COMPARABILITY OF ASSESSMENTS,
GRADES AND QUALIFICATIONS

D ROYCE SADLER
Faculty of Education, Griffith University

Abstract
Traditionally, the desirable qualities of tests and assessments that are most frequently referred to in the literature are reliability and validity. Reliability broadly speaking has to do with how trustworthy a result is, validity with how well a measurement achieves its purpose. In this paper, I draw attention to a complementary concept, comparability, that assumes considerable importance in assessment contexts involving different assessments by different teachers, sometimes on different content and skills. Such contexts include school-based assessment, examining in different disciplines or fields of study, and comparisons of different qualifications. It is the focus of attention in recognition of prior learning, that is, in the broad equivalence of performances or competencies developed through various formal or informal activities. My interest is in exploring what is meant by comparability, and whether it is possible (and useful) to develop an interpretation of the concept that will do service in different contexts. I am more interested in how the concept can be interpreted than in cataloguing the various approaches that can be, or have been, used to determine levels of comparability in particular educational settings. The latter are, however, instructive in that they provide some raw material for an analysis of the concept.

Introduction
Unlike reliability and validity, comparability cannot even in principle refer to a single measure or assessment. It is a phenomenon applicable to two or more entities, such as achievement results. This paper is about the comparability of assessments, grades and qualifications in broad terms, with special though not exclusive reference to students at the end of secondary schooling. Although the notion of comparability lies at the heart of much of the theorising and practice in
educational assessment, comparability somewhat surprisingly seems to have received relatively little explicit attention in the assessment literature. Some exceptions are Nuttall, Backhouse & Willmott (1974), Christie & Forrest (1981), and Good & Cresswell (1988) with brief mentions by Gipps (1994).

A concern with the comparability of results shapes many policies and procedures at all levels of educational systems. It is especially relevant in the present social context, which is characterised by limited employment opportunities for school leavers and high demand for places in colleges and universities. The environment is decidedly competitive. Parents naturally want their students to compete on an equal footing with other students who are leaving school at the same time, or who have already left school but are still seeking work or an opportunity for further education. The importance of comparability as an issue is directly proportional to the consequences of wrong or unjust decisions. Given these considerations, it is clear that care must be taken to set up and maintain procedures that are soundly based and that foster public confidence.

There are a number of reasons for focusing attention on comparability, all of them related to issues of equity and equal opportunity. For example, grades should necessarily be comparable if certificates are to enjoy high credibility and currency throughout a school system. A certificate issued by a central education authority has a significant social value in that it acts as a leveller, recognising the achievements of students no matter whether they study at a metropolitan public school with high reputation and visibility, an exclusive private college, or at the only secondary school in a country town. The imprimatur of the issuing authority gives the certificates wide currency, and protects the interests of students who attend small, isolated, or little-known schools. Achieving comparability is therefore an important factor in achieving equity in education. Wherever a centrally authorised certificate is issued, a persuasive case for a variation in standards from school to school or from district to district to take into account what might be called "local conditions" is difficult to make. If it is attempted, the outcome is difficult to interpret because new elements of potential ambiguity are inevitably introduced.

A second example of a context in which equity is important is when universities use grades in their admissions procedures to sort students into rank order for offering places or for awarding scholarships, or to determine whether students have
reached specified standards in prerequisite studies for entering certain undergraduate courses, assuming of course that special aptitude or entrance tests are not used for admission.

Underlying the development of the theme in this paper is the assumption that despite the differences in form, enough generic similarities exist for comparability to be worth exploring and clarifying. However, providing a comprehensive analysis of comparability in its various forms is hardly feasible in a conference paper. I will be satisfied if awareness of the concept is raised, and certain commonalities identified. I will, therefore, outline something of the general agenda that needs to be worked through and provide a tentative analysis of restricted aspects only. The latter will draw primarily from experience with secondary education and its interface with entry to higher education. It makes no difference to the discussion in this paper what the grade designations are (letters, numbers, or verbal descriptors), how many make up the scale, or whether one of them is regarded as the minimum for "passing" a subject. For convenience, assume that the grades are awarded on a letter scale "A", "B", "C", and so on. The nomenclature of awards similarly makes no difference to the general analysis.

Comparability analysed
We shall begin with an abstract but simple analysis that looks at the nature of comparisons and comparability, without worrying for the present what the things are that make up the subject of discussion.

There are two easily distinguished meanings for comparable recognised by compilers of most dictionaries. The first means "able to be compared", and follows directly from the etymology. The emphasis is on "able to". This meaning is used when there is some question as to whether it is possible to compare two things that are not identical, and is written here as compare-able. The second dictionary meaning is "more or less equal", and when used this way, the word is written here as com-parable. Compare-ability is a prerequisite for com-parability. That is, a judgment about the compare-ability of two potentially equivalent things is logically prior to a judgment about the equivalence itself. In the present context, the "things" are (mostly) various kinds of student achievements from different sources. With respect to drawing a conclusion as to similarity or equivalence, compare-ability is essentially prospective, and com-parability descriptive. Criteria play a decisive role in making such judgments.
Now consider compare-ability in a little more detail. Some things so obviously belong to the same class, and are so obviously compare-able, that the matter is not given a second thought. For example, performances of a set piece of music by different students are clearly compare-able. So are the acrylic bowls made by students in the manual arts workshop. The relative merits of things from within a particular category can often be decided without resort to explicit criteria. In other words, the criteria are implied by the context, although they have to be brought to the surface and made explicit if teachers' qualitative judgments are to be substantiated to students, other teachers or agencies, and parents.

The question of compare-ability for things from different classes (such as a watercolour and an oil painting in art, or a cartoon and a poster in English), or of complete folios of student work, is more complicated because the things may be compare-able with respect to some features or aspects and not with respect to others. An egg and a glass of milk are not compare-able with respect to shape (compare-ability in this case is simply nonsense), but they are compare-able with respect to nutritional value or cholesterol content. Here, a criterion by which to assess compare-ability is not immediately suggested by the context and has to be stated explicitly. Once it is stated, the compare-ability issue is settled and it becomes sensible to ask whether the things might also be com-parable.

The need for criteria that cut across several (but not necessarily all) classes is directly dependent on the degree to which surface dissimilarities may obscure the deeper commonalities. But it is precisely these commonalities (which are signalled by the criteria and standards in a subject) that help to characterise art as art, economics as economics, and (say) a first class honours thesis. Under school-based assessment, schools make choices with respect to such aspects as content and assessment instrumentation, within the general framework of a syllabus. As a consequence, subjects are "expressed" differently in different schools. It would be interesting and informative to research the criteria and standards that lie beneath the assignment of grades to student folios, and to delve into the similarities, differences, and trade-offs involved in arriving at an assessment. Such a study would involve inquiring into processes actually used by the teachers who made the original assessments. It would map the decision field, explore the
tolerances within which things are deemed to deserve the same grade, and attempt to identify any distinctive characteristics that appear to separate one grade from the ones on either side. The research would share some of the characteristics of so-called moderation meetings (common to school-based assessment systems), in that the reasons and rationales for decisions would be made explicit.

Clearly a different approach would be needed for unravelling comparability across subjects. One tack could be to create cross-disciplinary assessment panels in an effort to explore what it might mean conceptually to speak of comparability (actually, compare-ability) across subjects. In general terms, this would involve considering intrinsic conceptual complexity, the nature and depth of the intellectual demands made upon students (for example, critical thinking), and hence academic (or other forms of) rigour.

That a study of this type is conceptually feasible can be demonstrated by considering the extreme ends of the spectrum for a particular subject, that is, the very simple end (reflected in what eight-year-olds could master with ease) and the most sophisticated end (what current researchers in the discipline are exploring). Somewhere in between some patch must exist that could be judged as appropriate to (say) A-level students in the subject in the final year of schooling. Presumably, the same would be true for all disciplines. What makes this possible? What considerations are applied, even intuitively? The research task would be to explore this cross-disciplinary territory and come up with some sort of statement without trivialising either any common ground discovered, or the individual disciplines. The result could have worthwhile implications for curriculum development.

Faces of comparability in educational achievement
A number of distinct faces of comparability can be identified. Without claiming to be exhaustive, I list the following:

Comparability of grades
across assessment tasks (designed and required by different teachers in the same school);
across teachers of different classes (in the same subject and teaching context, with the same assessment tasks);
across subjects within a discipline or area (subjects such as Victorian literature, spoken discourse, writing & style, and modern drama within an Arts degree);
across disciplines (such as music, mathematics, ancient history, Japanese);
across time (within one subject or discipline over a period of years);
across institutions (schools, colleges and universities; for example, first class Honours in Physics from different universities); and

Comparability of qualifications across institutions and sectors (certificates, diplomas, degrees from universities, TAFE colleges and specialist providers).

The last of these is increasingly important as pathways through education and training become more numerous and diverse, as Recognition of Prior Learning becomes codified and commonplace, and as credit transfer gathers momentum (as evidenced through, for example, the recent establishment of the Australian Credit Transfer Agency under the Australian Vice-Chancellors' Committee). Observe that some of the classes listed above overlap or invite border disputes (such as the "across subjects" and "across disciplines"), that a number classes may be nested in a particular context, and that many situations in which comparability is important involve several classes simultaneously.

Two of the cases in the list above are now explored, namely comparability across subjects, and comparability within a subject across schools. Although it is not difficult to separate the two conceptually, in practice they often interact or one may be confused with the other.

Comparability across subjects It is natural to ask whether a particular performance level in one subject, say a grade of A in music, represents an equivalent accomplishment to a grade of A in another subject, say English. Is it easier to obtain a grade of A in some subjects than in others? Is a grade of A somehow worth more in music than in English, in the sense that it carries more weight (implicitly or explicitly) in the considerations for a composite rank order? This issue clearly concerns students as they choose their subjects for the final years of schooling. No one wants to be disadvantaged merely by inadvertently choosing a poor combination of subjects. If, for example, school leavers require a certain minimum number of As, Bs or Cs for entry into employment or training courses, it would not do for the As to be easier to achieve in some subjects than in others.
As indicated above, one approach to studying comparability across subjects is to examine the intrinsic conceptual complexity of the content of different school subjects, the nature and depth of the intellectual demands made upon students, and the characteristics of the testing program in relation to how the material is presented, learned and assessed. Clearly, different criteria for judging performances are appropriate for different subjects. In large measure that is why each subject is able to make its own unique contribution to intellectual, social and personal development, and why so many students choose different subjects according to their aptitudes and interests.

Although a community may generally expect that subjects should be pitched at a level appropriate to, say, the final year or years of schooling, it is a moot point whether it is possible to ensure that subjects are "equally rigorous academically", except in a vague sort of way. If an attempt were made to enforce strictly some such requirement, it might result in no As being awarded at all in subjects that attract students who are less able. Almost certainly, this would be negatively motivating for those students, in that hardly any of them would taste of success, for quite arbitrary reasons. It could be argued, therefore, that subjects are innately so different that the pursuit of this type of comparability is irrelevant. It should be sufficient to allow each subject to go its own way, and leave the interpretation of different grades in different subjects or disciplines to those who will make use of them. One would expect only naive educators and employers to look at a student's grades without regard to the subjects studied.

Before leaving this line of thinking, it is useful to explore one of its implications. If all pretence that different subjects can somehow be made comparable were to be given up, it would be necessary for all users of certificates to be scrupulous in taking account of the distinctive contributions made by different subjects, and not to equate them, even mentally. Provided comparability of grades within a subject across schools could be assured, the onus would be directly on decision makers to recognise the unique contribution of each subject, and not to aggregate at all. In terms of university entrance, this would open up a variety of possibilities for selecting students for courses using clearly articulated criteria related to grades in particular subjects. The practical disadvantage would be that students would have to make critical choices about subjects for their eventual careers with enormous care one or two years prior to
finishing schooling, and that no trading off of differential performance across subjects would be possible.

While it is interesting to speculate about how one might untangle the philosophical and practical problems of judging the relative rigour of what are in reality incommensurables, the exercise would certainly be quite difficult and the pursuit of that goal may not achieve enough to be worth the effort. This might be thought a convenient point to retreat from further consideration of comparability across subjects. Perhaps it would be, except for the fact that people are in the habit of interpreting events for themselves and placing their own constructions on outcomes.

Comparability across subjects is obviously important when the grades awarded in different subjects are treated as though they were interchangeable, with one subject being regarded as equal to any other for some purpose. Even where grades are reported on a non-numerical scale, the grades are often assigned numerical values and then averaged, added, or otherwise combined. When this is done (as it often is) without regard to subjects studied, for any purpose, and at any time whatsoever, effectively the same assumption is made. The matter deserves some attention.

The most common technique used for investigating across-subject comparability is fairly straightforward. Using the subjects from the beginning of this section as examples, for all students in a given population who study both music and English, the proportions receiving each of the different grades can be compared. If these proportions are similar, the subjects are deemed to be comparable. The assumption made in this approach is that over a whole school system, local variations balance out and the similarity of the distributions of grades provides a valid measure of the comparability of different subjects.

Of course it is possible to ensure (rather than just check after the event) the similarity of distributions across subjects, across successive years in one subject, or across both subjects and years. The simplest way to do this is to specify in advance what the distribution must be (along, perhaps, with some statement about acceptable tolerances). This is sometimes called "grading on the curve". The actual shape of the "curve" is immaterial. It could be bell-shaped (for example, Gaussian), rectangular, or something quite different again. On the other hand, its shape may not be specified at all; any shape would then serve, except that
high levels of skewing (positive or negative) may be identified as unacceptably aberrant and call for reanalysis. The rationale for grading on the curve is based on two premises, namely that (a) students should be neither advantaged nor penalised for choosing particular subjects or combinations of subjects, and (b) it is necessary to impose control on the distributions to minimise inflation of grades and consequent devaluing of the grades themselves.

Equalising the distributions of grades places a premium on the difficulty level of the subjects as experienced by whatever cohort of students chooses to study them, regardless of the intellectual demands made by the content and skills required in the particular subjects. As a consequence, the comparability that is achieved is on one dimension only, because difficulty level is only one of the many possible characteristics of assessment in a subject.

Grading on the curve obviously guarantees uniformity in grade distributions. It also has a certain superficial appeal, which is why the issue is being discussed in this paper. But uniformity is not necessarily the same as what most people would want to interpret as comparability in school-based assessment. This limited interpretation of comparability can be achieved by grading on the curve only if it is reasonable to assume (a) that all of the (overlapping) populations of students for the different subjects are equally able academically, (b) that the average performance of students in a subject remains constant from year to year, and (c) that one has some independent rational way of knowing what the distribution should be. These conditions are fairly stringent and are almost never met in systems where students self-select into subjects and courses, and where the proportion of the age group staying on at school past the official school-leaving age (say, to the end of secondary education) steadily increases (as at present) or decreases.

Before leaving the topic of comparability across subjects, it is perhaps useful to make a comment about the philosophy behind another common technique of ensuring comparability of grades, namely statistical scaling. A theoretical perspective on across-subjects comparability lies at the root of the practice of combining assessments in different subjects to form aggregates, which are then used to create rankings of students. The fundamental postulate is that, for any particular student, there exists an underlying generalised trait or characteristic which, broadly speaking, influences in a major way how that student performs in a variety of
subjects at school. What that trait should be called is immaterial, but for the purposed of this paper it can conveniently be called simply ability. Each subject assessment constitutes an estimate of this underlying ability. If a student studies five subjects, there are five available estimates of ability. All five estimates then have to be expressed in terms of a common metric to find the best estimate of the ability itself.

The statistical operations usually employed have the effect of projecting achievement measures in all subjects on to a common scale prior to aggregation. These techniques directly or indirectly align each subject with every other subject. The best estimate is defined by calculating the mean score, which is superior to any single measure. Variations of the estimates about the average are treated as measurement error. Broadly speaking this is the theoretical rationale for scaling and forming composites. In a certain sense, therefore, the scaling procedures achieve comparability across subjects, because a particular quantum (namely, the scaled subject measure) that is contributed to a student's aggregate has the same value regardless of subject from which it is derived.

Comparability within a subject across schools
The second face of comparability explored briefly in this paper is related to the grades within a subject across schools. Comparability of achievements across schools does not attract much public attention under external examination systems, of which there has been a long tradition both in Australia and overseas. Presumably, the assumption is that when the contents of the examination paper are kept secure until the scheduled hour, and when all students sit for the same examination paper under rigidly controlled conditions, all students have equal opportunity to demonstrate their achievements. But although the public may seem generally satisfied that all students face the same challenge, comparability is always a latent issue.

A particular reservation about the comparability of external examination results arises from the recognition that not all students are equally suited to the formality associated with external examinations, and as a result do not perform as well as they should. In addition, anomalies sometimes occur through the ways teachers prepare students for the examinations. For example, chief examiners have been known to comment about the surprising uniformity of scripts from particular schools in certain years. Sometimes the performances are uniformly on target, sometimes uniformly wide of the mark. Such a phenomenon occurs when teachers try
to anticipate particular questions, and coach or otherwise channel their students' efforts into producing model answers to the question expected. Other reservations have to do with choice of questions (a matter touched upon below), the consistency of markers (both over time, and from one marker to another), and the comparability of grades in a subject from one year to another. But by and large, external examinations are often accepted as providing a satisfactory degree of comparability.

In school-based assessment systems, comparability across schools has to do with whether the performances of all students who are awarded a particular grade in a subject are, within the range of performances associated with a designated grade level, of equivalent quality regardless of the school attended. This does not imply that students should tackle identical tasks in all schools, nor that test instruments and assessment programs should be standardised and similar in format. It does involve comparing things that are commensurable, for the reason that the "things" are achievements within a single subject. It is natural and necessary to ask how student achievements that are measured by means of different assessment programs, and sometimes on different content, can be considered equivalent, and therefore comparable.

In school-based assessment systems with a central certifying authority, achieving comparability across schools is typically carried out using what are known as "moderation" procedures. These procedures are essentially administrative, and control the composition and responsibilities of the review panels (or person), the timing and conduct of panel meetings, the sampling and submission of student work, how the moderation process itself should proceed, the security of the operations, how conflicts or appeals are to be resolved, and quality assurance mechanisms for the process as a whole.

Achieving comparability of outcomes within a subject across schools actually starts within each school. This is a prerequisite to achieving comparability across schools simply because it is pointless to examine between-schools variability if the primary source of variability is within schools. All students must be awarded grades that reflect accurately the quality of their work, and the assessments must not be affected by extraneous factors. Potential threats to comparability within a school may be grouped under three headings: (a) the teacher's personal grading practices, (b) the design of the assessment program, particularly in the
matter of choice in assessment tasks, and (c) the approaches taken by different teachers.

First consider teachers' grading practices. There is evidence in the literature that a teacher's assessments can be influenced by the teacher's values compared with those of the students, by teachers' knowledge of personality and other characteristics of students (including their diligence, classroom behaviour, and agreeableness), by the order in which they assess or grade student work, and by handwriting. Some of these can be minimised by suitable techniques. For example, the seriousness of order effects can be reduced by randomising before each grading session. Ideological bents and halo effects are more elusive, but their seriousness can be reduced by arranging for cross-marking with other teachers who do not know the students.

The second factor arises when students are given a choice of tasks. Suppose, for example, that students in a particular subject are given a choice of assessment items on an examination paper (whether for external or school-based examinations, assignment topics, or take-home projects). A student who chooses certain items may be able to submit work that is primarily descriptive. A different choice of items may require higher-level critical skills, and therefore be more demanding. If an unrestricted choice is allowed, it becomes difficult to judge the equivalence of descriptive versus critical responses from students, and hence to draw a conclusion about comparability. A practical solution to this problem is to insist that students choose similar proportions of items that make cognitively similar demands from the available categories.

The third element arises when classes in a subject are taught by different teachers, even when the same tasks or instruments are used. One teacher may teach to the test, deliberately or inadvertently. Another may prime students to give superb but atypical or artificial performances. Yet another may be overzealous in the opposite direction, even to the point of not preparing students adequately for the test. In addition, the wording in some assessment tasks may be more familiar to students in one class than in another, not by design, but because teachers who are thoroughly conversant with a body of subject matter are sometimes unaware that they teach using a particular, and to some extent idiosyncratic, vocabulary. The teacher setting a test may have pet phrases or forms of expression that are semantically equivalent to the forms of expression used by teachers of other classes,
but are nevertheless unfamiliar to those classes. Or the teacher setting the test may run so close to production deadlines that other teachers have little time to review the test thoroughly. It is obvious that these matters can be dealt with only within the context of the school, and that schools should develop policies that encourage internal moderation, openness, and consensus with respect to assessment to ensure fair and comparable assessments for all classes and all students.

Conclusion
Comparability is an abstract property of sets of achievement results (such as grades) or aggregations of them (such as qualifications) that is fundamental to equitable decision making by and on behalf of students, and to public confidence in a credentialling system. In this paper, a number of aspects of the concept have been identified and two (namely, comparability across assessments and comparability across subjects) explored. Further analysis of the nature and significance of comparability is called for, particularly in a context of multiple pathways towards certification of academic and practical achievements.

References