Designing a fair and effective assessment system

Paper presented at the 2007 BERA Annual Conference as part of the ARG symposium
Future Directions for Student Assessment

Wynne Harlen, University of Bristol

This paper arises from an Assessment Reform Group (ARG) project, The Assessment Systems for the Future, which was funded by the Nuffield Foundation, 2003 – 2006. Being concerned with assessment systems means considering the procedures and regulations that determine the roles of assessment and how it is conducted for various purposes:

(i) to assist learning (formative)
(ii) to report regularly on the learning of individual students to parents, other teachers, students (summative)
(iii) to accredit achievement as a qualification for further education or employment (summative for certification)
(iv) to monitor standards of student achievement at regional, national or international levels (summative for system monitoring).
(v) as a contribution to evaluating education provision at college or school or local level (summative for accountability)

The first three of these purposes are concerned with the performance of individual students, whilst the information for purposes (iv) and (v) concerns the performance of groups and indeed not all students in relevant groups need to be assessed to provide information for these purposes. Further, since purposes (ii) to (v) are concerned with reporting on what has been learned and only (i) concerns using the information as part of the learning process, the purposes can be reduced to two: formative and summative.

The ASF project focused on summative assessment of students in the context of a system in which the results might be used for: internal school tracking of student progress; informing other teachers, parents, students; certification/accreditation or selection; monitoring national standards; evaluating school provision. Although assessment for the last two of these purposes is not used to make decisions that directly affect individual students, as in the first three purposes, nevertheless, using the results of summative assessment for accountability and monitoring can, and does, affect students through impact on teaching and the curriculum. Thus procedures used to achieve some purposes can have consequences for how well other purposes can be met. For example, there is research evidence that:

- When school accountability is based on external summative assessment data, this impacts on the way that teachers conduct their own internal summative assessment and on how they use assessment formatively (eg Pollard et al, 2000).
- How teachers carry out formative assessment can influence their own internal summative assessment practice (Black et al, 2003).
- Using student attainment data both for accountability and for system monitoring restricts the range of information that can be used for either purpose (ARG, 2006).
Often unwanted effects of one part of the system on another arise from piecemeal changes in policy that ignore the relationship of parts within the whole. These interactions, and those between assessment and the curriculum and teaching methods (pedagogy), are indeed at the heart of current and past arguments for reform in assessment systems (Black, 2001; Black and Wiliam, 2007).

**Criteria for evaluating components of an assessment system**

To address the question of how well each of the functions of assessment was being carried out by procedures adopted within a particular system, the ASF project identified four criteria for evaluation. One obvious criterion is that any assessment should be valid for its purpose, that it assesses what it is intended to assess. Another is that it should provide reliable data. But there are also other matters to be taken into account; in particular, in view of the interdependence of the various system components, the impact on other assessment practices and on the curriculum and pedagogy. Further, there is the use of resources; assessment can be costly, both in terms of monetary resources and the time of students and teachers.

Thus assessment for any purpose should provide information meeting the criteria of validity, reliability, desired impact, and good use of resources. Each of these criteria is now briefly considered before looking at how they apply to different ways of conducting student summative assessment.

**Validity**

In this context validity means how well what is assessed corresponds with the behaviours or learning outcomes that it is intended should be assessed. Various types of validity have been identified, most relating to the type of evidence used in judging it (eg face, concurrent, content validity), but there is general agreement that these are contained within the overarching concept of construct validity (Messick, 1989; Gipps, 1990). The important requirement is that the assessment concerns all aspects – and only those aspects - of students’ achievement relevant to a particular purpose. Including irrelevant aspects is as much a threat to validity as omitting relevant aspects. Thus a clear definition of the domain being assessed is required, as is adherence to it.

**Reliability**

The reliability of an assessment refers to the extent that the results can be said to be of acceptable accuracy for a particular use. It is often defined as, and measured by, the extent to which the assessment, if repeated, would give the same result. The degree of reliability necessary depends on the purpose and use of an assessment. When assessment is used formatively, it involves only the students and the teachers. No judgement of grade or level is involved; only the judgement of how to help a student take the next steps in learning, so reliability is not an issue. Information is gathered frequently by teachers who will be able to use feedback to the student to correct any mistaken judgement. However, high reliability is necessary when the results are used by others and when students are being compared or selected.
Impact
Here impact means the consequences of the assessment, often referred to as consequential validity (Messick, 1989). It concerns the inferences drawn from the assessment information in relation to the uses that are made of it. As noted earlier, assessment generally has an impact on the curriculum and on pedagogy, which is greater the higher the stakes attached to the outcomes of assessment, so it is important that any potential adverse effects are minimised. Assessment can only serve its intended purpose effectively if this is the case.

Resources
The resources required to provide an assessment ought to be commensurate with the value of the information for users of the data. The resources may be teachers’ time, expertise and the cost both to the school and to external bodies involved in the assessment. In general there has to be a compromise, particularly where a high degree of reliability is required. As well as the trade-off between validity and reliability, as discussed later, there is a limit to the time and expertise that can be used in developing and operating, for example, a highly reliable external test. Triple marking of all test papers would clearly bring greater confidence in the results; observers visiting all candidates would increase the range of outcomes that can be assessed externally; training all teachers to be expert assessors would have great advantages – but all of these are unrealistic in practice. Balancing costs and benefits raises issues of values as well as of technical possibilities.

The properties of tests and examinations
We take as given the need for summative assessment for the purposes of recording, reporting, certifying achievement, monitoring standards of student performance and that it has a role in the accountability of teachers and schools. It is also accepted that how the assessment is conducted for these purposes can have considerable impact on the curriculum and pedagogy, for example -

When the results of tests and examinations are used to pass judgments on teachers and schools, they also affect the ways in which pupils are taught. Given their importance, it is essential that results of summative assessment should reflect and influence school learning in the best possible way.
(ARG, 2006: 1)

Validity
High validity in summative assessment used for selection and certification is essential, first because it sends a strong message about what learning is valued and second, because of the high stakes attached to summative assessment for use external to the school. Current thinking, world-wide, emphasises the importance of developing in students types of skill, attitudes, knowledge and understanding, that are regarded as more important than accumulating large amounts of factual knowledge. Content knowledge can be found readily from the information sources widely available through the use of computers and especially the internet. What are needed are the skills to access these sources and the understanding to select what is relevant and to make sense of it. So we need to provide students with understanding of broad, widely
applicable concepts and the ability to use them to solve problems and make decisions in new situations.

These are not the only goals of education, of course. There are things that students have to know, such as spellings, number bonds, procedures, conventions and some names and facts that require practise and memorisation until they become automatic. These goals are probably effectively assessed by short tests of a conventional kind. They are also amenable to computer-based self-testing. The concern here is with assessing important goals that are not currently included in what is assessed. It is abundantly clear that unless the assessment includes higher level learning, the attention these goals are given in teaching will not match the rhetoric of their importance.

Tests and examinations have inherent limitations in relation to the assessment of any goals and these limitations are particularly severe in respect of these learning outcomes. They are poorly assessed in external tests and examinations. Since teachers’ internal summative assessment tends to emulate the external assessment (Pollard et al, 2000), this also fails to reflect these important goals.

Reliability

One reason for using tests is that the use of assessment results for external purposes demands that they are seen to be as ‘fair’ as possible. Fairness in this context refers to technical reliability or the extent to which results can be said to be of acceptable consistency or accuracy for a particular use.

Black and Wiliam (2006) list three sources of error in tests and examinations:

- Any particular student may perform better or worse depending on the actual questions chosen for the particular administration of the test.
- The same student may perform better or worse from day-to-day
- Different markers may give different marks for the same piece of work.

(Black and Wiliam, 2006: 120)

The obvious way to minimise the third of these errors is to ensure that marking depends as little as possible on human judgment. This favours multiple-choice items and machine marking. Clearly, items that require students to be creative, or present arguments, or show understanding of a complex situation, do not fit this description very well.

The second of the sources of error listed by Black and Wiliam (2006) is unavoidable. They note that ‘It is highly unlikely that the score that someone gets on one occasion would be exactly the same as on another occasion, even on the same test.’ (p 120). This source of error is generally acknowledged. However, it is the first of the three points that is less well recognised and imposes an unavoidable limit on ‘fairness’. A test or examination of any reasonable length can only sample a syllabus or programme of study. A different sample of items would be found easier or more difficult according to many differences between students such as the difficulties they encountered and their general engagement with the relevant activities during learning.

The error due to sampling can be quite considerable. For example, Wiliam (2001) estimated that the difference this would make in the ‘levels’ awarded to students in
the national tests in England at ages 7, 11 and 14. At age 14, if the test has an overall reliability of 0.80, the proportion of students given the wrong level is estimated to be 43%. Similarly, Gardner and Cowan (2005) show conclusively that there was a high incidence of misclassification in the tests used until recently at age 11 in Northern Ireland to select students for grammar schools. The error can be reduced either by increasing the reliability of the test without changing the length (which inevitably means narrowing the range of items) or increasing the length of the test to include a larger number of items. But the return for increased length is small in terms of the increase in accuracy and to make a substantial difference would require an unacceptably long test.

These sources of error apply to all tests and examinations, making it all the more surprising that the notion that tests are necessarily more reliable than other methods of assessment is so prevalent. As Black and Wiliam (2006) note ‘the public in general and policy makers in particular do not understand or pay attention to reliability. They appear to have faith in the dependability of the results of short tests when they are in fact ignorant of the sizes of the inescapable errors that accompany this and any other measure’ (p 119).

**The interaction of reliability with validity**

To increase the reliability of a test of acceptable length, error due to marking has to be minimised. This has the effect of narrowing what is assessed, thus reducing validity in relation to the assessment of goals such as application of concepts, problem solving and critical thinking. The corollary is that increasing the validity of an assessment in relation to such learning outcomes, through using open-ended and extended tasks, would increase the need for judgement in marking and so potentially reduce reliability. Thus in practice a test cannot have both high reliability and high validity; there has to be some trade-off between them.

Where the balance is best struck in this trade-off between validity and reliability depends on the use made of the results of assessment. Where fairness to individual students is involved, it would seem that preference ought to be given to reliability. But high validity is also important because of the impact on teaching and learning. Thus the emphasis has to be on increasing reliability in the interpretation of the evidence whilst maintaining validity at the highest level possible.

The interaction between reliability and validity, and the effect of item sampling, show particularly clearly in test items and assessment tasks designed to assess problem solving and other higher level skills. Complex tasks, set in authentic contexts are, on the face of it, of high validity, but the performance of students is strongly affected by the context and subject matter. For example, in the Assessment of Performance Unit (APU) national monitoring programme in England, Wales and Northern Ireland for annual monitoring in the 1980s, the science assessment included short practical investigations carried out by students in the presence of an administrator/observer. Each investigation took a student about 20 minutes to complete, so only a few could be administered. It proved impossible to form a score for ‘performance of investigations’ as the variation in students’ performance across different investigations was so large. Other research has arrived at similar findings - that students who perform well in one item will not necessarily do so in another item testing the same skills but in a different context (for example, Pine et al (2006)).
Although Yeh (2001) argues that it is possible to turn authentic tasks into fixed answer questions, the validity of such items is likely to be threatened when teachers are under pressure of high stakes use of results. Experience and research evidence (for example, Gordon and Reese, 1997) show that in such circumstance teachers can be very effective in training students to pass tests even when the students do not have the understanding or higher order thinking skills that the tests are intended to measure. Whether or not this is true for performance assessment is arguable. Resnick (1994) accepts that all assessment will be ‘taught to’ and so ought to exemplify the important outcomes and standards.

**Impact**

The interaction between validity and reliability already takes us into the area of the impact on teaching and learning of how assessment is conducted. Indeed this impact is not necessarily a side-effect, but can be a deliberate policy, as, for example, in the approach of the movement known as Measurement Driven Instruction (MDI), which is to raise the stakes so that schools are obliged to teach what is measured (Airasian, 1988).

The rationale for national testing is embodied in the slogan that ‘testing drives up standards’. In an extensive review by Tymms (2004) of national test results for eleven year olds in England from 1995 to 2003, the pattern found was a rise over the first five years (1995 - 1999) followed by no change from 2000 to 2003. The pattern was the same for mathematics and English. While some other data supported a rise from 1995-1999, it was noted that the data from the Trends in Mathematics and Science Surveys (TIMSS) showed no rise in mathematics over this period. While Tymms (2004) could identify several reasons why standards of tests may have changed over this time (mainly related to how cut-off scores for levels are determined when tests change from year to year) he concluded that the effect of teaching test technique (new to pupils of this age in 1995) and of teaching to the test are very likely to have accounted for a good deal of the initial change. This conclusion is supported by trends over time in other test regimes. For example in the USA Linn (2000) found ‘a pattern of early gains followed by a levelling off’ (Linn, 2000: 6) to be typical across States where high stakes tests are used.

However, for reasons already noted, using national test results to monitor standards provides a very limited view of students’ achievement. So we cannot really tell whether or not standards are changing. A more useful picture would be obtained by a sample survey, where teachers to not know which students will be tested and students in the same class will not in any case all be given the same items, so results would not be distorted by practising what is to be assessed. Moreover, a wide ranging survey which can include a far greater number of items than any one student can take would considerably reduce the task sampling error and allow a more valid range of goals to be assessed.

**Resources**

Estimating costs of assessment is notoriously difficult; the uncertainty of concepts, the complexity of variables and the variation in practice among schools mean that all reports have to be treated with great caution. Some tentative estimates (ASF, 2006) for the indirect costs in terms of teaching and learning time in English schools.
indicate that the end of primary school tests mean that teachers are each year spending the equivalent of about a week of contact time on preparing and marking tests, and students about three weeks of learning time. The actual time spent on national testing is much smaller than the time spent on regular tests given by the teacher. Other evidence (e.g. Reay and Wiliam, 1999; Pollard et al, 2000; OfSTED, 2005) would suggest that this is the result of teachers giving students practice tests. The direct costs, of providing, administering, invigilating, marking and reporting tests and examinations in primary and secondary schools was estimated at £370m in 2003 in a survey carried out for the QCA by PriceWaterhouseCoopers. The 2007 figure is probably at least twice this, since entrance fees increased by over 70% in the years between 2003 and 2006.

**Alternatives to tests and examinations**

The unavoidable limits on reliability and validity of assessing achievement through tests and examinations forces a search for alternatives. Fortunately there are several other methods of summative assessment, all depending on the fact that the experience that students need in order to develop the desired skills, understanding and attitudes also provide opportunities for their progress to be assessed. In some approaches evidence of achievement is cumulatively collected in portfolios; in others evidence from regular activities is supplemented by embedded tasks or optional tasks specially devised to ensure opportunities for students to use the skills and understanding to be assessed. The key factor is the judgment of the teacher, using criteria which generally reflect a view of progression in learning.

Using teachers’ judgments for summative assessment is not new; indeed it is well established in systems in several countries where it is used for important external assessment (Harlen, 2007). It is built into existing practice in the UK for assessing pre-school children and for older students for assessing vocational skills. However, other parts of the UK are ahead of England in reviewing and making radical changes in the assessment systems for the compulsory years of school that have been in place through most of the 1990s, changes that are taking systems in Scotland, Wales and Northern Ireland away from frequent testing and towards greater use of summative assessment by teachers.

*Validity of summative assessment using teachers’ judgments*

Over the period of time, such as a term or half year, for which achievement is being reported, students have opportunity to engage in a number of activities in which a range of skills, attitudes and understandings can be developed. These same activities provide opportunities for achievement to be assessed by the teacher. What is assessed becomes closely matched to the curriculum experienced. The problem of sampling does not arise, since it is possible for all learning activities and tasks to contribute to the assessment. In other words, the limitation of the restricted time that a test provides does not apply when assessment is teacher-based.

Methods of assessment based on observation during regular work also enable information to be gathered about processes of learning rather than only about products. Such information is useful to those selecting students for advance vocational or academic courses of study, where it is as important to know about whether
candidates have learned how to learn and are likely to benefit from further study as about what they have already learned.

Reliability
The reliability of teachers’ judgements is widely held to be low and there is research evidence of bias (Hoge and Butcher, 1984; Wood, 1991; Koretz et al., 1994; Harlen, 2005). But research also shows that when assessment criteria are well specified (and understood) teachers are able to make judgments of acceptable reliability (eg Rowe and Hill, 1996; Cumming and Maxwell, 2004). The moderation procedures that are required for quality assurance can be conducted in a way that provides quality enhancement of teaching and learning as noted below.

An issue that all summative assessments schemes based on the accumulation of evidence from everyday work must face is how to bridge between the detail of that work and the broad categories that exist in national or state standards or criteria used in reporting. The use of embedded tasks (as in the BEAR system (Wilson and Scalese, 2004)) largely avoids the problem by ensuring that the evidence collected from the tasks is designed to relate to the criteria for reporting. This is not the case, however, when evidence is gathered from regular activities. Harlen and James (1997) point out that information gathered and used formatively in the course of teaching and learning, as in ‘interactive’ formative assessment (Cowie and Bell, 1999) or assessment ‘in the ZPD’ (Allal and Pelgrims Ducrey, 2000) cannot merely be aggregated to form a summative judgment at those time when achievement is to be reported. The reason for this arises from the way in which judgements are made in formative assessment, which includes an element of student-referencing (ipsative). (If this were not the case, and formative assessment were to be purely criterion-referenced, the effect would be profoundly discouraging to many students who would be faced with regular failure.)

The student-referencing does not matter as long as the information is used to help each student. It does matter, of course, for summative reporting, where all students are assessed on the same basis. What is needed, then, is that the evidence used in formative assessment (and not the student-referenced judgments) is re-evaluated against the reporting criteria. Harlen and James suggest specifically that

(i) it is reviewed strictly against the criteria of what students are expected to achieve at certain ages/stages
(ii) the criteria are applied holistically, using judgments as to the ‘best fit’
(iii) there is some way of ensuring that the judgments of one teacher are comparable with those of other teachers.

(Harlen and James, 1997: 373)

The third of these points is relevant to all alternative summative assessment which depends upon teachers’ judgments. If a system based on teachers’ assessment is to achieve acceptable levels of consistency in the ways that judgements are made, then effective quality management procedures need to be in place. These will include moderation of the process of assessment to prevent error and bias (quality assurance) and moderation of the outcomes of assessment to detect any inconsistency in the way that criteria have been applied (quality control). Various approaches to quality assurance and control are described in Harlen (1994). These include statistical moderation (using a short external test), group moderation, visits of moderators, use of exemplars (or benchmark tasks), and accreditation of a school or individual. The
choice of method should be decided with a view to ‘quality enhancement’, a concept that recognizes the interrelationships of assessment with the curriculum and with pedagogy. Quality enhancement refers to the wider benefit of schemes for quality assurance and quality control, going beyond the dependability of assessment results and affecting the quality of students’ learning opportunities. For example, collaborative group moderation through discussion of students’ work supports teachers’ greater understanding of goals and of different ways of achieving them.

**Impact**

One of the important criticisms of high stakes tests is the impact on teachers’ use of assessment to help learning. Pollard at al (2000), in a longitudinal study of primary students in England found that the introduction of national tests was associated with teachers’ own classroom assessment becoming more summative. Using teachers’ judgments can, in theory, avoid this since the pressure to prepare for specific tests and constantly assess against summative criteria is absent. The evidence gathered and used as part of teaching can be the evidence used for summative assessment, providing that evidence is reviewed against the reporting criteria, as suggested above.

However, some of the negative consequences of tests (Harlen and Deakin Crick, 2003) can equally follow from using teachers’ judgements for summative assessment (or any other approach) if the results are used for accountability. In such circumstances, moderation procedures can become over-elaborate and constrain teachers to collecting evidence using ‘simple and safe’ methods rather than more educationally valuable ones. Then there is likely to be a tendency to focus on a narrow interpretation of criteria and on performance rather than learning. The possibility of reducing the high takes therefore has to be considered.

**Resources**

The amount of testing conducted by teachers would be expected to fall if teachers’ judgments were more widely used and trusted, with teachers and students then being able to spend a significant proportion of their time in other ways. More teachers’ time would be needed for moderation, of course, and as this is out-of-class time it requires resources for reducing contact time. Saving on examination fees and costs of purchasing tests would go some way to providing funds for this purpose.

**Can the stakes be lowered?**

Teaching will inevitably be focused on what is assessed. If this were not the case, then tests and examinations would not necessarily limit the curriculum and pedagogy. A disconnection between what is taught and what is assessed – even if it were feasible – would not, however, mean that tests were more able to assess the skills and higher level thinking that are the goals of modern education. However, as noted, alternatives to tests do not necessarily achieve their potential for assessing these goals in the face of high stakes use of the results.

To look to change this situation two sources of high stakes need to be to considered separately. One is the pressure on teachers arising from the direct link between students’ achievement and accountability. The other is the pressure on students because decisions about their future may hang on the results.
**Accountability**

Being accountable means being responsible for one’s actions and being able to explain to stakeholders why and how certain things were done or why they were not done. People can only be held accountable for actions or outcomes over which they have control. In the context of students’ learning, Madaus (1993) and Shepard (2000) are among many authors who have pointed out the unfairness of evaluating teachers and schools using the same tests and criteria when there are gross inequalities between schools in social, health, family and education resources and support. Teachers can be held accountable for what they do in the classroom, what learning opportunities they provide and the help they give to students, and so on. They are not necessarily responsible for whether externally prescribed learning outcomes are achieved, since this depends on many other factors, over which the teacher does not have control, such as the students’ prior learning and the many out of school influences and conditions that affect their learning (Madaus, 1993). These factors need to be taken into account both by teachers in setting and working towards their goals for students’ learning, and by those who hold teachers accountable for students’ achievements. (It is noteworthy that in France, tests provided by the Ministry of Education are administered and marked by teachers at the beginning of the third year of primary and the first year of lower secondary school so that they can be used diagnostically and teachers are not held responsible for the results (Bonnet, 1997)).

It follows that the information used in accountability should include, in addition to data on students’ achievements, information about the curriculum and teaching methods and relevant aspects of students’ backgrounds and of their learning histories. The student achievement information should reflect the full range of goals, both academic and non-academic. Since this information is not used to make decisions about or to report on individuals it need not be collected from all students in a class. Whilst it may be convenient to use existing data, such as from internal or external tests, it will be more meaningful to gather evidence of a wider range from a sample of students. This can be done, for instance, by the teacher or a visitor interviewing and reviewing the work of a few students so that the students’ understanding and conative as well as cognitive outcomes can be probed. Also, it may be more informative to report on the progress of groups of students, such as the lowest and highest achieving students, than on the average of the whole class.

Schools are now expected to communicate and explain their philosophy, aims, policies and student performance to the wider community. The process of doing this, as in the case of assessment of individual students, can be either formative - helping an institution to improve - or summative, providing a judgement that is used for decision-making. Also as in the case of individual assessment, there is greater value from the process when those being held accountable are involved in the process as opposed to following externally devised procedures. Making schools responsible for self-evaluation can provide them with the information they need in order to improve, and requiring them to give an account of themselves to others provides stakeholders with information to judge how effective the school is taking into account its circumstances. Various school self-evaluation guidelines provide some good examples of what this means (SEED, 2002; DfES and Ofsted, 2004; Estyn, 2004a and 2004b)).
High stakes for students

Self-assessment is also a way of reducing the impact of high stakes tests and examinations on individual students. In some systems the participation of students in their own summative assessment may seem unrealistic at present, but it ought to be considered as part of the move towards the development of students’ understanding of what it is to learn and of their learning autonomy that is a foundation for lifelong learning (Black et al, 2006). The argument begins with the role of student self-assessment in formative assessment. This is seen by Sadler (1989) as a key feature in the theory of formative assessment. Self-assessment depends on students understanding the goals of their learning, being able to see where they are in relation to these goals and recognising what they need to do to achievement them. Black and William (1998) review several research studies of self-assessment, all reporting success but also pointing out that implementation requires considerable changes in teachers’ roles and in assessment and pedagogical practices. So self-assessment is fundamental to formative assessment. What role can it have in summative assessment?

When summative assessment takes the form of tests or examinations in which self-assessment has no role, then the result is likely to bring awareness of differences which are interpreted by the lower achievers as failure. More likely, practice test have already made this clear and these students have written off any chance of success and ceased to make an effort. This learned helplessness ensures that these students’ prophesy of failure will be fulfilled. Although it is far easier to see how students can participate when summative assessment is carried out by teachers, there is still much that can be done in relation to tests and examination. A good deal of the emotion that is aroused by summative assessment results from fear or suspicion of the unknown. Students and teachers know full well that there is a lottery in the selection of items for any test, since it cannot encompass all that has been studied. Openness about the need for and purpose of assessment and about how and why it is carried out will help to take the fear and suspicion away from summative assessment. Those closely involved ought to be fully aware of how evidence is gathered and how judgments are made. Even the youngest students can be given some explanation of what evidence they and their teachers can use to judge whether they are making progress. Transparency of procedures supports students in assessing their own work. Action that teachers can take includes discussing with students the purposes of assessment, how it is carried out and what criteria are applied. Students can experience using these criteria in judging their work so that they have some idea of what results they will achieve.

All this helps to demystify summative assessment and ensure that the process provides feedback that motivates effort and at the same time means that there are no surprises in the outcomes (Dweck, 2000). A further step is to reduce the frequency of testing, so that it takes place only as necessary so that it is an occasional event, not a prominent part of their school work. Such actions will not remove the high stakes of assessment that is used for certification and selection, which after all is a strong motivation for learning, but they can improve the quality and enjoyment students’ learning experiences and their understanding of the learning process.
References


