a whole.*

In most cases the analytic mark submitted will be the same as the mark that arises from the last unit studied, but the recommendation permits schools to routinely vary this mark in the case that such a variation is warranted and can be justified. This is the mark that should be used towards the TER.

Scaling and equating

Common elementary measurements in every day experience and in the physical sciences have a well defined origin and an arbitrary but well defined unit. This arbitrariness of the origin and unit is understood by children in primary schools and is part of the mathematics curriculum.

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. Compounding the irony is that there is a tendency for greater belief in the consistency of origin and units in social measurement than there is in physical measurement where their arbitrariness is made explicit. Numbers assigned to characterise degrees of a construct are used as if they had the properties of physical measurements. Cementing the irony is that there are abundant examples of quantification in the social sciences that lend themselves to this study, including of course assessment and measurement of student achievement and the current TER process itself.

In the case where different measurements need to be compared or reconciled, it is necessary to convert measurements from different assessments onto the same scale. The terms *equating* and *scaling* are used for such a purpose with the term *scaling* used in this report. In terms of the current TEE system the process of scaling termed statistical moderation places marks from different schools on the same scale. The process of scaling then places school marks and external assessment marks within a subject on the same scale. Finally, further scaling places the marks from different subjects on the same scale.

The principle behind scaling in the social sciences is that 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, that is, standard deviation. If two sets of measurements on the same group do not have the same average and standard deviation, and without independent other information, it is taken to reflect that they do not have the same origin and unit and therefore are not on the same scale. 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.

***Recommendation 4** That further 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 measurements

When two measurements are averaged, they define a construct which is more or less broad – to the degree that the content assessed in each is different, to that degree the average assesses a *higher order* and *broader* construct than either measurement alone. Thus the average of school based and external measurements assesses a broader construct than each measurement alone, and the average of measurements of different courses assesses a broader construct of educational achievement than the measurement of any course alone.

However, before measurements can be averaged, they need demonstrably to be on the same scale – that is, they need to have the same arbitrary origin and unit.

Policy implications for scaling and equating

The requirement of scaling arises from policy issues. Thus, for example, it is a policy issue that a course will be sufficiently broad to include both a practical and a written component and that these will have a particular weight. It is similarly a policy issue that each course will include both school based and external assessments and that these will have equal weight, and it is a policy issue that all courses are to be given equal weight in forming a TER.

Thus the scaling of school based and external assessment is essential to ensure the policy of 50% weighting of each, and the scaling of courses is essential to assure students that a higher measurement is not obtained in one subject compared to another just because the units and origin of the scales are different.

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.

***Recommendation 5** That further 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.*

Taking the average and standard deviation as points of reference requires that the numbers of measurements are sufficient. The minimum number required at present in

a school is 10 before the teacher needs to work with a teacher from another school. This minimum should be greater, perhaps 15. The benefits for teachers of small classes in interacting with other teachers in other schools teaching the same course is essential to providing consistent assessments for that group of students.

However, requiring consistent measurements with possible *differences in origin and unit of scale* from other schools or designated groups of schools which can therefore be scaled to each other, is a much less demanding task than requiring assessments to also be immediately on the same scale without common practices in the teaching, learning and assessing. In addition, it is impossible for the external assessments to be on a specified a priori scale which will be the same for all courses and on the same scales as the school based assessments.

The benefits of teachers with small groups working together to ensure a common assessment scale goes beyond just this provision. The benefits include giving the teacher and students confidence that their teaching and learning is equivalent to those in other schools. I understand that it is intended to provide more resources for support in school based assessment. A potent contribution would be to fund some staff release time for those who have to work with staff from other schools.

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 continue. This is assumed and not considered a special recommendation.

Consistency of classification and precision

Clearly, in any assessment and measurement, there is a need to have both consistency and a high enough level of precision for the task at hand. In particular, in some aspects of monitoring progress associated with OBE, there is a premium placed on teachers classifying students into levels. However, these classifications have inherent elements of uncertainty which may be too great for certain purposes.

An example of relative difficulties of mathematics items derived from outcomes

The uncertainty in assessing levels of achievement is illustrated using items constructed to assess mathematics achievement. The items were developed to monitor the progress of the students in the West Australian education system specified in terms of the 8 levels against mathematics outcomes. Generally, it would be expected that the higher the level of an item, the more difficult the item. The empirical pattern of relative difficulties shows that the *average* of the difficulties of any particular level is greater than the average of the previous level. However, the locations *overlap* considerably so that, for example, an item designated as level 2 was more difficult

than an item designated to be at level 6. Figure 1 of the report is produced below. This is a consistent pattern found with such data and indicates that classification into levels of achievement has considerable uncertainty for assessment.

It is also consistent with the generic nature of the level descriptors that the locations of items follow the levels *only on the average*, but that at a finer level of precision, the locations of individual items vary considerably from the levels. This point is returned to later in the summary.

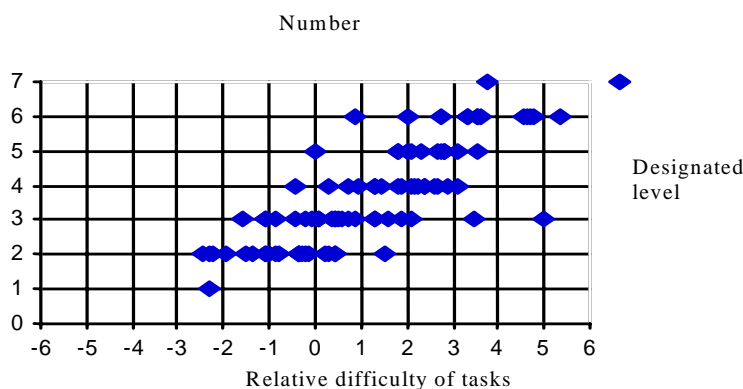


Figure 1 Relative empirical difficulties of tasks designated a priori at particular levels.

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

Potential sources of artificial consistency and imprecision

Consistency of classification can be achieved readily at the expense of precision of measurement. Such consistency is termed *artificial*.

Three sources of artificial consistency are elaborated in the main report as part of Case Study 1. First, the classification system into levels is relatively crude; second, the criteria and classifications are abstract and general and do not arise from the features of the assessment task; third, there is a halo effect across aspects of assessment.

Case Study 1 – the assessment of writing

Case Study 1 involved 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.

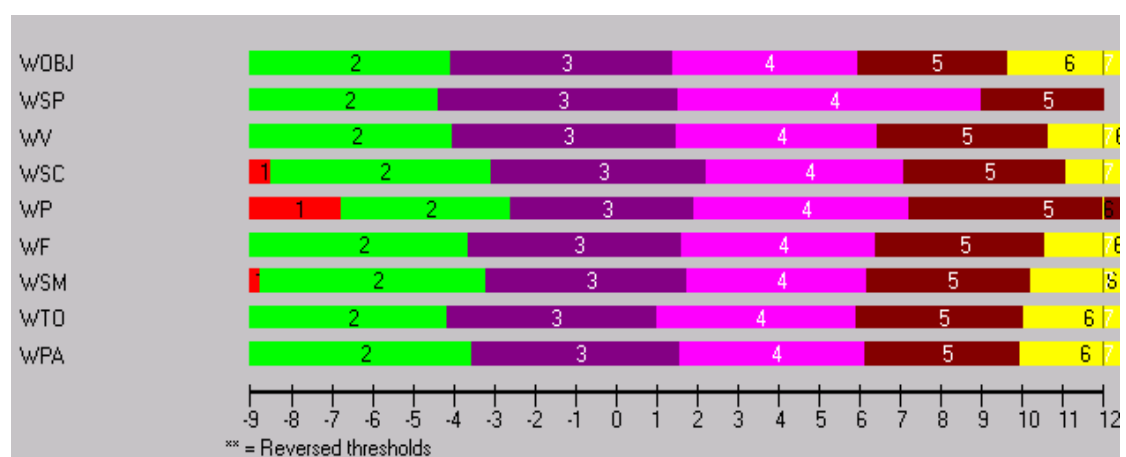
In the initial marking guide, the criteria were aligned *directly* with the *levels* of the outcome statements so that a score of 2 represented level 2, a score of 3 a level 3 outcomes, and so on. There was also an on balance judgement (OBJ) of the writing as a whole into one of 8 outcome levels. The data showed symptoms of artificial consistency. The halo effect was removed but there was still substantial evidence of further artificial consistency.

Heldsinger, Humphry, and the MSE program analysed the marking key and *it was evident that in assessing the performance on the particular task, that the generic framework of outcomes with the same number of levels on all aspects was deficient in omission of some features and commission of others*. They removed the resulting semantic and logical overlap and ensured that relevant aspects of the actual writing task were assessed into degrees of achievement that were distinguishable by the markers. This produced an *analytic* marking key. The original and revised marking guides are provided in the report. They include an on balance judgement of assigning the writing to one of the original 8 levels.

Conceptualisation of the levels as marks on a ruler

In the original classification system the levels were the same in number for each aspect. The statistical analysis of the data makes it possible to examine the relationship of these levels. 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 which mark the ranges of the corresponding levels on the achievement continuum.

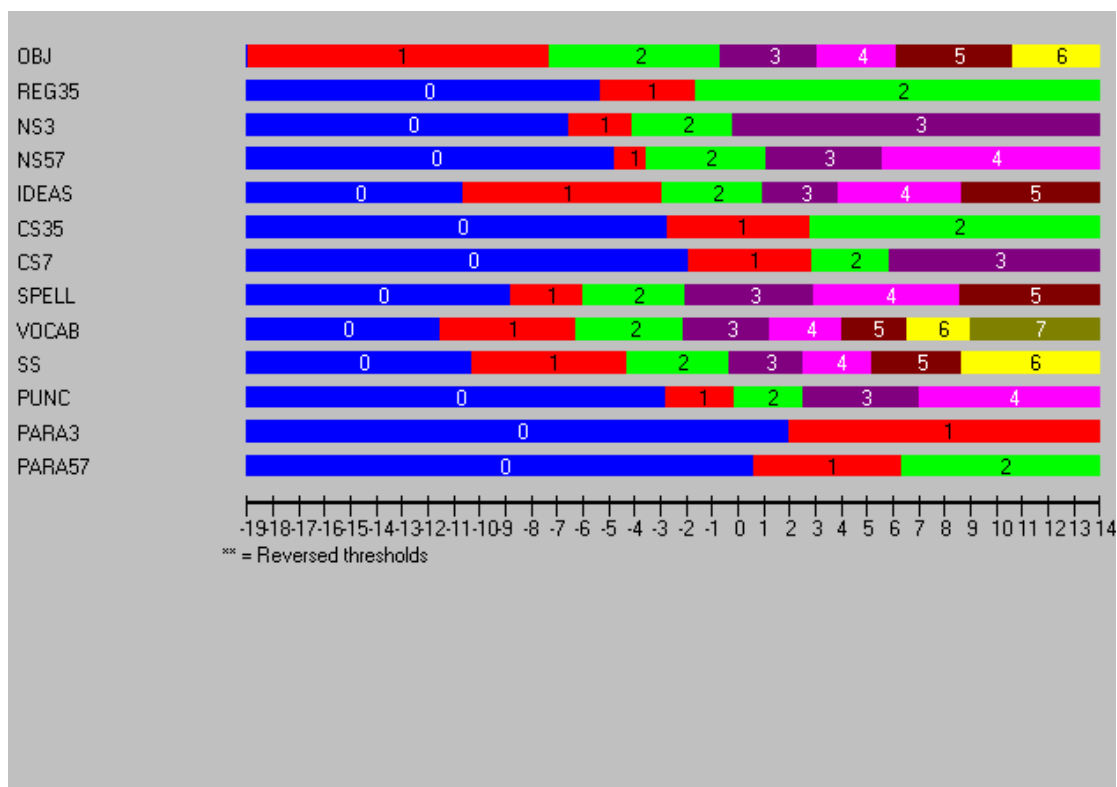
Figures 2a and 2c of the report are reproduced to show the threshold alignment for the first and third of the three analyses.



No reversed thresholds

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

Report to the Curriculum Council Executive Summary



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.

It is evident from these figures that in the first of these cases (a), the thresholds were very much aligned and that in the third case (c) they are very much non-aligned. *However, the alignments in the first case reflected artificial consistency which worked against precision.* In the third case, as is shown below and explained more fully in the full report, much greater precision of measurement was obtained.

Figure 2c also 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. For example, a score of 1 on the criterion REG35 corresponds to level 2 on the outcome statements.

This inference of artificial consistency evidenced in Figure 2a was confirmed by the distribution of persons relative to the distribution of thresholds for each of the three analyses referred to above. Figures 3a and 3c of the report are reproduced showing that there is considerable clumping of the person distribution in parallel to the clumping of the threshold distribution shown below the horizontal axis. The precision was not uniform across the continuum of achievement, and where there were no thresholds, the measurement was imprecise.

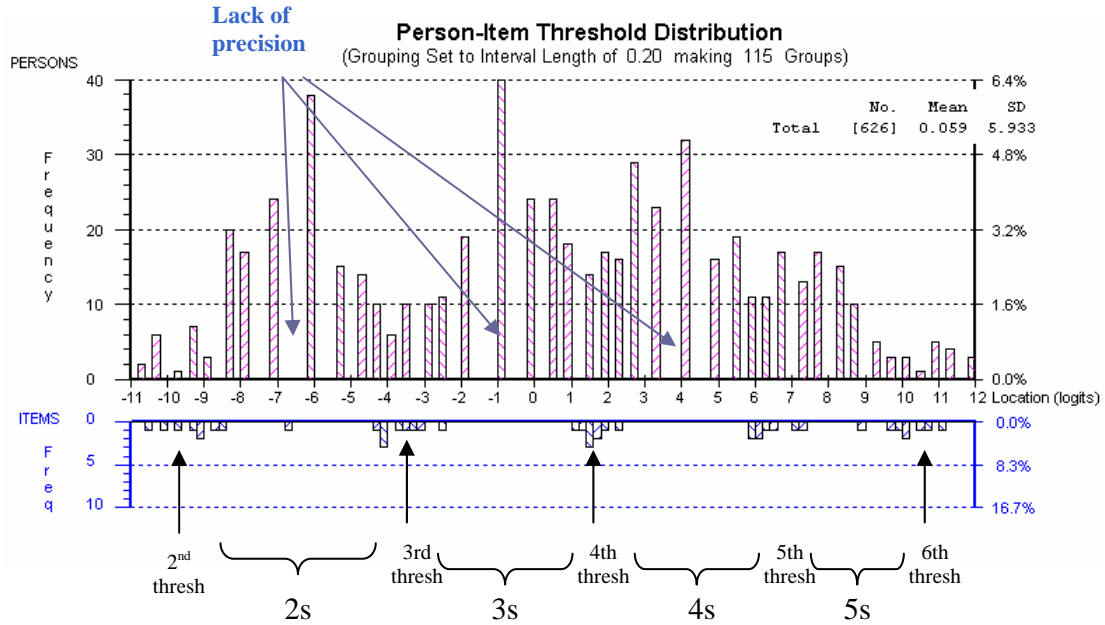


Figure 3a Person distribution and summary threshold map for original analysis when there is explicit alignment among categories. Thresholds are below the horizontal axis.

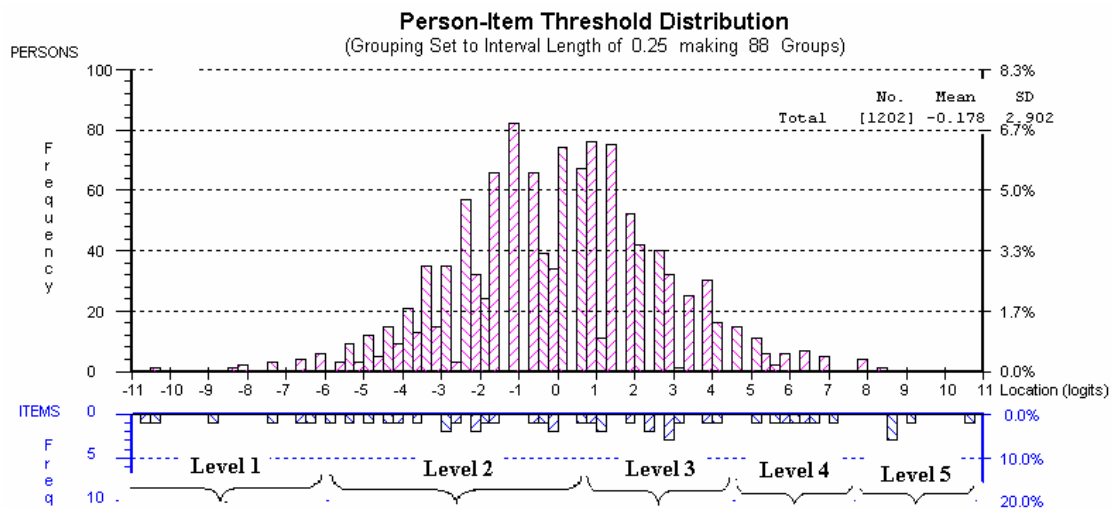


Figure 3c Person distribution and summary threshold map for the analysis when different aspects had different criteria – thresholds not aligned.

Observations on the assessment process

In Figure 3c, the OBJ classification in terms of successive integers is with respect to the original outcome statement level; the classification in the other criteria into *ordered* categories is also scored with successive integers but they are *not* based on outcome levels. It is evident from Figure 3c that with analytic marking, any particular student's performance can be located at a point *throughout the continuum*, and not only at the distinct locations far apart on the continuum. This is consistent with the level descriptors being general and abstract and characterising a range of achievement and not achievement at a point.

The thresholds mark off the region where the particular level *is most likely to be given*. Converted to the analytic scores, an OBJ level of 2 is most likely to be given to a student who has analytic scores between 6 and 19 and a level 3 to students with analytic scores between 19 and 27. It is emphasised that the level classification is not made with certainty, but probabilistically relative to the analytic marking. *Because the threshold between levels 2 and 3 is at the raw score of 19, in theory a 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 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 involve uncertainties. This uncertainty in assessing the writing outcome is consistent with the uncertainty in categorising mathematics items illustrated in Figure 1.

As indicated above, the marking keys for the different criteria of a performance on a task are in ordered categories and markers rate the criteria into one of the ordered categories. The raters do not need to be involved with the step of scoring, though it simply involves assigning a successive integer (0,1,2,...) to the successive categories. A special case is when the number of categories is just two, giving possible scores of 0 and 1.

The task set for students arises from the outcome statements and designated levels of achievement, and then the marking keys arising from the tasks, take this construction a step further. They build on and elaborate on the outcome and level statements but are also compatible with the tasks set, rather than being only generic. The analytic marks make possible assessment at a finer level of scale than possible with the levels themselves. This finer degree of assessment arises in the Case Study 1 data primarily because the markers are focussing on the categories within the criteria which are not generic, and therefore eliminate artificial consistency. In other situations it could arise because the number of ordered categories within a level can be made greater than just two or three. However, as found in Case Study 1, the number of a priori distinctions is not the only basis of obtaining greater precision – it is the way the categories across marking keys work together which provides the precision. The marking key for each criterion should have as many ordered categories as is possible for the markers to distinguish.

Analytic marking keys which arise out of the assessment tasks are constructed now with varying degrees of rigour. Constructing analytic marking keys that arise from

the tasks and are also compatible with the outcome statements and the level descriptors, can make the analytic marking keys substantially more rigorous.

Recommendation 8 *That for both external and school based assessments, analytic marking keys which are compatible with the tasks set for students be used in conjunction with a classification into one of 5 levels or sublevels permitting the mapping of the former onto the latter, and that the Council supports staff in developing such marking keys for school based assessments. Further, that the former marks be scaled as required to meet the policies for constructing a TER.*

Ratings relative to levels

Recommendation 8 means that I am indicating that the ratings into levels of achievement and sublevels, such as 6.2, 6.5 and 6.8 or 5.2, 5.5 or 5.8 not be used directly for purposes of constructing a TER, but be elaborated in conjunction with the specific tasks set in order to provide scores at a finer level. There are three complementary reasons for this recommendation.

First, because the levels are generic and cover a wide range of achievement the assessment tasks will not fall naturally into the 3 sub-levels any more than they will fall naturally into 5 levels. However, because it is important that the levels across schools also be commensurate within courses, the results from analytic marking obtained in conjunction with levels or sublevels can be used to help monitor the latter across schools. It is important that the Council provide support for analytic marking as it does for assigning levels.

Second, use of levels directly can give 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. This is not the case at the level of precision required for constructing a TER.

Third, and very importantly, the generic descriptors of outcomes and levels seem to be appropriate for communication and understanding *amongst teachers and experts in the field*. It is relatively specialised jargon of a professional field. I believe this is the source of some unfortunate press – that the formal language used within the profession for its own communication, is considered the only language for communicating with students. The Council should consider supplementing the language of the standards with the use of naturalistic descriptors used in analytic marking keys. The full report provides an example of such generic descriptors.

Recommendation 9 *That the language of outcomes and level descriptors be recognised explicitly as the technical language of the profession and that results from analytic marking keys which arise from tasks set for students for school assessments be used to supplement the feedback to students and communication to students on their progress and for communicating with parents.*

Case Study 2 – the assessment of drama

Case Study 2 involves the assessment of a *performance* and a *written* component of a course and illustrates a number of important features. The performance had four aspects of which all were compulsory; within the written, three questions of one aspect were compulsory and one of three questions in each of two other aspects was compulsory. The written and the performance components were each supposed to be weighted equally; hence the maximum score of 50 for each. Importantly, and consistent with *Recommendation 8*, different criteria did not all have the same number of categories - the number of ordered classifications ranged from 21 to 6. The report describes the assessments in some detail. In summary, the performance component seems to be the easier one on which to obtain a higher mark, by some 3 marks, than the written component. Figure 5 of the report is reproduced to show this.

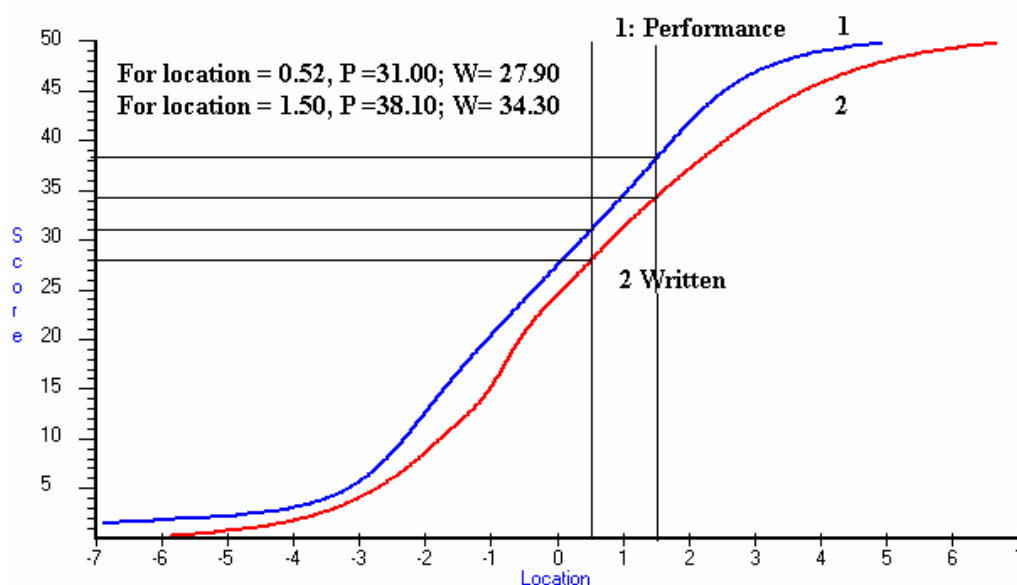


Figure 5. Score equivalence between the performance and written components

This example was also used to consider the implication of summarising a course score which has both a written and a performance component. The summary score reflects a *higher order and broader* variable than each of the components of assessment, the performance and the written.

An argument could be made that because these students have selected to study drama, many are genuinely excellent in drama performance and not as good at the written component of drama; therefore that the marks in the two components should not be scaled against each other – effectively that the greater weighting of performance should hold. 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. However, in considering this argument itself, the purpose of the summary score and its general frame of reference for constructing a TER needs to be considered.

The TER is a ranking for tertiary selection and is used as the only criterion for many courses that do not set prerequisites. It is a reasonable policy decision, therefore, to require all courses used for obtaining a TER to have a written component. Making the written component and performance component scores each out of 50 indicates that they are intended to contribute equally. Accordingly, it would be necessary to scale the performance and assessment data in Drama to ensure the policy is enacted. 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.

***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 be weighted not less than 50% for the contribution to a generic TER score. That it also negotiates with tertiary institutions to provide disaggregated scaled scores or differently weighted 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.*

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. These comments are concerned with a written external examination.

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 if relevant the questions in the external examination show the unit or units of study for which the question is most relevant.*

The Curriculum Council should consider using appropriate *open book* examinations in all external assessments where it is relevant. That is, that appropriate materials that students can take into the examination to aid the assessment be permitted and made explicit. 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. Indeed, given that the external examination is worth 50% of the final mark, it needs to be as valid as possible.

Recommendation 12 *That the Curriculum Council considers having 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. Further, 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.*

Policy implications for assessment

In summary, it is recommended that the marks provided by the schools for purposes of constructing a TER be based on analytic marking which operationalises further the ratings and levels. These could be marks out 100 for each unit, and then a summary course mark which is also out of 100. The analytic marking is complementary to ratings into levels of achievement and may be mapped back to these levels for purposes other than for constructing a TER.

This may appear to reflect elements very much like the status quo. However, this is governed in part by the need to assist teachers to use the standards and in part by retaining policies that there be a school based and an external assessment for each course, and that a single TER will be formed which is a summary of equally weighted performances in each of the courses. Furthermore, as indicated elsewhere, it is not considered that assessment for all purposes, and in particular for constructing a TER, is the defining feature of the OBE reforms.

If certain policies are the same, then it is inevitable that matters directed by the policies will remain the same unless their method of implementation was demonstrably inadequate. I do not believe this is the case, and indeed current methods would be the envy of most jurisdictions concerned with tertiary selection.

The policy that the school based and external assessments will be scaled against each other implies that the policy that this scaling be done automatically, at whatever schools or groups of schools are considered to assess on the same scale, should be retained. Particularly if the external exam is an open book exam, it will be the most valid kind of assessment for this purpose of scaling, and only because it is the same assessment that is conducted with all students, not because it is inherently superior to the school based assessment.

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

With experience, the school based assessments should not vary by very large amounts from the external assessments – that is a matter of teaching experience and opportunities for teachers to work with each other, something which occurs by and large now. This feature of professional competence is not something that arises out of

Report to the Curriculum Council Executive Summary

OBE or any other approach to organising teaching and learning, but something necessary to place assessments on the same inherent, relatively arbitrary, scale. If there were no scaling and there was an inherent discrepancy of only five marks out of 100 in each course for a school, but all the discrepancies were in the same direction, the accumulative effect would be substantial. This would imply that depending on which school students attended, there were substantial differences in the TER for no reason other than that the scale on which the measurements were expressed are different. I am sure that no one would sanction such circumstances.

Of course, this is not different in principle from the proposal that the levels in the different outcomes from the different courses be of the same intellectual demand – it is just that I consider that this cannot be achieved by decree at a fine enough level using 8 generic descriptors that span 12 years of schooling, even if finer ratings within levels are attempted.

Finally, if a general achievement test is considered for the purpose of scaling using the principles outlined above, rather than general monitoring, then it will be imperative that the score on the test contributes to the student's final TES. If it does not, then in the face of the many assessments that count for them, students will not take it as seriously as it is required for the purpose of formal scaling.

Bibliography

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.

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

Curriculum Council (1998) Curriculum Framework. Curriculum Council. Osborne Park

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.

Report to the Curriculum Council Executive Summary

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.