Assessment in Education
Vol. 12, No. 2, July 2005, pp. 105–123

Examination standards and the limits of linking

Paul E. Newton*
Qualifications and Curriculum Authority, UK

There is a tendency in the literature to characterize linking as equating done somewhat less rigorously. The ambiguity of this conception can lead to confusion amongst policy-makers and members of the public and can result in the proliferation of comparability myths. As the constructs assessed by two tests decrease in similarity, so the difference between equating and linking becomes one of kind rather than degree. To help make sense of linking in different contexts, a general model is proposed, based upon the idea of a 'linking construct'. This general model is used to define the limits of linking and to clarify what users and stakeholders need to know about linking and linked scores. Finally, a distinction is drawn between judgemental linking as a method (e.g., social moderation) and judgemental linking as a theory (i.e., the value judgement theory of linking). The latter presents a challenge to the general model, which is defended.

Introduction

> It is not much of an exaggeration to state that the assumptions in our models are all false except those that are true by definition. (Brennan, 1998, p. 5)

Pragmatism tends to be a characteristic of those who work in the educational measurement profession. This is inevitable, since educational measurement is a technology that aims to provide practical solutions to real-world problems within limited time-frames. Although we rely on theory, to an extent, we simultaneously accept that our theories are ultimately inadequate. For instance, we realize that the idea of ranking students in terms of their overall level of attainment at the end of an instructional course is based on an over-simplified caricature of learning, but we still consider it to be a useful caricature for many purposes (Mislevy, 1993).

One of the toughest dilemmas we face is when to say 'enough is enough', i.e., when to say that our ideas can be extended so far but no further. This is so awkward to call

*Qualifications and Curriculum Authority, 83 Piccadilly, London W1J 8QA, UK. Email: NewtonP@qca.org.uk

ISSN 0969-594X (print)/ISSN 1465-329X (online)/05/020105–19

DOI: 10.1080/09695940500143795

because *all* of our theories and practices are questionable to some extent. How should we draw the dividing line between problematic and too problematic? And, which is harder still, how should we defend such decisions when they would frustrate the aspirations of policy-makers? The idea of linking standards across assessments presents a paradigmatic example of this dilemma.

Educational measurement professionals have, arguably, not been forceful enough in drawing dividing lines between problematic and too problematic. We have allowed myths to develop which, due to their longevity, are now extremely hard to challenge. This has been particularly true in relation to the linking of examination standards over time. For example, in England, ever since the subject-based A level examination was introduced in 1951, the boards responsible for examining and certification have claimed—with each successive year—to have successfully linked standards from the previous year's examinations to the next. If so, then the 'obvious' corollary is that the 1951 pass standard in any subject must be comparable to the 2005 pass standard. Yet, the A level has changed radically over time, and no one who genuinely understood the implication of those changes would sanction such a simplistic inference. Unfortunately, persistent misunderstandings of the inferences that follow from linked standards over time have repeatedly perplexed policy-makers and the general public alike (e.g., Cox, 1975, p. 38; Orr & Nuttall, 1983, p. 14; Baker *et al.*, 2000, p. 9). At times, these misinformed debates have threatened to destabilize the system entirely; for example, during 2002, following the most recent radical reform of the A level system.[1]

The discussion that follows is not about the peculiarities of linking standards in England, it is about the peculiarities of linking standards, *per se*. It relates to any context in which linking relationships are established, in any country, and particularly when linking relationships are established between tests or examinations that are built to very different frameworks and specifications, e.g., assessments in different subjects. Having said that, the examination system in England does constitute a useful case study, especially given theoretical developments in England during the past decade, and it will be used as such.

The theoretical foundations for linking standards are not well articulated, and the following discussion seeks to explain why. This lack of articulation makes it hard to define the limits of linking—to identify when problematic becomes too problematic as far as linking is concerned—and it makes it hard to communicate the limits of linking, and the inferences which do and do not follow from linked scores, to policy-makers and members of the public. I will argue (contra some) that there *are* important limits on linking which we need to respect, even if that makes life politically hard for us.

A general model of linking

According to Kolen and Brennan (2004), the goal of any linking exercise is 'to put scores from two or more tests on the same scale—*in some sense*' (p. 423). But what sense, or senses, might we be talking about? The following section proposes a general

model through which to make *sense* of any linking exercise. It is based upon the idea of a 'linking construct'.

The model

Figure 1 is an attempt to illustrate the sense in which scores on any two tests could be said to be linked, via a linking construct. It is intended to model the conceptual basis of linking (rather than a generic methodology for linking).

In the most straightforward of situations, both test 1 and test 2 would be designed to exactly the same framework and specification, i.e., designed to assess exactly the same construct in the same way. This would enable the establishment of an equating relationship (see Feuer *et al.*, 1999).

According to the general model, equating is the special (or trivial) case of linking in which the linking construct is identical to the construct assessed by each of the to-be-linked tests. So, if both of the tests were designed to assess the same mathematics syllabus, then the linking construct would also be defined by that mathematics syllabus. From Figure 1, note how the level of attainment in mathematics on the linking construct scale corresponds to different scores on the test scales. The two test scales are linked via the linking construct.

As noted by Kolen and Brennan (2004), the term 'linking' is most often used to denote situations in which test 1 and test 2 are designed to different frameworks and

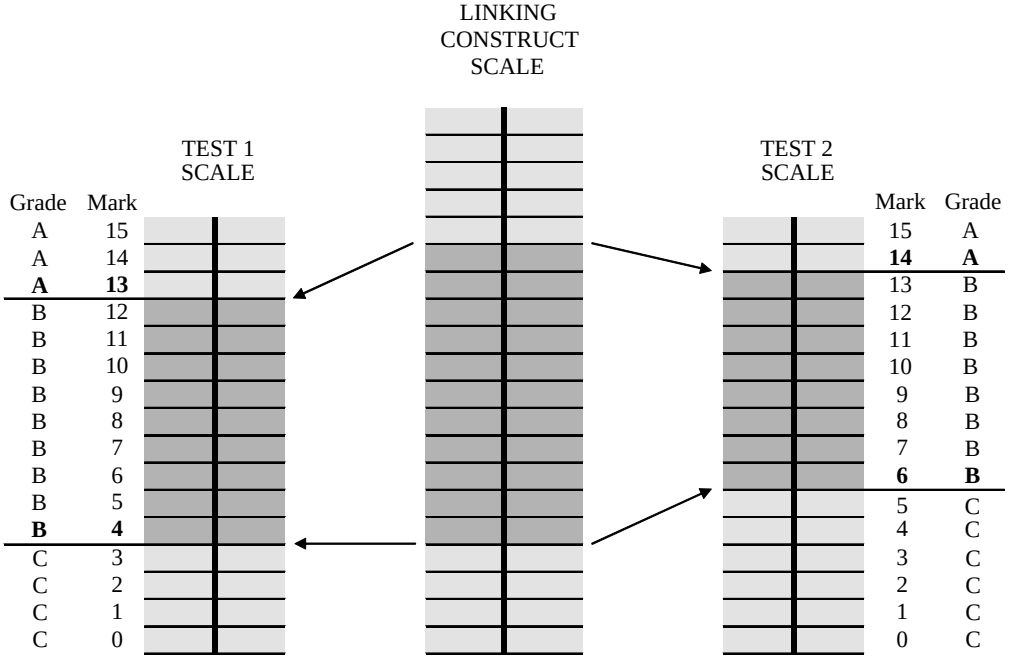

Figure 1. A graphical representation of linking standards

specifications. Here—and this is the crucial point—the linking construct is identical neither to the test 1 construct nor to the test 2 construct. Instead, the linking construct might represent higher-level skills/abilities shared by the test scale constructs, or common sub-domains of the test scale constructs, or perhaps a more distant construct still. For the link to have *some* meaning, though, the test scales would have to be at least partially correlated, i.e., they would have to share *something* in common.

A frequently cited example of a linking relationship is that between the SAT-M and the ACT mathematics tests, both of which are commonly used in the USA to support college admissions decisions. Here, the linking construct might be the 'developed abilities and skills in the domain of mathematics' which both are assumed to measure to some extent (see Kolen & Brennan, 2004, p. 428).

In the general model proposed above, it is the linking construct that makes the link *meaningful*. In other words, inferences drawn from linked scores (on different tests) should be expressed in terms of the linking construct. Or, alternatively, what students at linked scores (on different tests) could be said to share in common would be the same level of attainment/ability in the linking construct.[2] Indeed, this is *all* that could be said on the basis of the linking relationship.

Importantly, the linking construct might be significantly different from the constructs assessed by each of the to-be-linked tests. For example, it would be possible to link standards, between a test of mathematics attainment and a test of chemistry attainment, using IQ as the linking construct. This would not, in any meaningful sense, make a level of attainment in chemistry comparable to a level of attainment in mathematics. The process would simply mean that students at linked scores on the different tests would, on average, share the same IQ.[3] The tests would be linked in terms of average student intelligence, but not in terms of average student attainment. So, even though a mathematics student might know no chemistry, and a chemistry student might know no mathematics, scores on the two tests could still be meaningfully linked, using IQ as the linking construct.

By way of example, Western Australia statistically adjusts the marks of different Tertiary Entrance Score (TES) subjects onto a common scale. For most students, university admission relies on a single index of achievement, the Tertiary Entrance Rank (TER). The TER is based on the sum of the scaled marks achieved by each student in the particular combination of subjects which she studied. Subject marks are scaled prior to aggregation to ensure that students are not disadvantaged by choosing a difficult subject. The scaling is based upon a linking construct which has been referred to as 'academic ability' or 'potential' (see Partis, 1997); score distributions, for each subject, are scaled to reflect the ability distributions of the students which studied them.[4]

It is worth emphasizing that the linking construct might be significantly different from the constructs assessed by each of two to-be-linked tests, even when the tests actually shared much more in common. This might occur, for instance, if grade standards across a whole host of subject examinations were to be linked according to general academic ability. Even if pure mathematics and applied mathematics could be said to share a common 'mathematical ability' linking construct, they (and all the

other subjects) would still be linked according to general academic ability. This is the principle underlying the scaling of subject marks in Western Australia (and other Australian states).

In proposing the general model, a major intention is to emphasize that linking is not simply 'less rigorous' than equating, and comparability is not simply 'more approximate' than equivalence. In my view, this kind of talk is inappropriately woolly and can result in serious misunderstanding and negative impacts upon policy and practice. Instead, as the constructs assessed by two tests decrease in similarity, so the difference between equating and other forms of linking becomes one of kind rather than degree.

To appreciate the significance of this proposal we need to consider the inferences that are supported in each case. When two tests are designed to exactly the same framework and specification, the process of equating supports inferences concerning the single construct which both scales represent. However, when two tests are designed to different frameworks and specifications, the process of linking can only support inferences concerning the linking construct that encapsulates (at least something of) that which the two scales have in common. Students at linked scores from two tests can be considered comparable only in terms of the linking construct that the tests share in common, and not in any other sense. Note that the purpose of linking is to allow us to draw inferences about *students* and not about tests. When we say that test-standards are comparable we actually mean that students at linked scores are comparable (and the linking construct explains the sense in which they are comparable).

Different types of linking

A corollary of the foregoing discussion is that there could be no straightforward hierarchy of linking types. To some extent this challenges previous frameworks, particularly the formulation presented by Linn (1993) which was based upon the seminal work by Mislevy (1992).[5]

Linn (1993) listed five types of linking 'listed in order of statistical rigor, with equating being the most rigorous and social moderation being the least rigorous' (p. 85):

1. equating—tests measure the same construct, in the same way
2. calibration—tests measure the same construct, but somewhat differently
3. statistical moderation—tests do not measure the same construct, but scores can be linked statistically using an external measure (e.g., a measure of general academic ability such as the Australian Scaling Test)
4. prediction—tests do not measure the same construct, but an empirical relationship between scores can be estimated
5. social moderation—tests do not measure the same construct, but scores can be linked judgementally.

The problem with the idea of a straightforward hierarchy is that it necessarily confounds questions of conceptual foundation and methodological rigour. There are two basic questions to be asked of any linking exercise:

- what do the to-be-linked tests have in common, i.e., how should we characterize the linking construct (the conceptual foundation question)?
- which methodology is most suitable for linking standards in the relevant context and how reliable is it (the methodological rigour question)?

In principle, a judgemental methodology could be employed to link standards between two tests designed to exactly the same framework and specification (i.e., to establish an equating relationship). If so, would this elevate judgemental linking to a higher position in the hierarchy than prediction, statistical moderation or even calibration? The lack of a definitive answer signifies that the idea of a hierarchy is misleading. It emphasizes that the distinction between equating/equivalence and linking/comparability is not actually grounded in degree of 'statistical rigor' (Linn, 1993, p. 85) or ''strength' of the resulting linkage' (Kolen & Brennan, 2004, p. 429).

Linn (1993, p. 94) said of statistical moderation: 'although comparisons are in fact made among students based on their statistically moderated scores on different combinations of tests, the scores cannot be considered equivalent in any rigorous sense.' I would go further by arguing that such scores could not be considered even *approximately* equivalent. Inferences based on score comparability are of a different kind from inferences based on score equivalence; it is not simply that we should be less confident in the rigour of the linking methodology and less confident in our claims. This is an important point. Even when test scores have a wide confidence interval around them the reported score is still the best estimate and should be treated as such. By the same token, if we appear to talk about comparability as though it were akin to equating with a wide confidence interval, then we will mislead users into drawing exactly the same kind of inferences from linked scores (which do not measure the same construct) as from equated scores (which do).

Mislevy (1992) discussed the statistical moderation of tests in different subject areas slightly differently. He proposed that there: 'is no pretense that the two tests measure the same thing… *moderation* simply aligns scores from the two as to some measure of comparable worth' (p. 25). He later suggested that, when linking via a moderating test, 'the test's specifications determine the locus of value for "comparable worth" (p. 68). This could be interpreted as tantamount to my earlier proposal that inferences drawn from linked scores across different tests should be expressed in terms of the linking construct. On the other hand, he subsequently claimed that moderation forces an 'arbitrarily determined operational definition of comparability' (p. 72) which seems to suggest that he would prefer not to draw *any* meaningful inferences from linked scores. Yet, if no meaningful inferences followed, then why would we link scores at all? Surely this would be to introduce an unacceptable threat to validity, i.e., the introduction of apparent meaning where none actually existed? One response would be to argue that, although no meaning followed, the link might still enable the use of results for pragmatic purposes which would not be possible otherwise (e.g., the 'fair' use of results, from assessments in different subjects, for university selection). This is a red herring, though, since such

use would inevitably be predicated upon an implicit meaning even if that meaning was officially denied.

Judgemental linking

Before concluding discussion of the general model, the idea of judgemental linking needs to be considered in more detail. There is a need to clarify some of the terminology that has been used in the past, and to introduce a crucial distinction between judgemental linking as a method and judgemental linking as a theory.

Social moderation, as described by Linn (1993), describes a family of methods for bringing standards into line through the exercise of professional judgement. In countries such as England and Australia, it is typically associated with the evaluation of coursework, project or portfolio work by teachers. Coursework tasks in England tend to be somewhat more circumscribed than, say, portfolio evidence in Queensland. However, in both cases, different assessment evidence is judged according to a common national/state standard for the subject. In short, social moderation is generally associated with contexts in which the linking construct is either identical to the assessment construct (used by teachers in different schools), or has a great deal in common with the assessment constructs (used by teachers in different schools).

In many instances of social moderation, teachers from different schools gather together with samples of work that they have already assessed. Through a process of discussion and debate, they negotiate a shared understanding of the national/state standards and moderate (i.e., amend) their previous judgements as necessary. This is also known as consensus moderation, or peer moderation. In some countries, such as England, the process of coursework moderation tends to be somewhat less collaborative, and decisions rely more heavily upon the judgement of a moderator from an examining board.[6]

In terms of a typology of linking, it should be recognized that social moderation describes but one family of methods for linking standards judgementally; one in which teachers' judgements are brought into line. There are other families too; in particular, the family of methods for linking cut-score standards across test versions (e.g., judgemental grade boundary linking, described below in relation to the A level examination).

The limits of linking

Under the general model proposed above, the limits of linking are clear:

1. linking can only be achieved if a plausible linking construct can be identified (even if this construct can only be roughly defined and/or loosely defended);
2. inferences from linked scores can only be drawn in terms of the linking construct.

To facilitate understanding of linked scores and, thereby, to facilitate valid inferences and to prevent invalid ones, *the linking construct must be made explicit for all users and stakeholders*.

A challenge to the general model of linking

Much more interesting than the idea of value judgement as a *method* for achieving comparability, is the idea of value judgement as a *theory* of comparability. This proposal was developed in England in the 1990s, and was promulgated both by Cresswell (1996; Baird *et al.*, 2000) and by Wiliam (1996).[7] I shall focus on the work of Cresswell.

In preceding sections, I outlined a general model of linking, grounded in the notion of a linking construct. The linking construct provided a rational basis for making sense of comparability in different contexts. A corollary of the model was that, where no plausible linking construct can be identified, comparability cannot be said to exist.[8]

In stark contrast, Cresswell argued that comparability can genuinely be said to exist even when no rational basis for making sense of comparability can be established. The tension between these two perspectives will be explored below. The scene will be set through a brief description of present-day judgemental linking procedures in England—since this is the context in which Cresswell developed his theory—focusing on the subject-based General Certificate of Education Advanced level qualification (GCE A level).

Judgemental linking procedures in England

A levels were first examined in England in 1951. They were originally intended as qualifying examinations for university entrance and, as such, were graded as either pass or fail (although a distinction grade was introduced after a few years). However, they increasingly became instruments for competitive selection, and universities demanded greater levels of discrimination. Rather than making raw marks available, five passing grades were officially introduced in 1963, A to E. Nowadays, students tend to be offered university places contingent on their having achieved a particular A level grade profile (e.g., at least grade B in history and at least two other A levels at grade C or above).

The A level 'gold standard'

The A level was traditionally offered by examining boards which had emerged from universities (e.g., the University of Cambridge Local Examinations Syndicate). Different boards offered alternative examinations in the same subject; in fact, even within the same board, alternative examinations were offered for different syllabuses in the same subject.

The A level has changed considerably over the years. The number of boards has been reduced, as has the number of syllabuses offered, and greater regulation has been imposed. The most significant change of recent years was the Curriculum 2000 reform, which introduced modular assessment and a two-stage certification process.[9] What has notionally remained constant over the years has been the A level 'gold standard'—the standard required for the award of each grade in each subject.

For the A level 'gold standard' to be maintained, and for the A level to function effectively as a university selection device, linkage needs to be achieved (simultaneously) between different examinations:

- for the same syllabus (e.g., from one year to the next within the same board);
- for different syllabuses in the same subject (e.g., between one board and another, or from one year to the next within the same board following syllabus re-design);
- for different subjects (e.g., between psychology and chemistry).

Linkage is expressed through application of the same (A to E) grade scale across all examinations. This is achieved through a process of judgemental grade boundary linking. Although methods have evolved somewhat over time (and although, prior to regulation, different boards operated somewhat different procedures), the principle of judgemental linking has remained central.[10]

Present-day judgemental linking procedures

Recommendations for grade boundary marks are determined during grade awarding meetings. These meetings occur for each subject syllabus within each examining board and the primary objective of the awarding committee is the maintenance of specific subject standards over time. The awarding committee for each examination is chaired by the Chair of Examiners and also includes the Chief Examiner(s), Principal Examiner(s) and Principal Moderator(s). Examining board officers are required to advise the committee and direct its procedures.

The awarding committee, therefore, is made up of subject matter experts, who are generally experienced (and often practising) teachers, and who will have contributed to the development of the examination. They engage in a process of 'script scrutiny', which involves comparing scripts from the present year (at various marks) with 'archive' scripts from previous years (at grade boundary marks). In doing so, they use professional judgement of evidence produced in scripts to identify marks that correspond to comparable levels of attainment.

However, they are also required to pay heed to statistical information supplied by the examining board. This includes:

- technical information including mark distributions, relating to the question papers/tasks and questions;
- statistical information on candidates' performance in previous series (where available);
- details of significant changes in entry patterns and choices of options;
- statistical information on schools' and colleges' estimated grades for all candidates.

After the awarding meeting, the Chair of Examiners communicates her grade boundary recommendations to the examining board's Accountable Officer (who is ultimately responsible for deciding grade boundary marks for each examination). After review of the recommendations, in light of all the available evidence including outcomes for syllabuses in related subjects (e.g., comparing outcomes for history with

outcomes for other humanity subjects), the Accountable Officer is at liberty to decide upon alternative grade boundary marks should she consider this to be necessary.

The Code of Practice (e.g., QCA, 2004)—according to which the boards are regulated by the Qualifications and Curriculum Authority (QCA)—does not give explicit instruction on how the statistical and judgemental evidence should be combined, nor on whether to elevate one form of evidence above another in any circumstance. In addition, although judgemental decisions are supported by generic performance descriptions, there is no explicit definition of what is meant by the maintenance of examination standards. As such, the decision of the Accountable Officer to overrule recommendations from the Chair of Examiners can be a problematic one.

The value judgement theory of linking

Cresswell's thinking on value judgement as a theoretical basis for comparability was honed in the wake of a couple of decades of unsuccessful work on the development of grade descriptors for public examinations. Since the early 1980s, the examining boards had been under considerable political pressure to move from linking procedures which gave the impression of a kind of within-subject cohort-referencing (see Wiliam, 1996) to linking procedures which were grounded in the principle of within-subject criterion-referencing (see Secondary Examinations Council, 1984). The development of written grade criteria, and their use as a basis for linking exam standards over time, was intended to enable results to convey far more information about what students actually knew and could do (which would be good for employers), and to enable national and local monitoring of educational standards over time (which would be good for politicians and school managers).

Cresswell (1996, p. 57) introduced his argument as follows:

> By recognising the setting of standards as a process of value judgement, the analysis presented here explains why successive recent attempts to set examination standards on the basis of explicit written criteria have failed and, indeed, were doomed to failure. The analysis also provides, for the first time, a coherent theoretical perspective which can be used to define comparable standards in quite different subjects or assessment domains.

His sociological definition of comparability was couched as follows:

> Two examinations have comparable standards if candidates for one of them receive the same grades as candidates for the other whose assessed attainments are accorded equivalent value by awarders accepted as competent to make such judgements by all interested certificate users. (Cresswell, 1996, p. 79)

Wiliam (1996, p. 300), framing a similar analysis in terms of Searle's thesis on the construction of social reality (e.g., Searle, 1995), quoted Searle's story of a baseball umpire whose judgements were called into question:

> *Interviewer*: Did you call them the way you saw them, or did you call them the way they were?
>
> *Umpire*: The way I called them *was* the way they were.

In short, for Cresswell and Wiliam, those whom we empower to maintain examination standards do so *through their pronouncements* of comparability, as long as those pronouncements are generally accepted by society. According to the value judgement theory, there are no theoretical limits upon linking, only practical ones (related to what the empowered arbiters are prepared to claim and what society is prepared to accept).

This presents a direct challenge to the general model of linking presented earlier. At the heart of this challenge is an explicit rejection of the premise of grounding meaning via a linking construct. In contrast, Cresswell proposed that the judgement of equivalent value defies further 'epistemological justification' (Cresswell, 1996, p. 62), but that this does not actually hinder the effective use of examination grades. He was clear (Cresswell, 1996, p. 79) that judgements of comparability do not ascribe a property to the objects being judged (i.e., they do not construe levels of performance in terms of specific underlying student characteristics), they simply represent human responses to those objects (i.e., they purely and simply proclaim the value of particular levels of performance in different examinations).

Exploring the value judgement theory

To respond to the value judgement theory of linking, it is helpful to distinguish three different senses in which it appeals to value: value as the currency of examination results; the value judgement itself; and the principle of different underlying values contributing to grade boundary decisions.

Value as currency

In the theory proposed by Cresswell (1996), value is interpreted most fundamentally in the sense of currency. In accepting the pronouncement of an empowered arbiter of comparability, and in using examination grades as though comparability existed, users *create* the social fact of comparability (yet, at no point is there a formal requirement for agreement as to what comparability is actually supposed to mean). Thus, a grade has a certain currency as long as society agrees to treat it as though it had a certain currency. According to Cresswell, this even allows us to make sense of the requirement to link standards across examinations for different subject areas. It applies more generally, though, to all instances of linking.

While the currency analogy is certainly of philosophical interest, it can also result in apparent contradiction. The analogy suggests that, as long as users act as though a link has been correctly established, then comparability has genuinely been achieved. This results in the paradox that a link between scores which society accepted—following the pronouncement of its empowered arbiter—but then at a later date came to see as mistaken, would be genuinely valid (i.e., correct) for as long as the underlying error or false assumption remained unnoticed, and then genuinely invalid (i.e., incorrect) once the error or false assumption had been recognized. If this seems an unpalatable conclusion, then the notion of value as currency is insufficient to ground the value judgement theory.

The notion of value as currency would be deemed insufficient by anyone who wished to acknowledge the possibility that an arbiter of comparability might follow the broad due process of linking standards yet still fail to link standards effectively.

The value judgement itself

Perhaps it is not the use of results (as though comparability existed) that truly grounds the notion of comparability, but the nature of the value judgement itself. Perhaps society is prepared to act as though results are comparable because it is prepared to accept the judgement of a person/committee/institution whom/which they deem suitably qualified to make such judgements. So, society trusts the empowered arbiter to make *meaningful* judgements, even though this meaning may remain implicit rather than explicit. This does seem to be acknowledged by Cresswell (1996): 'acceptance is likely to be forthcoming from any particular user on a continuing basis as long as the awarders' evaluations differ only to some small extent from their own judgements, however informal or uninformed the latter may be' (p. 80). Importantly, then, trust would be rescinded if the presumed meaning of comparability seemed not to have been achieved, despite the prior pronouncement of comparability by the empowered arbiter.

On the other hand, by emphasizing the nature of the value judgement itself, we are at least implicating an underlying linking construct. And this seems antithetical to Cresswell's insistence that the value judgement proscribes further interpretation, as explained above. According to Cresswell, the task of the arbiter is to provide non-specific comparability judgements (i.e., the examination performances are of equivalent value) rather than specific comparability judgements couched in terms of a linking construct (e.g., the examination performances correspond to students of comparable attainment).

So, exactly what is it that motivates Cresswell's appeal to the theoretical construct of value judgement as a basis for linking standards? One reason seems to be a supposition that technically defensible linkages often cannot be made (when couched in terms of a specific linking construct such as attainment) so there is no option but to appeal to an explicitly non-technical perspective.[11] However, a further, or perhaps a related, reason seems to invoke the final sense of value.

Differing underlying values

In his most recent publication, Cresswell (2003) distinguished between the use of public examinations to qualify individuals and the use of the data they provide to monitor schools and the education system more generally. He proposed that the former use requires a 'comparable outcomes' perspective on comparability, to be fair to students, while the latter use requires a 'comparable performances' perspective to work at all.

Cresswell explored the implications of these two perspectives in the context of the new framework for A level examining, Curriculum 2000 (which, as noted above, represented a major change of curriculum and assessment arrangements). While the comparable outcomes perspective required that the first Curriculum 2000 candidates 'should receive, as a group, comparable grades to those which they would have

received had they followed the old courses' (Cresswell, 2003, p. 14), the comparable performances perspective required that 'performance similar to that awarded a grade x in the past should be awarded a grade x in the new examination' (Cresswell, 2003, p. 15).

Crucially, he noted how adverse effects arising from the introduction of new courses appeared to have resulted in a reduction in the quality of performances overall. If true, the two perspectives would result in radically different consequences for students' grades. Cresswell came down firmly in favour of the comparable outcomes perspective, since only this could be fair to individual students whose results would be used by selectors (who would, potentially, be selecting between pre- and post-Curriculum 2000 students).

Clearly, then, the different uses to which results may be put hold potentially conflicting implications for approaches to linking standards; more precisely, the different uses imply different definitions of comparability. In the terminology of the general model, *different users and stakeholders value different linking constructs*, wanting to draw different inferences from comparability.

Again, though, if it were possible to agree upon a single definition for comparability, i.e., to agree upon a particular linking construct, then why the need to defer to the more nebulous idea of value judgement? The answer seems to reside in a belief that it is not possible to agree upon a single definition: different users wish to use examination results for different purposes, and there are no strong grounds for deciding between those competing values/definitions:

> Ruling some such challenges out of court by insisting on a particular definition would be autocratic and, depending upon the size, power and influence of the particular constituency involved, only partially effective. Thus, choosing one of the definitions of standards discussed so far, even if it could be defended theoretically, would not be practically sensible. (Baird *et al.*, 2000, p. 224)

So, the value judgement theory of linking seems to divide into a series of claims. First, it is often technically impossible to link standards between examinations, so comparability can have no formal meaning in these circumstances (e.g., across subjects, or within subjects following major syllabus change). Second, there is a social obligation to link standards even when this cannot technically be achieved. Third, if comparability is to be achieved, then it can only be done by proclamation, and—if this proclamation is to be generally accepted—it must respect the multiple uses of examination grades. Fourth, this necessitates a procedure through which empowered arbiters make non-specific judgements of equivalent value (where the perceived value of an examination performance is relative to each arbiter's personal take on comparability).

Contrasting approaches to linking standards

The value judgement theory seems to necessitate a quite 'post-modern' approach to linking standards across examinations: the contest model of grade awarding.

The contest model

In the early 1980s, Christie and Forrest (1981) described the process of linking standards between public examinations in England as a contest between different interest groups, each elevating different definitions of comparability. In particular, they distinguished between those promoting attainment-based definitions (i.e., linking constructs) and those promoting ability-based definitions (i.e., linking constructs). Baird *et al.* (2000), writing two decades later, characterized grade awarding procedures in a quite similar manner. Thus, members of grade awarding committees come to each grade awarding meeting with their own implicit or explicit values/definitions. They debate between themselves, and with officers of the examining board who bring their own implicit or explicit values/definitions, before making recommendations to the examining board Accountable Officer on cut-scores that would link standards (it is actually the latter who is officially empowered to pronounce comparability, presumably in the light of her own particular values/definitions).

The approach recommended by Baird *et al.* (2000) appears very pragmatic: stakeholders use examination results for multiple purposes; these purposes prioritize different definitions of comparability; so a solution must be found which somehow recognizes all of the purposes and all of the definitions. Hence, the contest model. Metaphorically speaking, the linking imperative could be seen as an attempt to find a vector that maximized the validity of the full range of possible inferences. Wiliam (1996) characterized it slightly differently, envisaging users' needs as needles on a dial; as long as none of the needles entered the 'red zone', the resulting standard would be acceptable, if not ideal. Given this logic, some linking decisions would tend to support attainment-related inferences, some would tend to support ability-related inferences, some would tend to support IQ-related inferences, and others might tend to support none. Yet, the inferences that would follow (or would not follow) from each linking decision would never be transparent to users. In effect, the lack of a linking construct would prevent inferences from being drawn from linked scores.

The diktat model

At the other end of the spectrum lies an approach to linking that is differently pragmatic, and differently democratic; its premises can be characterized as follows. First, it is usually technically possible to link standards between examinations in a variety of ways, such that comparability can be given at least a plausible meaning. Second, in circumstances where comparability cannot be given at least a plausible meaning (i.e., where no plausible linking construct can be identified) there is a social obligation to be explicit about this and not to allow users to act as though meaning existed. Third, where multiple purposes imply multiple conflicting definitions of comparability, these purposes should be prioritized. Fourth, standards should be linked according to the definition which best supports the primary purpose; if this proscribes other uses, then other means should be sought for achieving those ends.

This approach recommends explicitly opting for a single definition of comparability (a single linking construct) in advance of any attempt to link standards. It might therefore be termed a diktat model, since the battle between definitions would happen only once, in advance of the very first linking exercise. This decision is ultimately a political one, presumably made at the highest level, with guidance from experts in assessment. The cost of this approach would be the need to make explicit that certain kinds of inference from comparability will not be warranted.

Choosing between approaches

Choice of the contest model over the diktat model seems to appeal to a notion of social defensibility, beyond that of rational defensibility. It appears to provide a way of keeping all stakeholders on board, by taking all of their perspectives into account when linking standards, at least partially (or, perhaps, ostensibly). However, the cost of the contest model is its inherent lack of specification. The distinction between right and wrong—in terms of both procedure and product—is blurred. This would seem to be a major obstacle to transparency, a major threat to public understanding and, potentially, also a threat to fairness for at least some candidates. Since the contest model does not really support the idea of 'valid inference' from linked scores, it does not lead to clear guidance on how results should or should not be used. The ultimate risk is that users may draw inappropriate inferences from comparability, and may misuse examination results, without even realizing that they are doing so. There is also the risk that, if assessment procedures and products were suddenly brought into question, the boards would have no purely rational basis for defending them. This could result in a serious loss of public confidence in the system.

The diktat model, on the other hand, seeks a rational basis for comparability, related to what has been stated as the primary purpose for which the assessment results will be used. Its aim is to maximize the validity of inferences in relation to that specific purpose. It does not necessarily proscribe the use of results for other purposes. However, it does make clear that comparability may not support those other purposes quite as well. Indeed, in proscribing certain inferences, it might effectively proscribe certain uses.

The central question is whether it is better to maximize rational defensibility (in relation to a single function and linking construct) than to downplay, or to abandon, rational defensibility and to defer instead to social defensibility.

Conclusion

> The public is clamoring for the measurement profession to develop such linking relationships, especially for tests with identical or similar titles. In fact, it appears to me that the public thinks that if the best people in the profession work hard enough they can obtain linking relationships that are strong enough to deflect all reasonable criticisms. Nothing could be further from the truth. (Brennan, 2001, p. 15)

All around the world, measurement professionals are called upon to create linking relationships between all sorts of different kinds of tests. Although we often

oblige—even when this stretches the idea of linking to its limits—we do not always succeed in explaining exactly what we have managed to achieve, i.e., we often fail to support users and stakeholders in drawing valid inferences from linked scores.

Part of the reason for this seems to be a lack of clarity within the measurement profession on how to conceptualize the essence of linking. This paper introduced the idea of a linking construct as an organizing principle for making sense of linking in different contexts. It proposed two key principles concerning the limits of linking:

1. linking can only be achieved if a plausible linking construct can be identified (even if this construct can only be roughly defined and/or loosely defended);
2. inferences from linked scores can only be drawn in terms of the linking construct.

While this may rate as no more than common sense, to some, the important point is that any discussion of linking—particularly discussions with policy-makers and members of the public—must be explicit concerning the linking constructs involved and the kind of inferences that do, *and do not*, follow from linked scores. If we fail to be explicit about such matters then myths of comparability will proliferate; for example, the myth that examination standards can be carried forward un-problematically across decades in time, despite major curriculum and assessment change. Indeed, had the English awarding bodies and regulatory authorities been more open concerning the inevitable limitations of linking with the introduction of Curriculum 2000, then the A level crisis of 2002 might never have occurred at all.

Although the general model of linking might appear commonsensical to some, others might consider it plain wrong. One of the most important linking theories of recent years—the value judgement theory—would seem to suggest just this. A central claim of the value judgement theory, when applied to complex linking contexts, is that defensibility ultimately requires compromise between stakeholders with potentially different definitions of comparability. If operationalized—as in the contest model—comparability could not be understood in terms of a single linking construct and no clear inferences would follow from linked scores.

Clearly, given the general model proposed earlier and subsequently expressed reservations, I find the contest model unsatisfactory. The alternative can be described as a diktat model, in which either a single linking construct is defined in advance or (where no plausible linking construct can be identified) linking is declared impossible.

Acknowledgements

I am very grateful to many colleagues (from the UK awarding bodies, the National Foundation for Educational Research, the Qualifications and Curriculum Authority and elsewhere) who have challenged and encouraged the ideas that eventually found expression in this paper, and especially to Mike Cresswell and Jo-Anne Baird. None of the views expressed should be taken to represent those of my present employer, the Qualifications and Curriculum Authority.

Notes

1. The first Curriculum 2000 A levels were awarded in August 2002. However, the release of results was greeted with rumours that students had been downgraded by the boards to avoid substantial pass rate rises, which was fuelled by claims that 'straight A' students were being awarded Ungraded on coursework units. The Education Secretary, Estelle Morris, ordered a public inquiry, to be headed by the former Chief Inspector of schools, Mike Tomlinson. The Tomlinson report uncovered an element of confusion between the boards and the regulator, and recommended remedial action, including the re-grading of certain units. By the beginning of October 2002, the Chairman (and acting Chief Executive) of the regulatory authority, Sir William Stubbs, had been forced to step down. By the end of October, the Education Secretary, Estelle Morris, had resigned. Ultimately, only 6 AS and 12 A2 unit grade boundaries were revised, resulting in 1,945 students receiving at least one revised A or AS grade.
2. The implication of the general model is that any straightforward standardization of attainment results, such as the conversion of raw scores to percentile ranks, would not count as linking (since there is no linking construct and, therefore, the scores are not linked in any *meaningful* sense).
3. Referring to Figure 1, the level of (mathematics) attainment in test 1 which corresponded to a mark of 4 might correspond to an average IQ of 100; if so, then the level of (chemistry) attainment in test 2 which corresponded to a mark of 6 should also correspond to an average IQ of 100 (for the scores to be linked).
4. Prior to 1998, the scaling process relied on students taking an additional test: the Australian Scaling Test. From 1998 onwards, an Average Marks Scaling system has been employed, where ability is not measured through a separate test, but is estimated through an iterative statistical process which models scores across combinations of subjects studied. Publications which provide information on the scaling process can be found on the website: http://www.curriculum.wa.edu.au
5. The idea of a hierarchy is less evident in Mislevy's original formulation; and he noted (of projection) that 'This is not merely a matter of stronger or weaker information but of qualitatively different information' (Mislevy, 1992, p. 24).
6. Maxwell (2002) makes a distinction between 'panel moderation' and 'peer moderation', arguing that only the latter is truly social (since the standard is negotiated), while the former is merely bureaucratic (since the standard is imposed).
7. Note that, although both Mislevy (e.g., Mislevy, 1992, p. 23) and Linn (e.g., Haertel & Linn, 1996, p. 69) discussed the role of value judgement in social moderation, neither developed this into a theory of linking, *per se*.
8. In the A level context, for comparability between 1951 and 2005 to be discussed, a linking construct would need to be posited (for each subject) which encapsulated that which was common to the assessed constructs in both years (and, presumably, the intervening years as well). Given experience from comparability studies covering much shorter periods in time (e.g., Christie and Forrest, 1980) this would be a daunting prospect, to say the least.
9. Curriculum 2000 A levels generally require students to study for three A1 units and three A2 units. Students typically study A1 units during year 1 and, if they also opt for examination in year 1, are able to 'cash in' these units for the award of an AS qualification. A2 units are normally studied in year 2. Once these have been examined, all six units may be 'cashed in' for the full A level award.
10. The judgemental linking principle had been central to school examining in England for decades prior to the introduction of the A level. Crofts and Jones (1928), for example, described how it operated in relation to the School Certificate.
11. This might actually be challenged with reference to the general model presented earlier, given the potential for linking constructs to differ significantly from assessed constructs. Of course, the 'no option' proposition also rejects what might be considered a more defensible stance: to pronounce that comparability cannot meaningfully be achieved in certain contexts.

Notes on contributor

Paul Newton is a Principal Assessment Researcher at the Qualifications and Curriculum Authority. His interests lie in the establishment of defensibility arguments concerning large-scale educational assessments and in the public understanding of assessment.

References

Baird, J., Cresswell, M. & Newton, P. E. (2000) Would the real gold standard please step forward? *Research Papers in Education,* 15(2), 213–229.

Baker, E., Sutherland, S. & McGaw, B. (2002) *Maintaining GCE A Level standards: the findings of an independent panel of experts* (London, QCA).

Brennan, R. L. (1998) Misconceptions at the intersection of measurement theory and practice, *Educational Measurement: Issues and Practice,* 17(1), 5–9, 30.

Brennan, R. L. (2001) Some problems, pitfalls, and paradoxes in educational measurement, Career Award Address to the *Annual Meeting of the National Council on Measurement in Education*, Seattle, 11–13 April.

Christie, T. & Forrest, G. M. (1980) *Standards at GCE A-level: 1963 and 1973* (London, Macmillan Education).

Christie, T. & Forrest, G. M. (1981) *Defining public examination standards.* Schools Council Research Studies (London, Macmillan Education).

Cox, C. B. (1975) Educational statistics, in: C. B. Cox & R. Boyson (Eds) *Black paper 1975: the fight for education* (London, J. M. Dent & Sons Ltd.), 36–39.

Cresswell, M. J. (1996) Defining, setting and maintaining standards in curriculum-embedded examinations: judgmental and statistical approaches, in: H. Goldstein & T. Lewis (Eds) *Assessment: problems, developments and statistical issues* (Chichester, John Wiley & Sons), 57–84.

Cresswell, M. J. (2003) *Heaps, prototypes and ethics: the consequences of using judgements of student performance to set examination standards in a time of change* (London, University of London Institute of Education).

Crofts, J. M. & Jones, D. C. (1928) *Secondary school examination statistics* (London, Longmans, Green).

Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W., Hemphill, F. C. (Eds) (1999) *Uncommon measures: equivalence and linkage among educational tests* (Washington DC, National Academy Press).

Haertel, E. H. & Linn, R. L. (1996) Comparability, in: G. W. Phillips (Ed.) *Technical issues in large-scale performance assessment* (Washington DC, US Department of Education, National Center for Education Statistics), 59–78.

Kolen, M. J & Brennan, R. L. (2004) *Test equating, scaling, and linking: methods and practices* (2nd edn) (New York, Springer Verlag).

Linn, R. L. (1993) Linking results of distinct assessments, *Applied Measurement in Education,* 6(1), 83–102.

Maxwell, G. (2002) *Moderation of teacher judgments in student assessment* (Queensland, Queensland School Curriculum Council).

Mislevy, R. J. (1992) *Linking educational assessments: concepts issues, methods, prospects* (Princeton, NJ, Educational Testing Services).

Mislevy, R. J. (1993) Foundations of a new test theory, in: N. Frederiksen, R. J. Mislevy & I. I. Bejar (Eds) *Test theory for a new generation* (Hillsdale, NJ, Lawrence Erlbaum Associates), 19–39.

Orr, L. & Nuttall, D. L. (1983) *Determining standards in the proposed single system of examinations at 16+* (London, Schools Council).

Partis, M. T. (1997) *Scaling of tertiary entrance marks in Western Australia*. Available online at: http://www.curriculum.wa.edu.au/files/pdf/114537_1.pdf (accessed 9 March 2005).

QCA (2004) *GCSE, GCSE in vocational subjects, GCE, VCE and GNVQ and AEA code of practice 2004/5* (London, QCA).

Searle, J. R. (1995) *The construction of social reality* (New York, Free Press).

Secondary Examinations Council (SEC) (1984) *The development of grade-related criteria* (London, SEC).

Wiliam, D. (1996) Standards in examinations: a matter of trust? *The Curriculum Journal,* 7(3), 293–306.