



Qualifications and
Curriculum Authority

Considerations in the design of summative assessment systems which incorporate teacher-led assessment

Working draft

Paul Newton, Research and Statistics Team

22 December 2004

Ref: ASF 01 05

Paper presented to the core group of the *Assessment systems for the future* project. 11th and 12th January, 2005. Cambridge, England.

Introduction

Presently, in the UK, there is considerable interest in the prospect of giving teachers a greater role in the summative assessment of their own pupils, particularly in relation to end-of-phase attainment certification (which has traditionally been associated with external tests or exams).

In this climate, there has been a tendency for advocates to proclaim: “teacher assessment good – tests and exams bad”. This creed has proved useful for evangelistic purposes, but there are a number of respects in which it has become inadequate, particularly as policy makers begin to take it more seriously.

There are at least three tasks to undertake to overcome the inadequacy of the proclamation and to begin working towards effective assessment reform. First, we need to clarify what we mean by ‘good’ (in terms of quality and impacts of assessment). Second, we need to clarify what we mean by ‘teacher assessment’. Third, we need to explore how the quality and impacts of teacher assessment may be affected by the overall assessment system within which it is situated.

What do we mean by ‘good’?

There are two main senses in which teacher assessment might be said to hold more promise than tests/exams, i.e., in which teacher assessment might be said to be ‘good’:

1. in supporting **valid inferences and actions** based on assessment results
2. in facilitating **positive educational impacts** (and minimising negative educational impacts) associated with the overall assessment system

The first concerns the narrow goal of assessment, while the second concerns the broader goals of assessment and, in particular, the broader goals of assessment reform.

It would be important to distinguish between these two senses if, for example, teacher assessment proved to be good in terms of educational impacts but bad in terms of the assessment goal (or *vice versa*). Of course, the situation is inevitably more complex than this; and decisions between assessment models typically require a trade-off between the accomplishment of positive educational impacts and the accuracy of assessment results. This trade-off is further influenced by resource constraints which establish parameters for assessment reform.

The narrow goal of assessment

The fundamental objective of any assessment system is to deliver results which support valid inferences and actions, where validity is understood in relation to a specific assessment **purpose**, for example:

- placement (e.g., decisions on which pupils to assign to which sets)

- certification (e.g., inferences concerning what pupils have learned by the end of an instructional course)
- selection (e.g., decisions on which students to offer further educational opportunities to)
- accountability (e.g., inferences concerning which teachers are under-performing)

Two major threats to valid inference and action are construct under-representation (i.e., not assessing the full range of characteristics associated with the to-be-assessed construct) and construct irrelevant variance (i.e., assessing characteristics not associated with the to-be-assessed construct). Evidence concerning threats to valid inference and action can be categorised using the traditional technical lexicon of educational assessment, including:

- content validity evidence
- predictive validity evidence
- reliability evidence
- comparability evidence
- etc.

The specific purpose, or purposes, to which assessment results are to be put affect the definition of **quality criteria** in relation to these categories. In principle, these quality criteria define out 'how much' content validity, predictive validity, reliability, comparability, etc. is 'enough' for the purpose at hand. Unfortunately, quality criteria cannot straightforwardly be specified (e.g., x% reliability, y% validity, z% comparability). This is for a variety of related reasons, including:

- the technical categories are not conceptually discrete (each one essentially characterises a different facet of construct validity)
- even though certain of them can be quantified (e.g., reliability), no overall index of technical adequacy can be computed
- although, in principle, it would be possible to specify minimum acceptable thresholds for certain of the technical constructs (e.g., reliability), the basis for making such decisions is not at all clear

The key variable in specifying quality criteria is the consequence of invalid inference/action. Consequences follow not simply for those who are assessed (e.g., pupils, schools) but for all stakeholders (e.g., employers, society generally). The key variable is generally described in terms of the '**stakes**' of the assessment.

Generally speaking, high stakes require more stringent quality criteria than low stakes. In short, the more important it is not to make mistakes, the greater the level of technical accuracy which the assessment results need to exhibit.

However, to complicate the matter, different purposes are also likely to favour different kinds of assessment error. For example, where the principal purpose is teacher accountability, comparability error – related to different standards being applied in different schools – would be serious (much more so than reliability error at the pupil level). In contrast, where the principal purpose was within-school placement of students, between-school comparability error would not be serious (although reliability error and validity error would be).

The broader goals of assessment reform

Assessment reform might be motivated by a perception that the extant system was not delivering sufficiently accurate results (for its intended purpose or purposes). However, it could equally be motivated by a desire to achieve particular educational impacts. Indeed, sometimes assessment reform will involve adopting a system which delivers *less* accurate results, but which delivers clear educational benefits.

Impacts might be classified at a variety of levels, including those on pupils, teachers, managers, parents, etc.. The following list illustrates possible positive educational impacts from assessment reform, given the introduction of an assessment model which:

- enabled *pupils* to learn better (e.g., by facilitating self-assessment practices, or by communicating learning objectives and assessment criteria more effectively)
- motivated *pupils* to achieve more, or simply to continue in education (e.g., by assessing smaller chunks of the course, or by making their progress more evident throughout the course)
- ensured that *pupils* learn what they need to (e.g., by minimising or avoiding opportunities for cheating/playing the system)

- enabled *teachers* to teach better (e.g., by facilitating a deeper understanding of learning objectives and assessment criteria, or by making students' strengths and weaknesses more evident throughout the course)
- motivated *teachers* to achieve more (e.g., by promoting a stronger sense of professional responsibility, or by allowing more control over how and what to teach)
- ensured that *teachers* teach what they need to (e.g., by ensuring that all aspects of the curriculum are assessed, or by minimising or avoiding inappropriate assessment-preparation strategies)

- enabled *managers* to manage better (e.g., by providing them with more, or more useful, information on the strengths and weaknesses of their pupils and teaching staff)

- motivated *parents* to engage more with their children's learning (e.g., by providing regular updates on progress)

Again, it is important to stress that these anticipated educational impacts are independent of the assessment goal; that is, there is no *a priori* reason to assume that an assessment system designed specifically to achieve these impacts would also deliver high quality assessment results. Realistically, though, the system would need to deliver at least reasonable technical accuracy, else negative consequences from using inaccurate assessment results would probably come to outweigh the broader positive educational impacts. The holy grail, of course, would be an assessment system which delivered both highly accurate results and a wealth of educational benefits.

Resource constraints

The trade-off between the narrow goal and the broader goals of assessment is not simply two-way, but three-way, since it will be mediated by resource constraints. Even if it was possible, in theory, to develop an assessment system which delivered both highly accurate results and a wealth of educational benefits, it might still not be possible in practice. The resource constraints set **parameters** for assessment reform and bear directly upon decisions concerning what kinds of systems would, and would not, be feasible. The nature of the trade-off between assessment goals, assessment reform goals and resource constraints can be illustrated through the oft-quoted plea from a high-ranking NASA official: “faster, better, cheaper... pick any two!”

Resource constraints fall into a number of categories, and some of the key ones are illustrated below:

- **financial** constraints – will the proposed system cost too much?
 - assessment instrument costs (e.g., exam papers, postage)
 - invigilation costs
 - script marking, or moderation, fees
 - training costs (e.g., travel, accommodation, teacher-release), etc.

- **time** constraints – will the proposed system deliver results quickly enough?
 - time spent completing assessment exercises (either in class, or during an examination)
 - time spent marking assessment evidence
 - time spent processing assessment data
 - time spent piloting the system, etc.

- **physical** constraints – will the proposed system be manageable?
 - ability to assess all students simultaneously (e.g., exam room space, or PC availability)
 - ability of markers or teachers to travel to training or moderation meetings, etc.

- **expertise** constraints – will there be sufficient assessment expertise to support the proposed system?
 - expertise in test design, pre-testing, standard setting
 - expertise to run training meetings
 - expertise to evaluate the system, etc.

- **social** constraints – will the proposed system be accepted?
 - stakeholder acceptance of the assessment model (e.g., perception of technical accuracy of results, perception of potential to cheat or play the system)
 - participant acceptance of new requirements (e.g., perception of workload, perception of validity of assessment model)
 - stakeholder and participant understanding of the system (i.e., whether the system is sufficiently open and transparent and whether the inferences which follow from results are clear), etc.

Occasionally, when parameters appear to have been exceeded within extant systems, resource constraints may feature as specific goals of assessment reform

(e.g., when confidence in the system has been lost, or when the assessment workload has become too high). However, resource constraints generally do not feature as specific goals, merely as pragmatic parameters.

Synthesising narrow and broad goals

The **defensibility** of an assessment system is probably best understood in terms of a trade-off between the narrow assessment goal and the broader goals of assessment. This would be to leave resource constraints entirely out of the equation which might, perhaps, be contentious. However, an assessment system which delivered inaccurate results, with negative educational impacts, could hardly be said to be defensible simply because it was the only one which could satisfy highly restrictive financial constraints.

To begin the process of system design for assessment reform, we need to:

1. define quality criteria for the narrow assessment goal
2. specify the intended positive educational impacts of assessment
3. quantify our resource constraints

During each of these tasks, and at all stages of decision-making process, underlying aims need to be **prioritised** or **weighted**. This might be done explicitly, although the complexity of the decisions to be made might invite a more implicit approach. Whether prioritisation occurs implicitly or explicitly, the decision-making process cannot be completed successfully in the absence of prioritisation. Prioritisation allows us to determine which losses are most tolerable, assuming that the holy grail of no losses is unattainable.

So, in defining quality criteria for the assessment goal, we first need to prioritise between assessment purposes (e.g., placement, certification, selection, accountability). It is tempting to argue that all should be given equal weight, implying that all assessment purposes should be supported to the same extent. The reality, though, is that any particular assessment system will be more effective at supporting certain purposes and less effective at supporting others. *So we do need to prioritise!* The prioritisation supports the definition of quality criteria by suggesting which types of error are most important to avoid (i.e., comparability, validity or reliability). Once this has been done, different assessment systems can be evaluated in terms of the likelihood of satisfying these criteria.

The specification of intended educational impacts requires a similar kind of prioritisation, for a similar reason: it probably will not be possible to achieve all desired goals through a single assessment system. Here, we need to consider each goal, not simply in terms of the likelihood of a positive impact, but also in terms of the likelihood of a negative impact. An assessment system which facilitated moderate positive impact on the highest-weight goal (e.g., student motivation), but which resulted in substantial negative impact on the second-highest-weight goal (e.g., teacher motivation), would not be particularly desirable. Again, once this prioritisation has been completed, different assessment systems can be evaluated in terms of likelihood of achieving the goals.

The quantification of resource constraints comes into play when the different assessment systems are being designed. Systems which cannot satisfy the resource constraints cannot be considered.

The final choice of assessment system is unlikely to be straightforward, since the system which best supports the narrow assessment goal may well differ substantially from the system which best supports the broader educational goals of assessment. As such, the decision-making process is likely to require some kind of iteration, whereby each of the prior tasks is revisited. This iteration might even necessitate a radical re-evaluation of central aspirations, if an adequate assessment system cannot be devised to satisfy all aims within resource constraints. (Alternatively, the resource constraints might be revisited, e.g., to allow more money or time.) Most likely, the iterative process will involve some lowering of expectations for the desired educational impacts, some reduction in ambitions for the number of assessment purposes and/or a lowering of quality criteria for the narrow assessment goal.

Clearly, the more assessment purposes and the more desired educational impacts, the harder the decision-making task will be, the more iterations will be needed, and the less satisfactory the final assessment system is likely to be. *Modest ambitions are always to be recommended when undertaking assessment reform!*

What do we mean by ‘teacher assessment’?

For the purpose of the following discussion I will draw a distinction between an assessment model and an assessment system, where the model is a discrete component of the system, tending to describe the key features of the assessment process. Crucially, the quality and impacts of assessment will be determined by the characteristics of the system (rather than by the characteristics of the model). So, although assessment quality and assessment impacts *can* be theorised at the level of an assessment model (see this section), they cannot fully be determined until situated within an overall assessment system (see next section).

This section deals with different kinds of teacher assessment models, noting that different quality and impact implications will follow. It provides only a very rough sketch, since the point is simply to indicate that the term ‘teacher assessment’ is not particularly useful until unpacked.

Perhaps the easiest way to conceptualise differences between teacher assessment models is in terms of **resource-intensiveness**; principally, the resources of time, money and effort. It is possible to envisage a rough continuum along which teacher assessment models differ in terms of how resource-intensive they are. Obviously, this is not the only way of differentiating between models, and it is certainly only of heuristic value, but it helps to illustrate a point.

At the lower end of the continuum might be a model in which very few resources were committed to the generation of summative assessment results. For example, a teacher assessment model in which:

- assessment evidence was generated opportunistically, from work conducted in class or at home, but no formal portfolio was constructed
- no external rules, or guidelines, were specified for how judgements should be reached on the basis of evidence
- judgements of performance were made holistically and impressionistically at the end of the period of instruction
- judgements were made according to a common grading scale, but without good exemplification of standards

- no training or guidance was provided to teachers in how to make their judgements
- no moderation of assessment judgements, or verification of assessment procedures, was undertaken
- students were not allowed to appeal against their results

At the upper end of the continuum might be a model in which many resources were committed. For example, a teacher assessment model in which:

- assessment evidence was generated from work conducted in class, according to a pre-specified assessment plan determined by a central assessment agency, and assembled by students into a formal portfolio containing every piece of assessment evidence generated (annotated by the teacher to indicate level of support received)
- rules and guidelines, specifying exactly how judgements should be reached on the basis of evidence, were determined by the agency
- judgements of performance were made according to an assessment matrix which specified both criteria and weighting, and detailed mark- and comment-based records were kept for every piece of assessment evidence generated
- overall judgements of 'mastery' were made according to a common grading scale, using explicit aggregation rules, and with detailed description and exemplification of standards provided by the agency
- high quality training and guidance documents were provided for teachers by the agency, to support teachers in making accurate judgements
- school and LEA consensus meetings were regularly held for training purposes
- moderation of school assessment judgements, and verification of school assessment procedures, was undertaken by a visiting agency representative
- students could appeal against their results to the agency

Clearly, the low-resource-intensive teacher assessment model would be quite practical, i.e., fast, cheap and easy. Furthermore, it might well be able to facilitate certain desired educational impacts; for example, it would leave plenty of time for teachers to hone their formative assessment practices. However, under this model, it would not be possible to provide any guarantee that teacher judgements were valid, reliable and/or comparable.

On the other hand, the high-resource-intensive model might well be able to deliver very valid, reliable and comparable assessment judgements. However, it might be so demanding of time, money and effort as to make it unworkable. If it were to be rolled out, it might result in negative educational impacts; for example, summative assessment might come to dominate the teaching and learning process unduly.

At neither of these extremes could teacher assessment necessarily be assumed to be 'good'. Good teacher assessment, for most stakeholders, would probably lie somewhere in between. Again, though, there are very many different manifestations between these extremes, each with their own particular pros and cons. And recall that the pros and cons will be relative to the context of intended purposes and impacts: a low-resource-intensive model might be best for supporting low stakes placement decisions and promoting assessment for learning practices; while a high-resource-intensive model might be best for supporting high stakes selection decisions and providing performance data for managers.

In short, contra the introductory proclamation, teacher assessment is not necessarily good; and certain teacher assessment models, in certain contexts, will be bad.

How might the wider assessment system affect the quality and impacts of teacher assessment?

It is important to recognise that the intended quality and impacts of a teacher assessment model are not necessary features, but will be affected by a variety of contextual factors. These can be explored at two levels: first, where the system is based upon a pure teacher assessment model; second, where the system is based upon a hybrid test/exam and teacher assessment model.

The following sections illustrate further considerations in establishing the likely defensibility of an assessment system.

Pure teacher assessment model

Perhaps the most important factor affecting whether a teacher assessment model will achieve the intended quality and impacts is whether participants are motivated to make it work as intended. Here, the principal participants are pupils and teachers.

Threats to participant buy-in are likely to be significant when, for example, a new model is perceived to require a heavier workload than the previous one, or when it is felt to be less empowering, or when there are clear opportunities for playing the system.

A particular threat would arise if teacher assessment alone was intended to support a school accountability framework. It would create perverse incentives for teachers to inflate their judgements; or, at least, to be generous in giving the benefit of the doubt. Although it is often argued that we just need to have more trust in teachers, even now many teachers openly admit to an undue focus on test preparation strategies or to an undue focus on those areas of the curriculum which are most heavily tested. Given this experience, it might not be appropriate to trust teachers with sole responsibility for assessing their own students if, in doing so, they would also be incriminating themselves.

A similar threat would arise if the new model of teacher assessment gave rise to more opportunity for cheating: be that on behalf of the teacher (e.g., giving undue assistance to students in producing assessment evidence); or on behalf of the pupil (e.g., downloading assessment evidence from the internet).

For both students and teachers, a perception of undue assessment workload would also pose a threat to successful implementation. Teacher assessment models which incorporate portfolios can appear to raise workload issues, but it is probably fair to say that the perception will depend upon: the kind of portfolio model (e.g., 'all evidence' versus 'latest and best'); its centrality to the teaching and learning process (e.g., compiled as part of normal teaching and learning versus compiled at the end of a year); and/or whether lines of student/teacher responsibility are managed effectively. Perceptions of undue assessment workload can also arise in relation to models which, ostensibly, make fewer demands; in particular, where the rationale for the workload is not fully appreciated or accepted (e.g., where teachers are required to complete numerous 'tick lists').

Finally, a factor which might mitigate the negative motivational impact of a high assessment workload is the degree of teacher empowerment within the model; for example, locating full responsibility for assessment and/or pedagogy with the school/teacher rather than with a central agency. Where the locus of professional responsibility lies primarily with a central agency (e.g., where the agency ultimately awards the result rather than the school, or where teaching methods are heavily prescribed) the process of assessment can be seen as an added burden rather than a professional responsibility.

Hybrid test/exam and teacher assessment model

It is seductively easy to assume that, as long as teacher assessment is *part* of the system, then the benefits of teacher assessment will accrue. This is undoubtedly not the case. In fact, hybrid systems which attempt to integrate both test and teacher assessment are particularly at risk of extracting the worst from both worlds. This is especially so because hybrid systems require the sub-division of limited resources. Running two very different systems in parallel can be very costly on all fronts, but most obviously in financial terms. The idea of 50% standardised constructed-response test and 50% internally and externally moderated teacher portfolio assessment might prove attractive to some, but it might well also cost a lot more money to operate effectively than either of the models in isolation.

There are other ways in which the anticipated quality and impacts of either a pure test/exam model, or a pure teacher assessment model, might be corrupted by combining them. The most obvious is the likelihood of increasing the assessment workload where, for example, teachers are both preparing students for tests/exams and ensuring that they generate a solid evidence base for sound teacher assessment judgements.

There are also many possible examples through which to illustrate the fact that just adding a component of teacher assessment will not solve problems caused by testing. For example, where a high stakes test is focused only upon a limited number of 'coachable' proficiencies, the addition of an element of teacher assessment is not going to prevent the teaching of inappropriate test-preparation strategies, nor will it necessarily prevent teaching-to-the-test. This would be particularly true if the results of test and teacher assessment components were reported separately, with the test results used for high stakes purposes and the teacher assessment results used only for low stakes purposes.

Conclusion

It is very easy to get sucked into debating along the lines: "teacher assessment good – tests and exams bad". Yet, not only is this far too simplistic, but it provides no support for policy makers who actually *do* want to strengthen the role of teacher assessment. It offers no distinction between good teacher assessment models and bad ones, and provides no indication that even good teacher assessment models can become bad when located within an ineffective assessment system. In fact, only once a full, contextualised system has been posited can assessment quality and impacts effectively be theorised. This is the level of debate which needs now to occur in the UK.