

Setting standards: Fitting form to function

Graham Samuel Maxwell

Educational Consultant

**Paper presented at the 34th IAEA Annual Conference
Cambridge UK, September 2008**

Abstract

This paper explores some issues concerning the representation and application of standards. The term 'standard' has a variety of meanings, with different consequences for practice. A key distinction is 'content standards' versus 'performance standards'. Another distinction is a 'range of standards' versus a 'targeted or expected standard'. Standards can be represented by cut-scores or ordered categories (or a combination of these). The traditional psychometric approach sees standard setting as an empirical exercise dependent on the assessed cohort performance; the emergent decision-based assessment approach sees standard setting as a judgement process dependent on prior description and example. Also, standards representing comparative performance on a particular task or course can be different from standards representing developmental improvement over time. Clearly, standards need to be represented differently for different purposes—form fitted to function. There is also a need to invent new ways of representing and managing standards that fit a personalised approach to learning.

Types of standards

The term ‘standards’ figures frequently in discourse on educational assessment. However, one of the difficulties in discussing standards is that the term can have many different meanings. These different meanings are not pedantic and benign. Failure to clarify which meaning is intended often results in a patina of agreement or a Babel of confusion. Having some way of identifying different meanings can lead to better communication. Also, it allows us more easily to examine their hidden assumptions and implications. This may lead us to some changes of direction in educational practice as we clarify options and invent new possibilities.

There are at least five types of standards (Maxwell, 2002a; forthcoming 1):

1. Standards as moral or ethical imperatives (what someone should do)
2. Standards as legal or regulatory requirements (what someone must do)
3. Standards as target benchmarks (expected practice or performance)
4. Standards as arbiters of quality (relative success or merit)
5. Standards as milestones (progressive or developmental targets).

Standards as moral or ethical imperatives (type 1) indicate something that is desirable but lacking regulatory force. That is, they offer principles or guidelines. Examples are the assessment for learning guidelines of the UK Assessment Reform Group (2002) and the various standards for school subjects in the USA (for example, NCTM, 2000). These are standards for schools and school systems to adopt in constructing and delivering their curriculum. At the student level, standards as moral or ethical imperatives typically relate to their moral or ethical behaviour. While some of these, such as no cheating or plagiarism may be regulatory requirements, others such as being polite and conscientious are merely encouraged (though rewarded informally).

Standards as legal or regulatory requirements (type 2) involve some form of compulsion. There are consequences for failure to satisfy the requirements. At an institutional level, an example is the ISO Standards (see the Standards for Statistical Methods, ISO, 2008, though this is only one of thousands of such standards). In this case, the standards must be satisfied to gain the ISO imprimatur. At a student level, the (minimum) requirements for being awarded a certificate or degree are often referred to as the standards for gaining the award. Usually, these kinds of standards involve a checklist of all the things to be satisfied for being awarded the certificate.

Standards as target benchmarks (type 3) define an expected or typical outcome (for example, a particular level or quality of performance). These can be requirements for a ‘pass’ or ‘satisfactory’ or ‘sound’ (or in the training sphere ‘competent’). Typically, this goes beyond the checklist typical of the previous two types of standards; what is needed is some representation of the point along a continuum that defines a minimum acceptable level.¹

The last two types of standards are concerned with differentiated levels orisarvaha sst tted leve34.15818ff39(n) at,

statements that each reference one criterion. Reading across a row in the matrix, we should be able to recognise successive increases in quality from one level to the next.⁸

How explicit should a rubric be? That depends on the circumstances of its use. The point is to provide sufficient detail about the desired performance characteristics to be able to make a consistent judgment about which level best characterises the performance.⁹ Necessarily, levels are broad categories and therefore somewhat fuzzy and imprecise. Yet, any degree of explicitness about the nature of performances typical of each standard sharpens the focus, fosters consistency and improves communication. Sometimes finer distinctions are made, for example, high, middle and low within each level, though usually without specific descriptors.

A criteria-and-standards matrix produces a performance profile: a specific level on each criterion. Sometimes, an overall level is reported, combining the performance across all criteria. There are two ways this can be done, by aggregating scores or by judgment. With score aggregation, a further decision is needed about cut-scores for the overall standards. With judgment, where performance differs across criteria, a best-fit judgment is required that allows trade-offs across the criteria. In both cases, the meaning of the overall standard is rather ambiguous, at least in the middle categories, because the trade-offs differ across students and the aggregate grade descriptions only characterise typical performance. This may be adequate for certification, where usually only the standards labels (e.g., A–E) are reported. However, for feedback (formative) purposes, the detailed profile of performance on separate dimensions is essential and the overall performance level is too vague and general to be useful.

How many criteria should a rubric have? This can be approached epistemologically, that is, through consideration of the inherent dimensions of the subject matter and/or the nature of the task (or portfolio, etc.) being assessed. However, there are pragmatic considerations too. It is difficult to keep very many characteristics in mind at the same time (Miller, 1956). For more than five criteria it is best to develop a hierarchical structure (sub-criteria embedded within main criteria) but the more the detail the more the cognitive demand anyway.

Despite the apparent benefits (and increasing popularity) of rubrics,¹⁰ there are some difficulties. Some of these have already been mentioned: they can only be interpreted in relation to a specific context; standards descriptors are necessarily fuzzy and imprecise; aggregate standards are ambiguous; lower standards tend to signify failure or deficiency; standards descriptors, even when accompanied by exemplars, are insufficient to ensure common interpretation and usage, which requires training and moderation.

There are some additional difficulties. First, there is a tension between generic and specific descriptors. Generic descriptors maintain consistency of language across different contexts (including different years); this creates interpretive difficulties, with the labels and descriptors referring to different observable features of performance (for example, ‘excellent’ refers to quite different performance in Year 5 and Year 12, or at the beginning and end of a course). Specific descriptors are tailored to the specific assessment event; this clarifies their meaning for that context but makes them one-off wonders.

Second, descriptors are often tautological, that is, they merely repeat the qualitative language of the standards labels (for example, limited, sound, high); alternatively, they are often vaguely quantitative (for example, few, some, many, all; or moderately, generally, very).

⁸ Wiggins (1998) distinguishes between holistic and trait-analytic rubrics. The former correspond with

between each level (that is, 24 categories overall). The levels represent typical progress at two-year intervals from Preparatory to Year 10.

VELS uses term ‘standards’ in three different ways: content standards—the knowledge and skills expected to be taught in each of the strands; developmental standards—the levels and progression points for assessing progress; and expected standards—the typical or targeted level for each year level. As a further complication, the Australian Government now requires all schools to report student performance to parents each semester on an A–E scale (Commonwealth of Australia, 2005). Under VELS, Victoria maintains an expectation that schools will continue to assess the standard (level) and progression point reached by each student, with computerised conversion to an A–E grade appropriate for each year level (and representation of the levels in terms of their year of typical attainment). These characteristics of VELS are both visionary and realistic, adhering to the benefits of charting student progress developmentally but acceding to governmental and parental expectations of merit grading within year cohorts. Whether this will be successful or confusing remains to be seen.¹³

In general, there are some clear benefits in using developmental standards:

- They provide explicit steps and targets for developmental progress.
- There is a language and expectation of progress.
- They make evident to students the progress they have made.
- They provide clear targets for further learning
- Student spurts and plateaus can be seen as natural and expected.

There are a couple of caveats. First, as for content standards, developmental levels descriptors depict typical performance but will not fit every student. Second, as for merit standards, levels can be holistic (cover several dimensions) with similar problems of best fit (tradeoffs) and imprecise meaning. Third, how slower progress is handled will affect student self-perceptions. There is a clear need for flexibility in using developmental standards.

There are also some challenges for developmental standards: how to promote acceptance of a new and different framework for reporting progress that breaks with traditional concepts of grading ; how to combine developmental levels with expected levels without reverting to a language of failure; how to talk about slower progress without creating negative self-perceptions; and how to develop school structures to support developmental progression.

Conclusion

This paper has explored some different meanings of the term ‘standards’ in educational assessment. It indicates a variety of ways in which we currently talk about and represent standards, each serving a different purpose and having different strengths and limitations. We should not confuse one meaning and purpose with another. We can reduce confusion and improve communication by being clear about the type of standard to which we are referring. This is a matter of fitting form to function. Rather than attempt to shoehorn one type of standard into all situations, that is, assume that ‘one size fits all’, we should recognise the strengths and limitations of each type of standard and tailor our practice accordingly.

However, that is not the end of the story. The analysis in this paper also suggests that there are some critical issues to address in relation to the way we talk about and frame educational standards. In particular, standards of any kind assume a ‘typical student’ (to set the pace and the expectations) and a ‘typical range of students’ (to represent different degrees of coping with the standard pace and expectations). The consequence is that we force-fit students to

¹³ Referents for A–E in Victoria are defined relative to the expected level for each year: well above, above, at, below, well below. Other Australian states and territories have adopted similar generic descriptors (for example, excellent, good, satisfactory, limited and poor) that offer crude comparative indicators (almost certainly inconsistently applied by different teachers and schools) but convey no information about what the student actually knows or can do. This may be sufficient for some purposes but not others.

