Designing and implementing a teacher-based assessment system: Where is the infrastructure?

Richard Daugherty
Oxford University Centre for Educational Assessment

Paper presented at the seminar
Teachers' judgments within systems of summative assessment: strategies for enhancing consistency

St. Anne's College, Oxford
20th to 22nd June 2011

Designing and implementing a teacher-based assessment system: Where is the infrastructure?

Introduction

For teachers to judge the attainments of their own students has been widely acknowledged as offering a potentially more dependable basis for summative assessments of student performance. Teachers' ability to draw on evidence from numerous events within a student's learning programme should, at least in principle, ensure that teacher judgments of such evidence will more accurately reflect student performance than would evidence from external tests and examinations. The case for teacher judgment to contribute to summative assessments has been succinctly made as:

> …teachers can sample the range of a pupil's work more fully than can any assessment instruments by an agency external to the school. This enhances both reliability (because it provides more evidence than is available through externally devised assessment instruments) and validity (it provides a wider range of evidence).
> (Mansell, James, et.al., 2009)

This argument has fuelled the case for performance assessments in the USA (Resnick & Resnick, 1992), for assessment by teachers in high school qualifications in the UK (Wilmut, 2004), and for reliance on teacher judgment in school-based assessment at the end of secondary school in Queensland, Australia (Maxwell, 2010). Harlen (2005) has reviewed the research evidence on the reliability of teachers' assessments for summative purposes, leading her to conclude that:

> When steps are taken to moderate the results, the reliability of teachers' judgments is comparable to that of tests.
> (Harlen, 2011)

What form should a system in which teacher judgment has a role take? In summative assessment systems that rely on evidence from tests and examinations much attention is focussed on test items, the responses to which will be the basis for conclusions about student performance. In systems where teacher judgment has a role the main focus of attention has usually been on how such judgments can be consistent when they are arrived at by a large number of teachers using diverse sources of evidence obtained in circumstances that are less constrained than test papers taken under examination conditions.

And yet both the student responses to tasks set during a programme and their responses to items in a time-limited test are in fact key points within an assessment *process* that begins with the design of a course programme and ends, for summative purposes, with the reporting of an overall measure of performance for each student following the programme. To focus only on the admittedly key point in the process where evidence is elicited and initial judgments are made is to overlook the fact such a process, whether teacher-based or test-based, needs to be structured in such a way as to maximise validity and reliability.

The *structure* of a test-based process is not always open to public view but anyone familiar with the work of testing agencies will be aware that there is a series of critical stages in that process, first to develop a set of appropriate test items and then to ensure that student responses are judged in ways that control for variations in assessor judgment. What appears to be less widely recognised is that if a system based on teacher judgment is to be credible it will require equivalent structures.

This paper will start by attempting first to identify the stages that need to be built in to a system of teacher-based summative assessment. It will then move on to explain and discuss how two such systems are being developed in Wales to assess students at the ages of 11 and 14. It will conclude by arguing that unless systems of teacher-based assessment are designed, implemented and supported by a suitable *infrastructure* they are unlikely to command the confidence of students, teachers or the wider public.

Stages in a process

As a starting point for discussing the process involved in teacher-based summative assessments I will refer to a small-scale piece of research from 25 years ago. A UK awarding body, then known as an 'exam board', allowed me access as a researcher to the decisions taken by teachers in schools and examiners employed by the board in two of the geography '16+ examination' schemes that preceded the introduction of the General Certificate of Secondary Education (GCSE) in England, Wales and Northern Ireland.

Looking through the literature at the time I couldn't find any account, still less a discussion, of the decision-making procedures on which awards at General Certificate of Education (GCE) Ordinary ('O') and Advanced ('A') levels were based. As a former chief examiner at GCE A level I had personal experience of how 'A' level papers were set and marked and of the subsequent steps, such as 'borderlining', that lay between the marking of scripts and the awarding of grades. But nowhere in the literature was there a description of the process.

Both the schemes in my research depended in part on test papers and in part on teacher judgment of what was called 'coursework'. The first aim of the study was to set out in diagrammatic form (fig. 1) how the exam papers were set and how student responses to the exam questions were translated into overall grades. The left-hand side of the diagram shows who was doing what at each stage and also which of those stages I was able to observe directly.

On the other side of the diagram is my attempt to do the same for what was then a relative novelty for the higher status GCE awards at 16+ though well established in the parallel Certificate of Secondary Education (CSE) examinations which had been originally targeted at lower attaining students. The part of figure 1 that is distinctive to coursework assessment shows the details of the process up to the point where the marks allocated for coursework were combined with the marks allocated for performance on the exam papers. It is this primitive piece of modelling on which I now want to elaborate in the light of subsequent experience of what in the UK is usually referred to as 'teacher assessment' in public examinations and in National Curriculum assessment.

Before I do so there is one other conclusion I came to from studying the 16+ geography examinations. My task there as a researcher was to analyse the decision-making process and the report of the research (Daugherty, 1986) is couched in those terms. However, the experience also opened my eyes to the potential strengths and weaknesses in such systems. I knew from talking informally to teachers that they didn't trust the coursework assessment outcomes for their students on one of the schemes. In contrast, the teachers with candidates on the second scheme were fully involved in the process and had confidence in the outcomes. Unfortunately most teacher-based summative systems in England and Wales, including GCSE, in the years since that study have had more in common with the first of those schemes than the second. And, I would argue, that is more a matter of failure to design and implement effective structures than of failure on the part of the teachers involved in those schemes.

Links in a chain?

A diagram (fig 2) from a seminal paper by Crooks and colleagues on validity (Crooks et al, 1996) uses the metaphor of links in a chain to discuss 'threats to validity'. In essence the argument is that a well-designed system of assessment should make provision for every link. Validity will be compromised if any one link in the chain is weak. Can the same metaphor be applied to teacher-based summative assessment and, if it can, what are the essential links in that chain for reliability and validity to be maximised?

I have proposed elsewhere (Daugherty, 2010) that there are a number of critical stages, each of which needs to be clear, in a structured process of moving from an already defined programme via the judgments teachers make to the drawing of inferences. Those stages in the process call for some specification of:

1 task type;
2 task conditions;
3 criteria against which student performance is to be judged;
4 performance standards.

But those stages are not the only features of a system that are important. Decisions at each stage are conditioned by the circumstances in which they are taken so it is not enough simply to plan each stage, for example by finding a balance between over-specification and teacher freedom in the regulation of task type. I have therefore suggested another four aspects of the framework for teacher judgment, the infrastructural factors, which also need to be addressed. These are:

1 explicit procedures for each stage;
2 existing teacher expertise;
3 ongoing teacher training and support;
4 quality assurance and control arrangements.

Taking this argument one step further, is it helpful to depict all these aspects of structure and infrastructure as links in a chain (fig. 3)? In other words, if any one link is missing or poorly designed or inadequately implemented the whole system of teacher-based judgment will be compromised. If so, how closely related are those links to the 'threats to reliability' identified in what the recent report (Baird, et.al., 2011) of the Technical Advisory Group of the Ofqual Reliability Programme refers to as 'teacher assessment'?

The Ofqual report calls for (p.49) a 'more complex account' of the variety of threats to the reliability of teacher assessment and identifies seven potential sources of unreliability:

1 design of the tasks;
2 presentation to pupils;
3 carrying out the task;
4 the product to be assessed;
5 assessment of the product;
6 procedures to reduce bias;
7 strategies to limit plagiarism.

It is understandable, given the remit of the Ofqual Programme, that these threats in the TAG report refer mainly, apart from the last two, to aspects of process and structure. However, it is arguable that some of the greatest threats to reliability are not from structural weaknesses but rather from deficiencies in the associated infrastructure, for example inadequacies in teacher support and/or in provision for quality assurance and control. So, for an overview of the necessary conditions for successful implementation of teacher-based systems, I would argue for including the infrastructural factors in my own version (fig. 4) of links in the chain of threats to the reliability of teacher-based summative assessment systems.

It should also be noted that there is a spatial dimension to be taken account of, in addition to this temporal dimension of stages in a process, when designing a large-scale system that relies on teacher judgment. Figure 4 illustrates the kind of layered structures that are required when the judgments made by numerous teachers are brought together within such a system. There has been a more general recognition of the need for these layered structures in the vocational education sector in the UK than has been the case in either higher education or the school sector.

National Curriculum key stage assessments in Wales

Until decisions on education policy were, with the establishment of a National Assembly for Wales in 1999, devolved to Wales that country had a minor role in a policy process that was 'England-based and London-centred' (Fitz, 2000). The new Welsh Government[i] set out its goals for education in *The Learning Country* (NAfW, 2001). Then, in the 2002 Education Act, provision was made for decisions on the school curriculum and on assessment to be made by the Welsh Government within the framework that had been established for both England and Wales in the 1988 Education Reform Act.

Neither the school performance tables that had become so prominent as an instrument of policy in England nor the tests for 7 year-olds at the end of 'Key Stage 1' had ever

received much professional or public support in Wales. The Minister responsible for Education in the National Assembly for Wales, Jane Davidson, soon decided to dispense both with school performance tables for all age groups and with tests at the end of Key Stage 1 in, respectively, 2001 and 2002. The fact that there was virtually no opposition to those changes presumably encouraged her to consider potentially more problematic, and more controversial, changes to the statutory assessment framework at Key Stages 2 and 3.

In 2004, the reports of two parallel reviews, one by the Government's own advisory body (ACCAC, 2004) and the other by an independent *ad hoc* group (Daugherty et.al., 2004), made similar recommendations that would form the basis for changes to assessment policies for schools in Wales. There are four main strands to those changes:

1 'Assessment for learning' as a central element in curriculum and assessment across all key stages;
2 'Skills profiles' for every pupil to be reported and developed from Year 5 onwards.
3 Assessment by teachers in the four core subjects at the end of Key Stage 2, backed by a system of cluster group moderation integrated with arrangements for pupil transition from primary to secondary school at age 11.
4 Assessment by teachers in all eleven National Curriculum subjects at the end of Key Stage 3, backed by a national system of secondary school accreditation.

Policies on all four strands have evolved since 2004 in ways that are summarised by James (2011). It is the third and fourth strands, policies for summative assessment by teachers at the end of Key Stages 2 and 3, on which this paper will focus.

At Key Stage 3, the reviews had recommended that secondary schools should submit sample evidence of students' work which would then be scrutinised through a process of moderation and verification leading in time to all schools being accredited to carry out these assessments. What has actually happened is that there have been differences each year in the aspects of external scrutiny that have been undertaken. For the first two years (2006-8) schools submitted portfolios, for moderation and feedback, of evidence of student attainment in the four core subjects. For the next three school years (2007-10) a similar portfolio-moderation-feedback process was applied to evidence in the nine non-core subjects. Then, during 2008/9 and 2009/10, every school's internal assessment systems and procedures were investigated and reported upon by a team of external verifiers. Thus, over the years since teacher judgment replaced external tests as the source of evidence on student attainment across all National Curriculum subjects, the building blocks have been put in place for what could be, should the Minister so decide, a sustainable system of teacher-based summative assessment at the end of Key Stage 3.

The different educational context for students at age 11 led the 2004 reviews to recommend a different model of summative assessment for the end of Key Stage 2. Students at age 14 in Wales will be moving on to make subject choices for Key Stage 4 having followed a broad National Curriculum up to that point but students at age 11 will be changing schools, transferring from primary to secondary school. All students in state schools transfer at age 11 and all state secondary schools are non-selective. The main policy priority for the Minister in 2004 was not 'how best to use aggregated Key Stage 2 assessments of students to judge the performance of schools?' but rather 'how can we

assess each student's attainments in her/his final year at primary school in ways that facilitates progress as s/he moves on to secondary school?'.

The system of cluster group moderation that has been introduced since 2005 reflects this overall aim for Key Stage 2 assessment in Wales. Each secondary school was linked to several feeder primary schools with a view to developing, through standardising and moderation at cluster group meetings, a shared understanding in each cluster of standards across the gulf that often seems to divide primary from secondary school teachers. Establishing those procedures was managed at the local, rather than the national, level although the Welsh Government did publish guidance materials and made provision for two additional in-service days specifically for this purpose.

Unsurprisingly the school inspectorate in Wales (Estyn, 2010) has reported that implementation of this system is patchy. The response of most schools has been positive but the goal of shared standards remains elusive:

> In three quarters of primary schools surveyed the confidence of teachers to award teacher assessment levels in line with National Curriculum level descriptions has improved through the cluster based process.
> …..schools state that it is extremely difficult to ensure that the cluster's shared understanding of standards is actually applied in the assessments that teachers subsequently undertake.

More surprising is the fact that the inspectorate's report makes no reference to the nature or extent of any provision for the training of teachers to prepare them for the central role they have in the system. This is in spite of the fact that 'professional learning' (Gardner, et. al. 2010), both generic and system-specific, is generally recognised as critical to successful innovation in teacher-based systems.

The system for cluster group moderation continues to evolve. A role for external moderation was signaled by a pilot project in 2009/10 involving a nominated Key Stage 2/3 cluster from each of the 22 local authorities (LAs). That is to be followed up by a roll-out of external moderation across all the clusters in 2011/12. A significant adjustment to the policy on cluster groups came with the announcement from the Education Minister, in the course of a wide-ranging speech in February 2011, that:

> We will expect all local authorities to ensure that Key Stage 2 teacher assessments are robust and consistent with the nationally defined standards, especially in respect of literacy.
> (Andrews, 2011)

This would seem to mark a shift from the original system of 'locally-owned' assessments of students at age 11, with teachers taking responsibility for defining and sharing standards within clusters, towards a system where LA staff are increasingly prominent in seeking to ensure that such assessments are 'consistent with nationally defined standards'.

Discussion

This case study throws some light on the problems encountered when implementing a system of summative assessment that relies on the judgments teachers make on the

attainments of their own students. Those problems are inevitably exacerbated by what is required of all stakeholders – parents, local authority staff and politicians as well as students and their teachers – in adjusting to a lower stakes teacher-based system replacing a higher stakes test-based system. A number of questions arise relating to the feasibility, fitness for purpose and sustainability of teacher-based summative assessment systems. (Reference is made here mainly to the Key Stage 2 cluster group system although the similar questions could equally be addressed to the Key Stage 3 portfolio-based system.)

Is it *feasible* to rely on the judgments made independently of each other by thousands of teachers across hundreds in school in 22 LAs supported only by briefing documents prepared at national level? The inspectorate's report on the cluster group arrangements reported that only about half of the schools surveyed were using the full two in-service days allocated for this purpose. Moreover, the great majority were using the time on standardising activities rather than on the other part of their remit, to moderate the judgments made by each year 6 teacher within the cluster about the attainments of her/his students.

Unease about the consequences of judgments at age 11 being made in each school and moderated only locally lay behind the Minister's recent decision to give greater emphasis to the role of LA staff in helping develop consistency of teacher judgment across the clusters. In the speech referred to above the Minister also highlighted the poor performance of students in Wales on 'reading literacy' in the most recent PISA survey (Bradshaw, J. et. al., 2010) as a cause for concern and announced his intention to ensure that end-of-stage teacher judgments were not the only source of evidence on attainment in reading:

> We will introduce a national reading test which will be consistent across Wales and will be designed to ensure that far fewer pupils are falling behind their designated reading age. (Andrews, 2011)

The arrangements for introducing reading tests and for any use of the results for any purpose other than the diagnosing of individual students' reading problems are unclear at the time of writing. For some teacher organisations the tests are being seen as a return to the 'performativity' culture to which was Wales was subjected in the pre-devolution era. For other interested observers of the system the use of diagnostic testing alongside teacher-based summative assessments is a considered response to the deep-seated weaknesses in students' reading.

It should be noted in relation to feasibility that this recent policy focus on developing moderation arrangements within and across clusters only addresses one aspect of the system's infrastructure (see fig. 3). The low priority currently given to other aspects, for example ongoing training and support for teachers, will continue to pose a threat to the system's dependability.

A second question, *fitness for purpose*, has been less widely discussed to date even though it will be critical to the long-term survival of a teacher-based system for reporting attainments. As was referred to above, the main reason in the minds of those who framed the policies that led to changes in assessment practices at the end of Key Stage 2 was to facilitate the educational transition of individual students from primary to

secondary school. Whether the arrangements that are now in place are well designed for that purpose has been the main focus of this paper. But in the medium- to long-term all those concerned with the system, especially the teachers called upon to operate it, will be asking: 'Does it contribute to a smoother transition in educational terms?' If there are not tangible benefits that are widely acknowledged by both primary and secondary school teachers the system will be in danger of degenerating into another set of bureaucratic procedures to which schools and teachers reluctantly conform.

'Mission creep' is a potential threat to the system's fitness for purpose. It has already been evident in some areas of Wales that LA staff who were accustomed to using aggregated end-of-stage test results as an indicator of school performance did not adjust readily to being told that the new teacher-based system was not designed for that purpose but rather to help individual students progress in their learning. The Key Stage 2/3 cluster group arrangements were designed to meet the purpose of smoothing the educational transition of students at age 11. That purpose will need to be kept to the fore in the minds of all concerned if the system which emerges is to be seen to be fit for purpose.

A third question to be considered is will the systems be *sustainable.* In their review of innovations in assessment in the UK, Gardner and colleagues concluded that:

> Successful innovation in assessment involves sustaining a climate of development where policy-makers, academic researchers, schools and teachers seek collectively to improve learning. More importantly, it involves teachers and schools in a culture of reflection and review….
> (Gardner, et.al., 2011, p.112)

Five years after implementation of the two systems in Wales began there would seem to be two main threats to the sustainability of the system. First, it is doubtful whether each of the elements identified in figure 3 above, especially the infrastructural elements, has been sufficiently considered for there not to be several potential weak links in the chain. For example, how much attention is being paid by policy-makers to the need either for explicit procedures to guide the activities of cluster group meetings or for quality control arrangements to confirm that adequate procedures are in place in every cluster? More fundamentally, has any account been taken of the expertise teachers bring to the operation of the system? Research in one LA in England by Black and colleagues (2011) investigating assessment arrangements within schools has revealed the variable levels of expertise that teachers bring to those activities.

The second threat to sustainability lies in what would appear to be the prevailing mindset of policy-makers in the UK when setting up teacher-based systems of summative assessment. It is difficult to imagine in a test-based system that policy-makers would decide to invest resources in different aspects of the system each year; they would surely see it as essential for the system, perhaps following a pilot phase, to be fully effective in all respects for each student cohort if the outcomes are to be credible. And yet piecemeal development (see above) appears to be what has been happening in Wales since implementation of the recommendations of the two reviews began in 2005. Is it perhaps that a system of teacher-based summative assessment is not seen as needing to be fully effective from the outset and is subjected instead to a series of policy initiatives from year

to year?

If the teacher-based systems of summative assessment in Wales are to be sustainable ongoing and effective structures will need to be in place to support teacher judgment. There will also need to be, as Gardner and colleagues have argued, a climate of development where policy-makers, academic researchers, schools and teachers continue to seek collectively to improve learning.

Conclusion

I have argued that the judgments made by teachers within large scale systems of summative assessment should be seen as only one amongst a number of stages in a *process.* Teachers' judgments in such a system need to be supported by a *structure* if all the relevant stakeholders are to be given good reason to believe in their dependability. However, those structures, placing any one teacher's judgment within a framework that links programme design to assessment outcomes, are not sufficient in themselves. There should also be an *infrastructure* that supports the development, implementation and ongoing operation of the assessment system.

Doubts have often been expressed as to whether we can expect even well-designed systems that depend on teacher judgment to gain ground in a world dominated by tests as the main or sole source of evidence of student attainment. However, there is, even in the schools sector in the UK, extensive experience of innovation should those responsible for policy choose to refer to the evidence. Summarising the findings from the Analysis and Review of Innovations in Assessment (ARIA) project, Gardner and colleagues reported two major conclusions:

> The first related to the lack of comprehensive planning ("under-designing") of many of the initiatives and the second related to perceptions of what constituted quality assessment practice.
> (Gardner et.al. 2011, p.112)

In contrast to the experience in vocational and higher education in the UK, the history of initiatives in UK schools to introduce teacher-based systems of summative assessment is not encouraging. Advocates of teacher judgment in summative systems tend to make reference to Queensland, where high-stakes school-based assessments at the end of secondary school have been in place since 1972, for want of examples in the UK or the US of teacher judgment having an accepted role in systems where the stakes are high. Black and Wiliam's conclusion from a review of large scale systems in seven countries is also worth noting:

> …while many systems rely on teacher judgment for assessments that are high stakes for students, there are … no systems that rely on teacher judgment for assessments that are high stakes for teachers.
> (Black & Wiliam, 2007; 11).

This paper has focussed on the *technical* attributes of assessment systems. But the social and political *context* within which the systems are initiated and then evolve is at least as important for successful implementation. The original design of the 'coursework'

element in the GCSE systems of certificating 16 year-olds in England was technically flawed even though it had emerged from more than twenty years of increasingly widespread use of teacher judgment in the examinations for 16 year-olds that preceded it. In the mid 1980s the political climate for introducing a teacher-based examination had been favourable. There was a broad consensus, led by a Minister on the political right, in support of greater reliance on work undertaken by students during the course because the GCSE grades would then more accurately reflect what each student 'knew, understood and could do'. Ten years later the political context of the 1990s, with the increasing use of aggregate data from the examination as a high profile indicator of school performance compounded by all-too-obvious technical flaws in some coursework assessments, led to a reduced dependence on coursework and its eventual replacement by 'controlled assessments' .

In the optimistic world of post-devolution Wales the then Education Minister argued:

> The informed professional judgment of teachers, lecturers and trainers must be celebrated without prejudice to the disciplines of public accountability
> (NAfW, 2001)

Such was the prevailing political climate at the time when the decision was taken to abandon national testing of 11 and 14-year-olds in Wales. Ten years later her successor as Minister is under pressure, in response to evidence from the 2009 PISA survey, to improve the performance of schools in Wales. It remains to be seen whether the National Curriculum assessment arrangements that replaced national tests in Wales at the end of key stages 1, 2 and 3, with teacher judgment having a central role, will be sufficiently robust to survive those pressures. The Minister will undoubtedly be aware of the value that students and their parents, in Wales as elsewhere, continue to place on test results, attitudes which Harlen (2011) refers to as representing 'formidable obstacles to serious use of teachers' assessment'.

The development, introduction and evolution of large-scale systems of assessment will always be a matter for public policy. The need for an infrastructure to support teacher judgment within such systems can be argued, as I have argued here, in technical terms. But it is the social and political context that will determine whether such systems are credible in the eyes of a sometimes sceptical public and, if implemented, will prove to be sustainable.

Acknowledgements: My thanks to Jo Hazell for the diagrams in this paper.

---

[i] The devolved government in Wales was referred to initially as the National Assembly for Wales and subsequently as the Welsh Assembly Government. Now, since May 2011, it is the Welsh Government.

References

ACCAC (2004). *Review of the school curriculum and assessment arrangements 5-16: a report to the Welsh Assembly Government*. Cardiff: ACCAC (Qualifications, Curriculum and Assessment Authority for Wales).

Andrews, L. (2011). *Teaching makes a difference.* Speech by Wales Education Minister, Leighton Andrews; 2nd February 2011. Available at: http://wales.gov.uk/topics/educationandskills/allsectorpolicies/ourevents/teachingmakesadifference/;jsessionid=zDkvNBbKs4NZHXYzTJswvTRTVpnjTx1SGVt6BXGBkC2PYQQ7Zv1b!1895062788?lang=en (accessed 9th June 2011)

Baird, J. et.al. (2011). *The Reliability Programme: Final Report of the Technical Advisory Group.* London: Office of Qualifications and Examinations Regulation.

Black, P. et.al. (2011, forthcoming). Can teachers' summative assessments produce dependable results and also enhance student learning? *Assessment in Education.*

Black, P. & Wiliam, D. (2007). Large-scale assessment systems: Design principles drawn from international comparisons. *Measurement* 5(1), 1–53.

Bradshaw, J. et.al. (2010). *PISA 2009: Achievement of 15 year-olds in Wales.* Slough: National Foundation for Educational Research.

Crooks, T., Kane, M., and Cohen, A. (1996). Threats to the valid use of assessments. *Assessment in Education* 3(3), 265–285.

Daugherty, R. (1986). *Examining Geography at 16+.* London: Secondary Examinations Council.

Daugherty, R. et.al. (2004). *Learning pathways through statutory assessment: key stages 2 and 3. Daugherty Assessment Review Group final report.* Cardiff: Welsh Assembly Government. Available at: http://web.archive.org/web/20040810192853/www.learning.wales.gov.uk/scripts/fe/news_details.asp?NewsID=1226 (accessed 9th June 2011)

Daugherty, R (2010). Summative assessment: the role of teachers. In Peterson, P., Baker, E. & McGaw, B. (eds.) *International Encyclopaedia of Education.* Volume 3, 384-391. Oxford: Elsevier.

Estyn (2010). *Evaluation of the arrangements to assure the consistency of teacher assessment in the core subjects at key stage 2 and key stage 3.* Cardiff: Estyn. Available at www.estyn.gov.uk (accessed 9th June 2011)

Fitz, J. (2000). Governance and identity: the case of Wales. In Daugherty, R., Phillips, R. & Rees, G. (eds.), *Education policy making in Wales: Explorations in devolved governance,* 24-46. Cardiff: University of Wales Press.

Gardner, J. et.al. (2010). *Developing Teacher Assessment.* Maidenhead: Open University Press.

Gardner, J. et.al. (2011). Engaging and empowering teachers in innovative assessment practice. In: Berry, R. & Adamson, B. (eds.) *Assessment Reform in Education,* 105-120. Springer: New York.

Harlen, W. (2005). Trusting teachers' judgment: research evidence of the reliability and validity of teachers' assessment used for summative purposes. *Research Papers in Education,* 20(3), 297-313.

Harlen, W. (2011, forthcoming). Taking charge of assessment. *Education Review.*

James, M. (2011). Assessment for learning: research and policy in the (dis)United Kingdom. In: Berry, R. & Adamson, B. (eds.) *Assessment Reform in Education,*15-32. Springer: New York.

Mansell, W & James, M. & the Assessment Reform Group (2009). *Assessment in Schools. Fit for purpose? A Commentary by the Teaching and Learning Research Programme.* London: Economic and Social Research Council, Teaching and Learning Research Programme.

Maxwell, G. (2010). Moderation of student work by teachers. In Peterson, P., Baker, E. & McGaw, B. (eds.) *International Encyclopaedia of Education.* Volume 3, 457-463. Oxford: Elsevier.

National Assembly for Wales (2001) *The learning country: a comprehensive education and lifelong learning programme to 2010 for Wales*. Cardiff: National Assembly for Wales. Available at: http://web.archive.org/web/20011112070144/www.wales.gov.uk/subieducationtraining/content/learningcountry/tlc-contents-e.htm (accessed: 9th June 2011)

Resnick, L. and Resnick, D. (1992). Assessing the thinking curriculum: New tools for educational reform. In Gifford, B. & O'Connor, M. (eds.) *Changing Assessments: Alternative Views of Aptitude, Achievement and Instruction*, 39–75. Boston, MA: Kluwer.

Wilmut, J. (2004). Experience of Summative Teacher Assessment in the UK: A Review. London: Qualifications and Curriculum Authority.

TERMINAL EXAMINATIONS						COURSEWORK				
Stage	**Participants**					**Stage**	**Participants**			
	Chief Examiner	Team Leader	Asst. Examiner	Board Officer	Moderator		Chief Moderator	Board Officer	Teachers	Asst. Moderators
DRAFTING EXAM PAPERS	●					DEVISING OF TASKS			●	
MODERATING EXAM PAPERS *	●			●	●	APPROVAL OF TASKS	●			●
DRAFTING MARK SCHEME	●									
	Candidates answer question papers						Candidates complete coursework tasks			
EXAMINERS MEETING *	●	●	●	●						
MARKING	●	●	●			MARKING			●	
						MODERATING MARKING *	●	●		●
STANDARDISING	●					STANDARDISING	●	●		

Marks for terminal examination and coursework are combined			
Stage	**PARTICIPANTS**		
	Chief Examiner	Team Leader	Board Officer
AWARDING *	●	●	●
BORDERLINING	●		●

* indicates a stage observed in this study

Fig. 1: Stages in the examining process

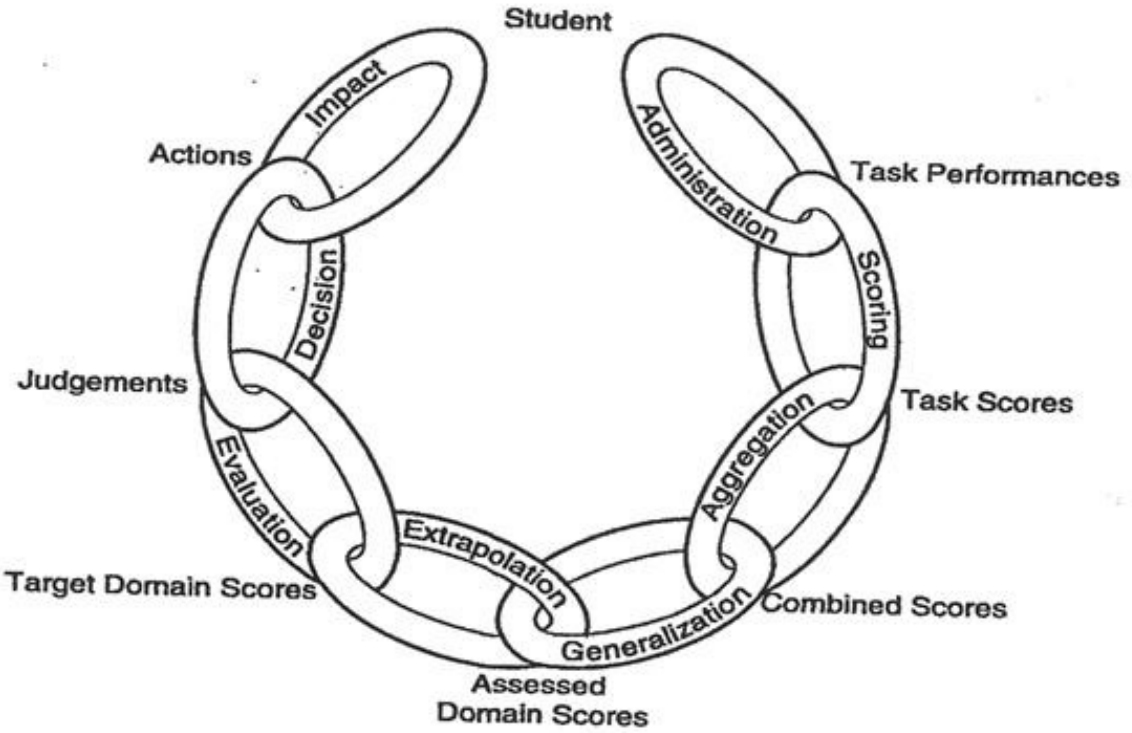

Fig. 2: Crooks, T. et. al. (1996) 'Threats to the Valid Use of Assessments' ***Assessment in Education*** **3 (3) 265-286**

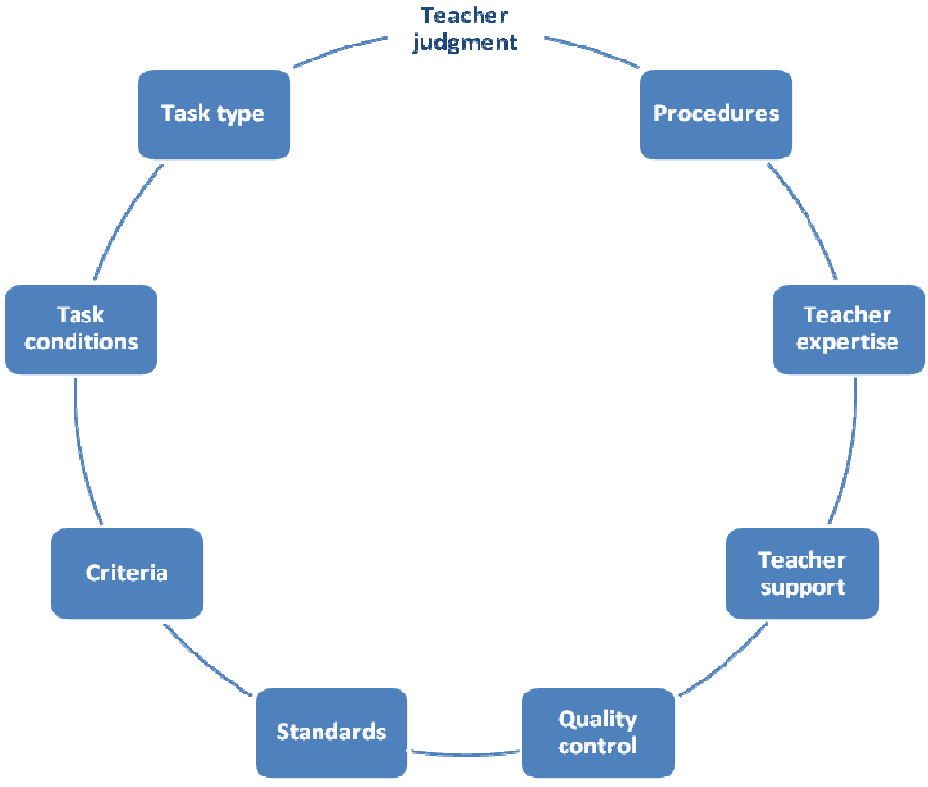

Fig. 3: Adapted from Crooks, T. et. al. (1996) 'Threats to the Valid Use of Assessments' ***Assessment in Education*** **3 (3) 265-286**

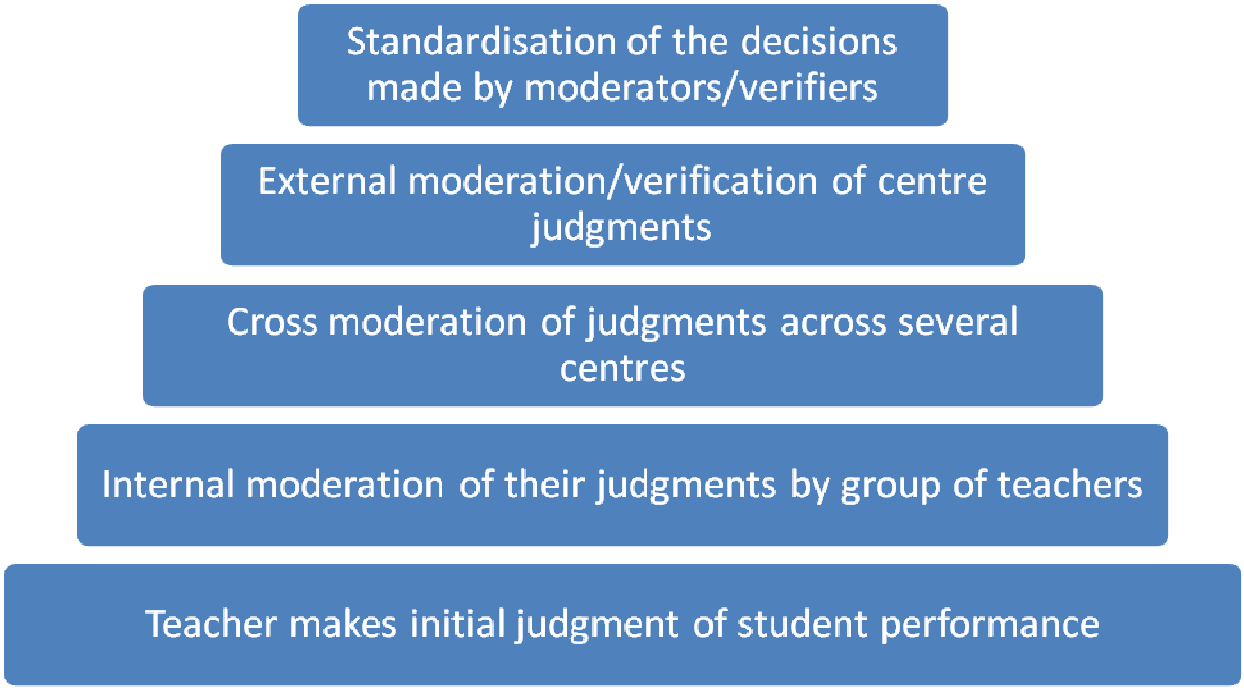

Fig. 4: Tiered structure of moderation