Teachers’ assessments in a standards world

Gordon Stanley

Introduction

Today most assessment schemes in education systems around the world are attempting to report in a more meaningful manner than simply grades or marks alone (Tognolini & Stanley, 2007). Increasingly assessment schemes are required to provide evidence that is credible to employers and end-users. At the heart of contemporary schemes is the reporting of student achievement in terms of a series of standards or verbal descriptions that indicate the characteristics of the learning, i.e. “what a pupil knows and can do”.

A standards world has generally required that mandatory curriculum be specified in more detail than formerly. This has been done with the National Curriculum in England, and State level curriculum in the USA and Australia. The level of prescription has varied across systems, ranging from very specific content in US state standards to broader outcome statements in the UK and Australia.

At the same time as this move to a focus on reporting student results in terms of standards, concern about student progress and school system accountability has led many countries to introduce or re-introduce system wide testing. A major design requirement for quality assessment, whether involving external tests or teacher judgments, is the need to ensure alignment to curriculum goals and expected outcomes (Biggs & Moore, 1993). We need the assessment to report progress towards and/or attainment of the curriculum outcomes.

Those school systems which have external tests and examinations at system level have been criticised constantly for such tests inevitably narrowing the curriculum and leading to ‘teaching to the test’ (Volante, 2004). However, it cannot be assumed that teacher assessment necessarily avoids the problems associated with external testing without some assurance of consistent professional standards for effective teacher assessment. As Phelps (2008) has argued, teachers may teach a narrow curriculum by selecting their favourite bits for emphasis and may employ classroom assessment practices which create similar pressures.

It may be in the nature of national and system level assessment that such corruption of the intent of curriculum delivery occurs, especially when student results are used for performance management of teachers and schools. To counteract such trends requires a strong professional ethic and system level investment in developing and maintaining high quality teaching and assessment regimes.

School systems need to support processes to develop assessment competence in teachers, a task that is often not maintained effectively beyond an initial implementation cycle. The professional status of teachers is enhanced when assessment occurs at the level of the classroom (Baker, O'Neil and Linn, 1993; Black et al, 2011). Moreover, teachers’ involvement in assessment moderation and
standard setting are invaluable in helping them to assign performance levels correctly according to national standards.

For effective teacher assessment for an education system there are a number of issues that need to be addressed, ranging from workload to ensuring consistent and quality engagement by teachers and students.

**Manageability and Evidence Records**

Tensions arise between the activities performed by teachers in embedding assessment into their daily teaching and the requirements for quality assessment. Classroom assessment practice can involve judgments based on formal written work, such as essays and assignments of varying structure and content, and those based on dynamic interactions in classroom performances.

Inherent in different classroom teaching and learning situations are varying opportunities to observe and record information to inform judgments about student achievement. Differences in the degrees to which teachers assist in eliciting evidence can have a significant effect on outcomes. These differences include aspects such as structure and prompting, editorial assistance, resources provided and time allowed for the task.

The amount of evidence collected and how it is recorded is a significant issue. In an attempt to provide transparency many systems have over-specified assessment criteria to the point where they are unmanageable in practice. This was a criticism expressed by teachers in the pilot studies for the APP scheme in England (Stanley, et al, 2009). Practices differ in the extent of data collection and recording ranging from detailed assessment protocols to ‘on-balance’ judgments of attainment of assessment criteria.

Practical tradeoffs need to be made with respect to the degree of task standardization, the number of tasks on which assessment is based, the degree of recording and storage of information to justify the teacher judgment and the impact on teachers’ workloads. Since a prime focus of teacher assessment is to provide an opportunity to make judgments about student progress and to guide the next step in learning it is important that the assessment regime is manageable in terms of teacher workload and provides timely feedback (Hattie & Timperley, 2007).

**Effectiveness**

The effectiveness of teacher assessment models needs to be evaluated in the policy context in which they are operating so that effectiveness can be measured against the purpose/s required by the education system. Much of the literature on the reliability or validity of teacher assessment has been based on relatively small scale studies embedded in contexts which may not fit well into a system wide reporting and accountability framework.

Where teacher training in assessment has been strongly embedded teacher assessments have found to be more reliable (see Falk et al, 2007; Lindstrom, 2007).
than when they are not (Koretz et al, 1994). In her literature review Harlen (2005), reported that the reliability of teacher assessments has been found to be within the range of the reliability of traditional tests. We came to a similar conclusion in our literature review as part of our APP study for QCA (Stanley, et al, 2009).

Kane (2010) and Lamprianou and Christie (2009) have pointed out that when teacher assessments are spread out over time and then aggregated with a final test score into an overall judgment of student performance we need to be careful about assumptions of all variability in scores being attributed to error. It is reasonable to expect student’s levels of skill to develop over time as a function of instruction and practice and the rate of such development may vary significantly across individuals. That is why some examination authorities advise schools to weight school assessment components closer to the summative external examination higher than earlier assessed components.

Alignment to standards

The research literature often assumes that middle to high correlation between two sets of ratings for a sample of portfolios of student work is sufficient to establish that the portfolios have been correctly aligned to standards. However this is not the case, as pointed out by Linn et al. (1992).

Supovitz, MacGowan and Slattery (1997) studied the mean scores allocated to portfolios in reading and writing and reported that for all six areas studied, classroom teachers awarded a higher mean than external markers, although the small sample sizes prevented some of these from being statistically significant. Shapley and Bush (1999) looked across ten areas for children in kindergarten and Grades 1 and 2 and found classroom teachers awarded higher scores than the external markers. The effect sizes reported were quite large, with eight of the ten being greater than 0.4. The authors report that “the discrepancies…may reflect teacher bias; that is, the classroom teachers are more lenient in assigning scores” (p. 119).

In a recent study of School Certificate assessment at Year 10 in New South Wales, MacCann and Stanley (2010) compared school assessment grades in science with grades on the external test for science. The school assessment programme is reported with respect to performance descriptors for five levels as is the external test. The test performance descriptors are similar to those for school reporting though somewhat more narrowly focused on scientific literacy.

The alignment of scores on school assessment to grade descriptors is made at school level and submitted to the examination board whereas the external test score alignment to grades is made centrally by a small team of experienced judges. With an annual candidature of around 80,000 one might expect reasonably stable results.

Two points from that study are of interest. First school assessments varied little from 2001 to 2007, whereas there was seemingly random variation over time for the external grade setting. Schools assessment allocated around 13 per cent of the candidature to the top grade while the external judges allocated around 4-5 per cent. This suggests that schools with a richer source of evidence are more consistent in their alignment, even if their judgment of top achievement is more generous.
Lamprianou and Christie (2009) compared the results from school assessments in contexts where there were external examinations and where there were only school assessments. They found:

Where there is no external examination to constrain the teachers’ judgments the variance in scores is significantly smaller. Inspection of the raw data reveals that in every subject regardless of condition full marks were awarded, sometimes liberally. In just over half of the subjects without external examination however no mark awarded was below ten, a truncation of range that occurred in only one subject where there was an external examination (p335)

Such results may reflect a natural tendency for the classroom teacher to see their students in the best possible light, despite the training they received in aligning the portfolios to the statewide standards. Clearly there is a need for external moderation of some sort, depending on the degree to which the assessments are high stakes.

Performance descriptors and standards are tools for aligning student performance to a common currency. When there are apparent discrepancies in grade assignment we need to develop procedures to resolve these discrepancies. The validation of standards setting is an important issue for assessment authorities. While there are a number of technical options for implementation finding practical solutions that are feasible and cost effective is a major issue (Tognolini & Stanley, 2011).

Conclusion

Teacher assessment to be respected in a standards world has to demonstrate that it is based on evidence that is consistent with appropriate alignment to the curriculum standards expected by end-users. The advantage of teacher assessment is the capacity to ensure that evidence about student performance is well-grounded and supported by the teaching and learning programme at school level.

However for quality assessment to occur there needs to be real investment in developing assessment competency and ensuring common standards are being implemented. Some school systems have demonstrated that school based assessment meets the requirements for rigour in high stakes contexts. Such systems invest in professional development and moderate outcomes to achieve consistent results. Whether or not school based assessment can withstand accountability pressures without some external moderating test regime is an open question.

References


Black, P. et.al. (2011, in press) ‘Can teachers’ summative assessments produce dependable results and also enhance classroom learning?’ Assessment in Education


**Author email:** gordon.stanley@sydney.edu.au