Dadansoddi ar gyfer Polisi
YMCHWIL GYMDEITHASOL Y LLYWODRAETH
GSR
GOVERNMENT SOCIAL RESEARCH
Analysis for Policy

Llywodraeth Cymru
Welsh Government

www.cymru.gov.uk

Ymchwil gymdeithasol
Social research

Rhif/Number: 32/2013

An investigation into Key Stages 2 and 3 teacher assessment in Wales

An investigation into Key Stages 2 and 3 teacher assessment in Wales

Prepared by the Australian Council for Educational Research.

(Views expressed in this report are those of the researcher and not necessarily those of the Welsh Government.)

For further information please contact:
Stephen Page
Department for Education and Skills
Welsh Government
Cathays Park
Cardiff
CF10 3NQ
Tel: 029 2082 5751
Fax: 029 2082 6014
e-mail: Assessment@wales.gsi.gov.uk
Welsh Government Social Research, 2013
Digital ISBN 978 0 7504 9542 4

WG18867

Table of contents

List of figures

List of tables

List of acronyms

ACCAC	Qualifications, Curriculum & Assessment Authority for Wales
ACER	Australian Council for Educational Research
AfL	Assessment for Learning
ARG	Assessment Reform Group
CM	Chief Moderator
DCELLS	Department for Children, Education, Lifelong Learning and Skills
DCM	Deputy Chief Moderator
DfES	Welsh Government’s Department for Education and Skills
EAS	Education Achievement Services
ERIC	Educational Resources Information Centre
FSM	Free School Meals
GCSE	General Certificate of Secondary Education
HKEAA	Hong Kong Examinations and Assessment Authority
HPE	Health & Physical Education
IAEA	International Association for Educational Assessment
IEA	International Association for the Evaluation of Educational Achievement
INSET	In-Service Education for Teachers
KS	Key Stage
LA	Local Authority
LEA	Local Education Authority
LCE	Local Consensus Event
NC	National Curriculum
NCA	National Curriculum Assessment
NFER	National Foundation for Educational Research
NRC	National Review Council
OECD	Organisation for Economic Cooperation and Development
PISA	Programme for International Student Assessment
QA	Quality Assurance
QBSSSS	Queensland Board of Senior Secondary School Studies
QSA	Queensland Studies Authority
SABER	Systems Approach for Better Education Results
SAT	Standard Assessment Task
SEF	School Effectiveness Framework
SEN	Special Educational Needs
SBA	School-based Assessment
TA	Teacher Assessment
TIMSS	Trends in International Mathematics and Science Study
WJEC	Welsh Joint Education Committee
WFL	Welsh First Language
WSL	Welsh Second Language

Executive summary

1. Aim of the investigation

The aim of this investigation was to provide the Welsh Government's Department for Education and Skills (DfES) with insight into how the end of Key Stages 2 and 3 (KS2 and KS3) teacher assessments are currently being conducted and whether they are fit for their intended purpose. This report outlines the findings of an investigation that was conducted in 2012 by the Australian Council for Educational Research (ACER) and provides discussion and recommendations to inform future policy and practice.

The investigation focused on a set of questions that relate to the reliability, validity, impact and operability of the current teacher assessment system.

2. Methodology

The investigation employed a mixture of qualitative and quantitative research methods that included the following:

2.1 Documentary and literature review

The investigation began with a wide-ranging documentary review encompassing policy documents and reports from the time the system of teacher assessment was introduced, and procedural documents generated during the implementation of the approach. The second stage of the investigation involved a detailed review of the literature on teacher assessments and moderation, particularly at the KS2 and KS3 levels. The review identified what works (and does not work) in teacher assessment and moderation and determined what models of assessment exist internationally for similar teacher assessment and moderation schemes. The review also addressed whether there was a relationship between type of assessment regime internationally and performance on international tests. Information from the literature review was used in answering some of the research questions.

2.2 Semi-structured interviews

Seventy-four stakeholders from 49 schools involved in the teacher assessment process were interviewed. This included Year 6 teachers, secondary subject leaders/assessment coordinators and headteachers. In the primary/secondary school clusters interviewees included cluster coordinators (core subjects). Also interviewed were Local Authority (LA) personnel responsible for moderation and KS2/3 transition. Chief and deputy chief moderators, contract managers and key staff were interviewed from the Welsh Joint Evaluation Committee (WJEC), who were responsible for conducting the external moderation process.

2.3 Process evaluation of the moderation procedures

The research team conducted a process evaluation of the moderation for English, Welsh First Language, and Welsh Second Language that took place in May/June 2012.

2.4 Questionnaires and interviews

Four separate but linked questionnaires were distributed to primary school headteachers, secondary school headteachers, primary/secondary school cluster coordinators, and local authority personnel responsible for moderation and the KS2/3 transition. The questionnaires were distributed according to a sampling plan to ensure that all sectors were represented. Questionnaire responses were received from 48 primary schools, 18 secondary schools, 17 cluster coordinators, and 14 of the 22 LAs. The response rates were lower than desired and a number of factors seem to have contributed to this. Because of the low response rate some of the results should be taken as indicative only.

2.5 Observations of the external moderation

In April−May 2012 the external moderation of Welsh First Language, Welsh Second Language and English Key Stage 2 and 3 tests organised by WJEC were observed and interviews were held with the Chief Moderators, Deputy Chief Moderators and WJEC staff.

3. Findings

The findings from the investigation are presented here as responses to the research questions.

3.1 Reliability of the current system

1. What works in terms of securing accurate teacher assessments?

Reliability is generally defined in terms of consistency across assessments or “the extent to which assessment can be trusted to give consistent information”. The central idea is that the results of an assessment would be the same or similar if the procedure were to be repeated under equivalent conditions. In the context of school-based standards-referenced assessment, reliability is related to consistency of teacher judgments and the comparability of reported results.

There must be some measure of understanding about consistency of teacher judgments and comparability of reported results if the results of school-based assessment are seen to be reliable in the sense of being dependable. Obtaining reliability in teacher assessment systems involves the use of a standards schema, criteria/standards matrix, grid or grading master that teachers apply for making decisions about the standard or level of student work.

The system in operation in Wales is a social or consensus moderation scheme, which is the most common form of moderation for school-based assessment systems. Social or consensus moderation is a quality assurance process that brings teachers together to review and discuss judgments across examples of student work, often in different assessments, and reach some level of agreement about the application of standards to that work. Such systems can produce levels of reliability comparable to other forms of assessment, but mechanisms need to be in place to ensure consistency of teacher judgments.

A recurring theme in the literature on teacher assessment is the level of expertise required of teacher-assessors. The skill sets that can be identified as necessary for a reliable system are (a) expertise in their disciplines, (b) total immersion in their level statements, (c) the ability to evaluate, and (d) the ability to negotiate when seeking consensus at moderation meetings. Teacher assessment systems typically have a component in which the original judgment by a teacher is checked in some way to ensure that there is a consistency of application of the standards.

2. Do the assessments demonstrate reliability?

a. What is the reliability and consistency of judgments made in the KS2 and KS3 assessment and levelling?

b. How does the KS2 system that is focused on the school cluster level compare to the KS3 system?

As it was beyond the scope of this investigation to perform a full-scale measurement and analysis of the reliability of the judgments in the current system, measures of the confidence of the primary-school headteachers, secondary-school headteachers, LA advisers and cluster coordinators were obtained from extensive questionnaires sent to each sector.

There is a problem with the reliability of teacher judgments in the assessment system. Confidence in the accuracy and reliability of all four components of the teacher assessment system is not very high.

The four components[1] of the teacher assessment system in Wales are:

1. Pupil assessment records
2. Internal standardisation procedures
3. Internal moderation procedures
4. Cluster moderation procedures

Primary schools expressed the most confidence in the reliability but even then, only 50−60% expressed this view. In contrast, about 80% of local authorities expressed some confidence in the accuracy and reliability of the system. Possible causes for the lack of confidence in the reliability of teacher assessment and ways that this might be addressed are described in other parts of the investigation.

The current system does not systematically gather data on the reliability of the judgments made by teachers. It is impossible to know how well the system is doing in regards to the consistency of teacher judgment until methods by which reliability can be measured are available. Some of these methods are described below. It is recommended that measures of the actual reliability of judgments be embedded into the teacher assessment systems so that it can be tracked and, if necessary, the implementation of the system can be refined to optimise the reliability.

A comment, which is relevant here but applies in the validity section also, relates to new forms of assessment that have been emerging (e.g., teacher assessment, assessment of higher-order thinking skills, assessment using a variety of instruments, and assessment tasks that emulate the kind of

[1] Not to be confused with the five elements of an effective assessment system described elsewhere

process-based tasks thought to represent good practice). Movement along a continuum from assessments where pupils select the correct response (as in multiple-choice tests) to assessments where pupils are required to generate a response (as in large open-ended tasks) can be associated with a decrease in reliability at the expense of an increase in validity. The challenge for assessment systems then becomes how to maximise validity and reliability simultaneously or how to trade off one against the other.

Consideration should be given to the following methods that are employed in other teacher assessment systems around the world to determine which might be suitable to introduce in Wales to enhance the reliability.

The description that follows of methods used to establish and monitor reliability in other systems does not assume that these methods or variations thereof are not used in or have not been considered in Wales. Nor does it assume that all are appropriate for teacher assessment at KS2 and KS3. These methods include:

Establishing reliability

Immersion

In an "immersion" process teachers study samples of student work to locate instances of the desirable features of work at a particular level rather than being given examples of student work at that level. An advantage of this approach is that it addresses some of the issues arising from the traditional and comparatively passive approach whereby teachers have to confront exemplars for which trading off has already occurred, which is especially problematic for the middle levels.

Marking rubrics

One of the ways to improve reliability in the marking process is to use marking rubrics. Marking rubrics are descriptive marking schemes developed by teachers or other assessors to guide the analysis of the products or processes of students' efforts. The descriptors should not to be too wordy, and they must convey meaning with clarity and precision.

Monitoring reliability

Marker monitoring for tests of written expression

The marker monitoring process involves the comparison of many different pairings of markers on the particular responses they have both marked in order to identify markers who are discrepant with other markers, with the purpose of re-calibrating them to enhance their accuracy.

Marker monitoring for constructed-response items

Another method for improving reliability of teacher judgments in the marking of constructed-response items is to check if the differences between the grades assigned to a student's response to an item (or assessment task) by a pair of markers are within a pre-set tolerance level for each item or if the differences are random. In the latter case markers would be required undergo further training.

Inter-marker agreement

A tool that has been used in the quality control procedures in the assessment of written expression and short-response items is weighted kappa. This is a statistical measure of the degree of agreement between two or more markers who assign scores to the same piece of work; it is weighted to take account of different levels of agreement, which is especially useful when marking guides have several levels.

Paired comparisons

The method of paired comparisons is useful in cases where objects (e.g. student portfolios) have to be judged only subjectively because it is impossible or impracticable to make relevant measurements in order to decide which of two objects is preferable or of higher quality (in the case of assessment, worth more marks or ranked higher). In this procedure objects are presented in pairs to one or more judges with all possible pairings being considered.

Post-hoc consistency check

Random sampling is designed to provide feedback and research data about consistency across different assessment periods. Sample folios of student work are selected and distributed to moderators who undertake a review in much the same way as they would as part of the routine quality assurance processes.

Comments on establishing and monitoring reliability

No mechanism, of itself, can guarantee reliability in teacher assessment. Reliable assessment only occurs after large-scale implementation strategies, or experience over time, or a tacit understanding amongst the practitioners. Only the first-mentioned of these is a transparent mechanism for disseminating standards. It is important, however, to recognise the second of these (experience over time) when making a realistic evaluation of an evolving system. Vital, therefore, are (i) teacher professional development (including in-built professional development of the type described in “immersion” above) and (ii) ensuring that marking rubrics, whatever form they take, provide teachers with a simple structure for assessment, written in such a way that multiple interpretations of the standards are not likely to occur but rather that the intended standards are the applied standards.

Recommendation 1 – Reliability

The current system does not systematically gather data on the reliability of the teachers’ assessments within clusters. External moderation of assessments depends upon the choice of portfolios submitted by clusters. Until methods by which reliability can be measured are available, it is impossible to know how well the system is doing in regards to the consistency of teacher judgment. It is recommended that measures of the actual reliability of judgments be embedded into the teacher assessment systems so that it can be tracked and, if necessary, the implementation of the system can be refined to optimise the reliability.

3.2 Validity of the current system

1. Do the current assessments accurately reflect the actual ability of the learner?

Overall, the current assessments were judged by the participants in the investigation to accurately reflect the actual ability of the learner, although there were differences in the opinions of the LAs, cluster coordinators, and the secondary and primary schools. Primary schools had most confidence in the accuracy with 93% feeling that the assessments were an accurate reflection of student ability. This drops to 75% confidence in the secondary schools and to 50% at the LA level. Cluster coordinators had 87% confidence, which is somewhere between the levels of confidence in the primary and secondary schools. In addition, there were complaints from secondary schools that pupils' functional literacy on entry to the secondary school in KS3 did not always correspond to the NC level they were awarded, particularly in relation to English and Science. This erodes the validity of the assessment because it casts doubt on the judgments being made by teachers.

2. What are the threats to validity of the current system?

Validity is generally understood in terms of the extent to which an assessment can be seen to assess what it is intended to assess; that is, that the evidence produced in response to the assessment is likely to be a suitable demonstration of the targeted aspect(s) of learning. Although there are many views on validity of assessment systems, they are all underpinned by fundamental notions about establishing the appropriateness of assessments; namely, that assessments should (a) measure what they purport to measure, (b) demonstrate predicted relationships with other measures of the intended constructs, (c) contain content consistent with their intended use, and (d) be put to purposes that are consistent with their design and are supported by evidence.

As the teacher assessment system is designed to measure pupil work that is the result of work in the classroom in lessons aligned to curriculum standards, the validity of the system is built upon a sound base. However, in the operation of the system, circumstances can arise that lead to an erosion of the validity, and that is what is reported as happening in the current system.

Another threat to the validity of the system is emerging from the fact that, with schools being compared to one another, there is pressure on them to raise their pupils' performance. This pressure seems to be leading to an inflation of the teachers' judgments of pupils' work and this, in turn, results in a lack of confidence that a judgment is accurate.

Another identified threat to validity is the use of taped oral evidence and other "less robust evidence of pupil attainment" which secondary teachers felt disadvantaged pupils by not preparing them for KS4 where pupils undertake externally assessed examinations based on writing and comprehension only.

Other teacher assessment systems around the world employ different strategies to enhance the validity of the system, and consideration should be given to which of them might be suitable for introduction to the Welsh system. The description that follows of methods used to establish and monitor validity in other systems does not assume that these methods or variations thereof are not used in or have not been considered in Wales. Nor does it assume that all are appropriate for teacher assessment at KS2 and KS3.

The methods include:

Internal moderation

Internal moderation is a form of peer-review whereby teachers of the same subject in the same schools meet to share and discuss student work and provide feedback to each other about the way standards have been applied. Internal moderation is already implemented in the Welsh system although, as reported elsewhere in this investigation, it is not being implemented consistently.

Alignment of assessment with curriculum and pedagogy

Valid assessments are aligned with curriculum and pedagogy. Most teacher assessment systems establish validity by using assessment techniques that reflect classroom experiences, not only in assessment format but also by allowing unlimited time (within reason), computer-generated text as well as or instead of handwriting, and so on. Wales does allow for diverse forms of evidence in the assessment process.

Construction matrix

A construction matrix is intended to ensure a range and balance of items and tasks across the portfolio of pupil work that is assessed. Range and balance can be represented by a construction matrix or grid in which characteristics of the assessment instruments[2] are tabulated, characteristics such as perceived difficulty, estimated time for completion of instrument, curriculum element(s) or objective(s) being assessed, and nature of the text that dominates in the instrument (e.g., verbal, numerical, spatial) to name a few.

Face validity

Face validity[3], while based on opinion rather than facts, is particularly important in new assessment systems. The opinions of parents, the general public and government cannot be overestimated for the ultimate success of an initiative. As identified above and elsewhere in this report, the face validity of the teacher assessment system in Wales is under threat because of perceptions of components that are seen to not be working.

Panelling

Panelling (or reviewing) is primarily a validation exercise often used in a test development cycle. Experts work collaboratively in small groups, both at the item level and the test level, to review the features of a test that can affect validity. Although panelling for teacher assessment is not

[2] The complete set of the instructions, test questions, answer keys, marking guides and other components that make up an assessment

[3] A basic form of validity in which it is determined if an assessment appears (on the face of it) to measure what it is supposed to

widespread, getting a range of opinions from different types of expert can add validity to the assessment.

Accreditation of assessment tasks

Assessment instruments (or at least blue-prints of them) that are designed by teachers are submitted to an external panel for approval before administration to pupils in order to alleviate the need for special consideration at the end of the assessment programme should the assessments be then deemed invalid (say in content or difficulty). The purpose is to ensure that pupils are not disadvantaged because the assessments (possibly set by inexperienced teacher−assessors) were not capable of bringing forth evidence of learning.

Statistical evidence

Some teacher assessment systems convene a technical panel after the assessment has occurred to study and evaluate data relating to instances of possible bias against sub-groups of the population (differential item functioning). The purpose is to measure whether different sub-groups defined by gender (or other indicator of interest) and ability level differ systematically in their performance on the assessment. If they do, the assessment may be biased in favour or against a particular sub-group.

Recommendation 2 – Validity

There is evidence that the face validity of the current system is already under threat because some schools have lost trust in the judgments made in the teacher assessment process. While face validity is not the only form of validity, it is important that those who need to operate the teacher assessment system should believe that its outcomes are valid. Without that confidence in the system, it will falter. To restore confidence Recommendation 3 on the impact of the teacher assessment system and Recommendation 4 on the operability of the system that are contained in this report should be implemented.

3.3 Impact of the teacher assessment system

1. Is the implementation and delivery of the moderation programme in line with best practice?

a. How do the systems at KS2 and KS3 compare to alternative systems that have been shown to work effectively in other education systems around the world?

The Welsh teacher assessment system contains the five elements that have been identified in the research literature as being the essential elements of effective teacher assessment systems as identified by Allen (2003).

The five elements[4] as translated into the particular features of the Welsh system are:

1. There are numerous documents that provide guidance and support to teachers implementing the curriculum and assessment system;
2. Schools set out plans for covering the content and skills required by the National Curriculum;
3. Schools are required to produce folios of student work as part of the moderation process;
4. Teachers are required to judge student work against a set of pre-set standards; and,
5. Moderation sessions are held to validate the judgments that teachers make.

However, as demonstrated in the literature, having all of the requisite elements is important, but not sufficient to ensure the success of the system in practice. When considering the list of necessary elements in the process of assessment (Meiers et al., 2007 and Gipps, 2002), the implementation of the Welsh system is not yet meeting best practice, based upon the evidence gathered in other parts of this investigation.

This situation, however, is not surprising since the review of literature demonstrates that introducing such teacher assessment systems requires considerable change across schools and local authorities because it has implications not only for the assessment, but also for pedagogy. It requires a strong commitment to professional development and support and thus to the resourcing of these.

As is being experienced in the Welsh system, one of the most challenging features is ensuring the quality and consistency of teacher assessments of student work. It is evident from the experience of countries that are further along the path than Wales that it takes years and several iterations of the process to develop the levels of teacher skills and experience to achieve the desired levels of reliability of judgments.

Even in the system that has been running in Queensland for 20 years, which achieved high degrees of reliability of scoring and widespread acceptance by schools and universities, there is a culture of continuous theory-building. In comparison, the Welsh system is still in its infancy, having been in existence since 2005, but only since 2010 with the external moderation component. The evidence from this investigation is that there are certainly areas in the system where improvements can be made, but given the history of the development of other systems this is to be expected. The Queensland case study shows that high reliability is ultimately achievable with sustained training of teachers and continued refinement of processes.

2. Are the assessment and levelling procedures being implemented as planned?

a. Are schools actually following the procedures as designed? If not, why not?

Deficiencies in the implementation of the system are the source of the lack of reliability in teacher judgments and of the threats to validity within the system. Most areas of implementation within the

[4] Not to be confused with the four components of the Welsh system as described elsewhere

system could be improved upon, although some more than others. Overall, implementation at the secondary school level seemed to be more in line with the planned scheme, while primary schools lagged in their implementation. In particular, there seems to be a weakness in policy documentation in a proportion of the primary schools, which is of concern given that it is four years since the statutory introduction of the policy documents.

The levels of understanding of the teacher assessment system in both school sectors is less than ideal, with some schools indicating that teachers were still unsure of the difference between standardisation and moderation. In particular, schools reported low levels of familiarity with DfES publications. There are problems with the current scheme of training which assumed that once some teachers were trained, they would go back to their schools and “cascade” what they had learned to other teachers. In reality, this was not implemented evenly.

Generally, the tracking of pupil attainment data across subject areas seems adequate, but there is a range of different pupil tracking systems in use across the primary and secondary schools, which can lead to lack of compatibility that hampers transfer of data across schools.

Primary schools gave more weight than did secondary schools to standardised test data when they made best-fit judgments on NC levels.

Although internal standardisation seemed strong in the secondary schools, it was lagging among primary schools. While the majority of secondary schools have standardisation portfolios in place across most subjects, primary schools have much lower levels of standardisation portfolios across subject areas – an issue which grouping in clusters is intended to overcome at least in part. This cannot happen, of course, where some primary schools are not involved in cluster moderation, as is permitted in the current arrangements for schools that are small and send less than 50% of their pupils to a given secondary school.

Not all teachers have a robust understanding of the level descriptors, of the range within a level and of the process of standardisation and moderation. Primary schools felt that even where there were standardisation portfolios, they did not always represent an appropriate range of evidence. The lack of standardisation portfolios in primary schools and their poor quality will make it harder for primary teachers to make reliable judgments about pupil work. Although some schools indicated a developing awareness of these aspects, this is some five years since their introduction as statutory elements of assessment.

Secondary schools are implementing internal moderation but implementation in primary schools is very low in comparison. Similarly, secondary schools have learner profiles in place across most subjects, but primary schools have much lower levels of learner profiles in place. Interviews in the cluster focus groups revealed that small primary schools do not have internal moderation procedures in place for mathematics and science. This is due mainly to time constraints. The process does not wholly comply with statutory requirements.

Also the range of evidence in learner profiles could be improved. If primary schools are not engaging in internal moderation, the quality of teacher judgments about pupil work will be inconsistent and lead to a lack of reliability in the system.

From the observations of the external moderation process conducted as part of this investigation, it became apparent that the LA consortium did not seem sufficiently clear about their role in the process and that this role was not prioritised enough to ensure accuracy and reliability of the levels awarded at school and cluster levels.

3. How does the current cluster moderation programme for English, Welsh and Welsh Second Language work to improve assessment?

The majority of the secondary schools participate in cluster moderation, but for primary schools the participation was lower. Overall, clusters tended to have standardisation portfolios and learner profiles in place, except in mathematics and science. The majority of schools thought that the standardisation portfolios had appropriate commentaries and that the evidence in them reflected shared understanding of the NC levels.

Although the external moderation is already functioning well, there were some weaknesses observed in the moderation process and there are some improvements that could be made to bolster the system. One improvement is in the timing of the submission of pupil profiles. Also, the assessments were based on a narrow range of pupils' work and generated an attainment level based on one or two pieces of work. Another area for improvement concerns the fact that the guidance from DfES does not stipulate the number of pieces of work required for submission from clusters of schools.

The samples submitted by school clusters for assessment were of variable quality and were not always annotated to evidence clearly why they had been awarded a particular level. It was also observed that no uniform agreement existed amongst schools and amongst clusters of schools as to what was required to comply with a best-fit notion for awarding a particular level.

There was no requirement for schools/clusters to resubmit their samples/portfolios when issues were raised, or crucially when moderators could not verify the level awarded by the cluster. Schools were free to suggest levels based on their "professional interpretation" which, in some cases, may be wide of the mark.

It was noted that there was minimal impact arising from the local authority cascading back the moderators' comments about improvements looked for in next year's submission. Another potential improvement is related to the fact that chief moderators and their deputies are seconded from their substantive posts for the period of moderation only.

Recommendation 3 – Impact

The teacher assessment programme in Wales has in place all of the main components that high-quality teacher assessment systems across the world have, but there is room for improvement across the system in the implementation and delivery of the programme. There is a lack of consistency in how it is being implemented and so it is not operating fully as planned. As such the operation of the system is not fully in line with best practice. The areas of operation that require attention are:

3.1 – Schools' understanding of policy

There is a need to ensure that all schools' policy documents on assessment clearly define standardisation and moderation, and distinguish between them as well as describing arrangements for internal standardisation and moderation. Similarly, it should be ensured that school policy clearly describes the arrangements for cluster standardisation. Without the policies having this level of clarity, there is an insufficient base upon which to continue to operate the teacher assessment system and to train school staff in how it should be implemented.

3.2 – Transition plans

Transition plans and actual practice in the system need to be matched, and that match needs to be monitored more effectively.

3.3 – Teachers' understanding of the system

After four years of operation, teachers' understanding of the system and how it should work is still low. National training of teachers across both secondary and primary sectors is needed to improve understanding, which will set the basis for fuller implementation. Training in particular needs to cover the difference between standardisation and moderation, and on making best-fit judgments.

3.4 – Pupil tracking systems

Although standardisation of the pupil tracking systems may not be possible, steps should be taken to ensure that data can be transferred between systems, particularly between KS2 and KS3.

3.5 – Internal standardisation

Steps need to be taken to improve the implementation of internal standardisation among primary schools and to ensure that they have standardisation portfolios in place that represent an appropriate range of evidence.

3.6 – Internal moderation

Steps need to be taken to improve the implementation of internal moderation within primary schools and to ensure that they have learner profiles in place that represent an appropriate range of evidence.

3.7 – Cluster moderation

More primary-school teachers should be participating in cluster moderation meetings. Changes should be made to ensure that no feeder primary schools end up outside of the cluster moderation process as some do under the current 50% threshold rule. Also, attendance at cluster moderation meetings should be monitored closely enough to ensure that all who have to attend do so. Effort needs to be made to ensure that clusters have standardisation portfolios in place for all subjects, especially in mathematics and science that have been ignored until now.

3.4 Operability

1. Do the assessments and the moderation programme represent an effective use of resources?

While the majority of respondents to the questionnaires agreed the current teacher assessment procedures are an effective use of resources, about one-quarter of respondents across the education system do not feel that. This may change for the better if the assessment processes identified as needing improvement in other parts of this report are addressed. Just under two-thirds of schools, LAs and clusters saw external moderation as an effective use of resources, but over one-third felt that it was not.

2. How useful are the national curriculum level descriptors?

While the majority of schools considered the NC level descriptions useful for assessment, less than two-thirds of the LAs agreed. It is unclear why they have a different view to the schools. External moderators also viewed the NC level descriptions as useful for assessment and moderation, but they felt that some descriptions were imprecise and all would benefit from review.

3. What are the barriers to reliable and consistent teacher assessment?

a. What is working well and what is not working as well as desired?

Many of the schools, LAs and cluster coordinators commented in their questionnaires on many aspects of the barriers to reliable and consistent teacher assessment as they saw them, and contributed ideas about ways to improve the system. Themes that emerged from their comments were the reliability of teacher judgments; inflation of results; lack of external checking; lack of time; staff training and inclusion; simplification of the system; allowing time for the system to bed down; and pupil tracking systems as areas where changes need to be made to improve the effectiveness of the current system.

4. Is the moderation programme fit for the intended purpose?

a. Is the moderation programme assuring reliable and consistent teacher assessment outcomes nationally?

The moderation programme has satisfied a clear and worthwhile purpose at system and school level and within the wider school community. Its design − consensus moderation − is in line with effective moderation models used elsewhere. It satisfies its primary objective (to apply standards to level student work) and its secondary purpose (to develop teachers' assessment skills). The criteria for fitness for purpose have therefore been met. The quality of the programme, as opposed to its fitness, can and should be improved in ways described throughout this report.

Four ways of going about social moderation are given in Maxwell's (2006) work, together with the advantages and disadvantages of each. The types of social moderation are: (1) external moderators, (2) external moderation panels, (3) assessor meetings, and (4) assessor partnerships. They vary in terms of the location of control of the assessment system and the stakes of the assessment for which they are designed. Each has its own advantages and disadvantages. For example, in Type 1, external moderators would offer authoritative advice (a possible advantage) but incur substantial costs (a possible disadvantage). On the other hand, in Type 4, assessor partnerships can be locally organised and not need bureaucratic support (a possible advantage) although it should be noted that nothing might happen if it is voluntary (a disadvantage). Furthermore, in the external moderator model, external advice may not always be appropriate, while in the assessor partnerships model, partners may simply reinforce each other's errors and misconceptions.

One of the approaches (Type 3 – "assessor meetings") involves groups of teachers looking at examples of student work, discussing the extent to which these meet the expected standard or level, and coming to agreement on the level of attainment represented by each example. The group may comprise staff from different sub-groups within an established group from different schools (such as in cluster moderation).

KS2 and KS3 appear to demand different approaches, each of which is located somewhere between the category "Assessor meetings" (Type 3) and the category "External moderation panels" (Type 2) depending on the required degree of control and the high/low-stakes nature of the programme. Both types provide powerful professional development for those involved.

Experience from elsewhere demonstrates the added value of an external component.

Barriers to reliable and consistent teacher assessment outcomes nationally include the lack of a national training exercise for teachers and the verbosity and/or vagueness of level descriptors, both of which detract from the aim of reliability and consistency.

Recommendation 4 – Operability

Given that there is a need to increase the reliability of the teacher assessment system it would be wise to retain the external moderation scheme and extend it to other core subjects, as planned. It is a mechanism for ensuring that teacher judgments across the system are consistent and one of the few places where data are being gathered on reliability of those judgments at present.

While the phenomenon of teachers being unable to envision student work that matches a level description is not unusual in standards-based assessment, a shared understanding of standards in writing and in application is a cornerstone of moderated teacher assessment. There is a need to revise the level descriptions using a combination of curriculum experts, assessment experts, pedagogues, and editors with skills in instructional language and clear expression.

While national in-service programmes for teachers can be costly in both financial and human terms, teachers must not be subjected to conflicting messages from within the system and/or between schools and clusters. The costs involved in nation-wide professional development would be covered in the future because less time and energy would be expended in dealing with anomalies of interpretation and/or the consequences of colleagues inadvertently exchanging inaccurate second-hand information.

4. Conclusion

The picture that has emerged from this investigation is of a teacher assessment system that has the main components of successful systems elsewhere in the world. It has adequate levels of documentation about how the system should work and the responsibilities of the key participants in the process. However, the implementation of teacher assessment is an enormous task and there are many parts that must be functioning smoothly for the system to produce reliable teacher judgments and effective educational outcomes. The system has not yet achieved that level of functioning and it still requires some attention to the parts of the system that are not operating as designed. This will require careful scrutiny of the system and consultation with those involved in its operation. One thing that has become evident from this investigation is that the problems that need to be fixed are already known by many in the system, and what is more is that they have ideas for how to solve them.

List of recommendations

Recommendation 1 – Reliability

The current system does not systematically gather data on the reliability of the teachers' assessments within clusters. External moderation of assessments depends upon the choice of portfolios submitted by clusters. Until methods by which reliability can be measured are available, it is impossible to know how well the system is doing in regards to the consistency of teacher judgment. It is recommended that measures of the actual reliability of judgments be embedded into the teacher assessment systems so that it can be tracked and, if necessary, the implementation of the system can be refined to optimise the reliability.

Recommendation 2 – Validity

There is evidence that the face validity of the current system is already under threat because some schools have lost trust in the judgments made in the teacher assessment process. While face validity is not the only form of validity, it is important that those who need to operate the teacher assessment system should believe that its outcomes are valid. Without that confidence in the system, it will falter. To restore confidence Recommendation 3 on the impact of the teacher assessment system and Recommendation 4 on the operability of the system that are contained in this report should be implemented.

Recommendation 3 – Impact

The teacher assessment programme in Wales has in place all of the main components that high-quality teacher assessment systems across the world have, but there is room for improvement across the system in the implementation and delivery of the programme. There is a lack of consistency in how it is being implemented and so it is not operating fully as planned. As such the operation of the system is not fully in line with best practice. The areas of operation that require attention are:

3.1 – Schools' understanding of policy

There is a need to ensure that all schools' policy documents on assessment clearly define standardisation and moderation, and distinguish between them as well as describing arrangements for internal standardisation and moderation. Similarly, it should be ensured that school policy clearly describes the arrangements for cluster standardisation. Without the policies having this level of clarity, there is an insufficient base upon which to continue to operate the teacher assessment system and to train school staff in how it should be implemented.

3.2 – Transition plans

Transition plans and actual practice in the system need to be matched, and that match needs to be monitored more effectively.

3.3 – Teachers' understanding of the system

After four years of operation, teachers' understanding of the system and how it should work is still low. National training of teachers across both secondary and primary sectors is needed to improve understanding, which will set the basis for fuller implementation. Training in particular needs to cover the difference between standardisation and moderation, and on making best-fit judgments.

3.4 – Pupil tracking systems

Although standardisation of the pupil tracking systems may not be possible, steps should be taken to ensure that data can be transferred between systems, particularly between KS2 and KS3.

3.5 – Internal standardisation

Steps need to be taken to improve the implementation of internal standardisation among primary schools and to ensure that they have standardisation portfolios in place that represent an appropriate range of evidence.

3.6 – Internal moderation

Steps need to be taken to improve the implementation of internal moderation within primary schools and to ensure that they have learner profiles in place that represent an appropriate range of evidence.

3.7 – Cluster moderation

More primary-school teachers should be participating in cluster moderation meetings. Changes should be made to ensure that no feeder primary schools end up outside of the cluster moderation process as some do under the current 50% threshold rule. Also, attendance at cluster moderation meetings should be monitored closely enough to ensure that all who have to attend do so. Effort needs to be made to ensure that clusters have standardisation portfolios and learner profiles in place for all subjects, especially in mathematics and science that have been ignored until now.

Recommendation 4 – Operability

Given that there is a need to increase the reliability of the teacher assessment system it would be wise to retain the external moderation scheme and extend it to other core subjects, as planned. It is a mechanism for ensuring that teacher judgments across the system are consistent and one of the few places where data are being gathered on reliability of those judgments at present.

While the phenomenon of teachers being unable to envision student work that matches a level description is not unusual in standards-based assessment, a shared understanding of standards in writing and in application is a cornerstone of moderated teacher assessment. There is a need to revise the level descriptions using a combination of curriculum experts, assessment experts, pedagogues, and editors with skills in instructional language and clear expression.

While national in-service programmes for teachers can be costly in both financial and human terms, teachers must not be subjected to conflicting messages from within the system and/or between schools and clusters. The costs involved in nation-wide professional development would be recovered in the future because less time and energy would be expended in dealing with anomalies of interpretation and/or the consequences of colleagues inadvertently exchanging inaccurate second-hand information.

1. Background

1.1 Aim of the investigation

The aim of this investigation was to provide the Welsh Government's Department for Education and Skills (DfES) with insight into how the end of Key Stages 2 and 3 (KS2 and KS3) teacher assessments are currently being conducted and whether they are fit for their intended purpose. This report outlines the findings of an investigation that was conducted in 2012 by the Australian Council for Educational Research (ACER) and provides discussion and recommendations to inform future policy and practice.

The investigation focused on a set of questions that relate to the reliability, validity, impact and operability of the current teacher assessment system.

1.2 Key questions addressed in the investigation

The investigation focused on answering the following set of questions, which are grouped into four clusters that address reliability, validity, impact and operability.

Reliability

1. *What works in terms of securing accurate teacher assessments?*
2. *Do the assessments demonstrate reliability?*
    a. *What is the reliability and consistency of judgments made in the KS2 and KS3 assessment and levelling?*
    b. *How does the KS2 system that is focused on the school cluster level compare to the KS3 system?*

Validity

1. *Do the current assessments accurately reflect the actual ability of the learner?*
2. *What are the threats to validity of the current system?*

Impact

1. *Is the implementation and delivery of the moderation programme in line with best practice?*
    a. *How do the systems at KS2 and KS3 compare to alternative systems that have been shown to work effectively in other education systems around the world?*
2. *Are the assessment and levelling procedures being implemented as planned?*
    a. *Are schools actually following the procedures as designed? If not, why not?*

3. *How does the current cluster moderation programme for English, Welsh and Welsh Second Language, work to improve assessment?*

Operability

1. *Do the assessments and the moderation programme represent an effective use of resources?*
2. *How useful are the National Curriculum level descriptors?*
3. *What are the barriers to reliable and consistent teacher assessment?*
    a. *What is working well and what is not working as well as desired?*
4. *Is the moderation programme fit for the intended purpose?*
    a. *Is the moderation programme assuring reliable and consistent teacher assessment outcomes nationally?*

The next section of the report describes the methods that were applied to answer these questions and the results of the investigation follow in the subsequent sections that deal with the questions of reliability, validity, impact and operability.

2. Methodology

The investigation used multiple qualitative and quantitative methods to answer the evaluation questions. Methods included a documentary and literature review, an evaluation of processes, detailed questionnaires and interviews, plus observations of the external moderation

2.1 Documentary and literature review

To provide an understanding of the current system of teacher assessments at the end of Key Stages 2 and 3 in Wales, we began the investigation with a wide-ranging documentary review of the policy documents and reports from which the existing system originated, and of the subsequent procedural documents generated during the implementation of the approach.

A list of relevant documents for the review was drawn up in consultation with DfES and included the following:

- The Daugherty Report: Learning Pathways through Statutory Assessment: Key stages 2 and 3
- Ministerial statements
- National Curriculum for Wales and supporting materials
- School Effectiveness Framework
- Assessment Policy – NFER
- PISA results
- Estyn reports on teacher assessment
- Transition Plans Evaluation
- 2010 External Pilot and Reports
- WJEC Review of the 2010 External Moderation Pilot
- Arrangements for the 2011−12 External Moderation

These documents were analysed and the key points relevant to this investigation and the evaluation questions were synthesised. The results are presented in section 3 of this report.

The second stage of the investigation involved a detailed review of the literature on teacher assessments and moderation, particularly at the Key Stage 2 (KS2) and Key Stage 3 (KS3) levels. At the beginning of the literature review ACER researchers in discussion with DfES defined the areas of interest and the search terms to be used. Articles that matched the search terms were located through ACER’s Cunningham Library. Filter criteria were then applied in the selection of those that would be used in the review. See Appendix 1 for details of review procedures including search criteria and sources.

2.2 Semi-structured interviews

In the third stage of the investigation we identified stakeholders and invited a sample of them to attend focus group sessions. The stakeholders invited included Year 6 teachers and headteachers involved in the assessment and levelling process, and at the secondary level it included subject leaders/assessment coordinators. In the primary/secondary school clusters interviewees included cluster coordinators (core subjects). It also included Local Authority personnel responsible for moderation and the KS2/3 transition. Ten focus group sessions were conducted and these were attended by 74 representatives from a total of 49 schools.

Interviewed at the WJEC level were chief and deputy chief moderators and WJEC contract managers. Also interviewed were key WJEC staff at their head office. These interviews occurred at two distinct points in time – WJEC before the questionnaire and all others after it.

2.3 Process evaluation

We conducted a process evaluation of both the assessment and levelling procedures that occurred in the moderation for English, Welsh and Welsh Second Language that took place in May/June 2012. We designed an observation instrument that looked at two aspects of implementation. The first aspect related to whether the programme is being implemented as designed and the second as to whether the elements of best practice were observable in the programme. The observation instrument also allowed for ad hoc comments and judgments as issues arose.

2.4 Questionnaires and interviews

We designed and distributed four separate but linked questionnaires (see Appendix 2). One was for primary school headteachers and asked about KS2 assessment procedures in the school. The second questionnaire was sent to secondary-school headteachers to ask about assessment procedures at that level. The third was aimed at primary/secondary school cluster coordinators to enquire about KS2/3 cluster assessment procedures. The fourth questionnaire was sent to Local Authority personnel responsible for moderation and the KS2/3 transition to ask about that process. The questionnaires were distributed according to a sampling plan to ensure that all sectors were represented.

Table 1 shows the number of schools, clusters and local authorities in the samples and the response rates from each sector.

Table 1: Questionnaire response rates

Sector	Sample size	Responses	Response rate (%)
Secondary schools	20	18	90
Primary schools	219	48	22
Cluster coordinator	33	17	52
Local authority	22	*18	82

* One response included the views of the five local authorities that make up the Consortium

De Ddwyrain Cymru (Southeast Wales EAS) – Blaenau Gwent, Caerphilly, Monmouthshire, Newport and Torfaen.

2.4.1 Local Authority Questionnaire

A questionnaire that comprised 53 questions was sent to advisers in all 22 Local Authorities (LAs). The questions were designed to assess what LAs know about the assessment system in their schools and what training and other support they provide for its implementation.

Questionnaire responses were received from 18 of the 22 LAs. One covered the five LAs which now comprise the Education Achievement Service (EAS) for South East Wales – Blaenau Gwent, Caerphilly, Monmouthshire, Newport and Torfaen. No returns were received from most of the LAs that make up the Central South Consortium (Bridgend, Cardiff, Merthyr Tydfil and Vale of Glamorgan). Rhondda Cynon Taf was the sole exception.

Appendix 3 provides a list of respondents to the questionnaires.

2.4.2 Primary and Secondary Schools Questionnaire

The questionnaire sent to primary and secondary schools comprised 64 questions that dealt with the full range of topics on the operation of the teacher assessment programme in the school. It covered questions on policy, transitions, teachers' understanding of the system, the tracking system, standardised tests, internal standardisation, internal moderation, learner profiles, cluster moderation, and standardisation portfolios.

Eighteen of the twenty secondary schools that were sampled returned a questionnaire, which represents a 90% response rate. However, only 48 of the 219 primary schools sampled responded, which is only a 22% response rate.

The response rate for primary schools was much lower than desired and a number of factors seem to have contributed to this. School reorganisation and changing of staff roles was reported by some as being the reason for non-completion. A possible attribution of the observed apathy among some headteachers and their feeling that it was not worthwhile contributing to research on what they considered was a flawed system might be the move from local authority support to a consortium.

Many LA school support officers, some with responsibility for teacher assessment arrangements, retired at the end of August to be replaced by consortium staff. This led to uncertainty as to who in the consortia was responsible for teacher assessment. In addition, the first half of the autumn term, when questionnaires were sent out, was a very busy time for schools. They were responding to a host of new initiatives and may have seen the request to complete the questionnaire as adding to an already packed agenda.

The low response rate for primary schools means that the data collected from the questionnaires must be regarded with caution as it is impossible to know how much it represents the views of the primary schools as a whole. It can only provide some indications of their views.

2.5 Observations of the external moderation

In April−May 2012 the external moderation process was organised by the Welsh Joint Education Committee (WJEC) and two members of the investigation project team observed the moderation of Welsh First Language, Welsh Second Language and English at WJEC's offices, Upper Boat on 30th April, 10th May and 24th May 2012 respectively. During those observations, discussions were held with the Chief Moderator, Deputy Chief Moderators and WJEC staff.

3. Findings

3.1 Origins of the current system of teacher assessment

The current system of teacher assessment at the end of Key Stages 2 and 3 in Wales has evolved over several years and it is important to understand its origins before reporting on the rest of the findings in the investigation. This section, which draws upon the documentary review described in section 2.1, covers the following topics:

3.1.1 The policy background – assessing the National Curriculum for Wales

3.1.2 Supporting materials for the National Curriculum in Wales

3.1.3 The School Effectiveness Framework

3.1.4 Benchmarking against PISA

3.1.5 Estyn's reports on teacher assessment

3.1.6 Transition planning

3.1.7 The External Moderation Pilot, 2010

3.1.8 WJEC's Review of the External Moderation Pilot, 2010

3.1.9 Arrangements for External Moderation in 2011−12

3.1.10 Description of the current system

3.1.1 The policy background – assessing the national curriculum for Wales

From 2000 to 2004, statutory end of Key Stage 2 and 3 tests based on the Programmes of Study within the National Curriculum were taken by all eligible pupils in Wales at age 11 and 14 respectively. Statutory testing took place in the four core subjects of English, Welsh, mathematics and science. Tests in mathematics and science were available to schools as both English-medium and Welsh-medium versions.

Statutory end of KS1 (pupils aged 7) external tests in the core-subjects of Welsh, English, mathematics and science were removed in 2000 by the Welsh Assembly Government. At the end of KS1 teacher assessment has been relied upon to assess and report on pupil progress and development during the period post 2000. Prior to 2000, standard assessment tests and tasks for English, mathematics and science were, in general, developed jointly with England. Since 2000, all statutory tests at Key stages 2 and 3 in Wales have been developed and published independently of the test materials for England.

In 2003, Professor Richard Daugherty was invited by the Minister for Education and Lifelong Learning to undertake a review of the National Curriculum assessment arrangements for 11- and 14-year-olds in Wales. The Daugherty Assessment Review Group's Final Report was published in May 2004. It had many parallels with the Qualifications, Curriculum & Assessment Authority for Wales's (ACCAC) Report to the Welsh Government in April 2004 (Review of the School Curriculum and Assessment Arrangements 5−16).

In the Foreword to his Assessment Review Group's Final Report, May 2004, Daugherty said:

> "The proposals in this Final Report are centred on the educational needs of pupils whilst also ensuring that evidence from assessment is available to inform those who need to know about the learning of individuals and of groups. The proposals build on the foundations of the current system but also include innovative elements. Each element contributes to what we believe is a coherent set of recommendations. When fully implemented the assessment arrangements will be able to play their part in helping young people make the most of their learning opportunities during the vital middle years of schooling." (*Learning Pathways through Statutory Assessment: Key stages 2 and 3, p. 4*)

Seven of the Review Group's 26 Recommendations are relevant to the current investigation. They are

> Recommendation 10: A set of moderation procedures should be introduced, based on local school clusters, to build the confidence of all concerned in the consistency of statutory teacher assessment.
>
> Recommendation 16: A system for the moderation of statutory teacher assessment at Key Stage 3 should be introduced in order to ensure that data on pupil attainment is sufficiently robust to be fit for the several purposes associated with its use.
>
> Recommendation 17: Secondary schools should be accredited as having in place procedures to maximise the consistency of statutory teacher assessment in each National Curriculum subject.[5]
>
> Recommendation 18: Teachers should have access to a range of support materials and professional development activities that are necessary for the effective moderation of teacher assessment.
>
> Recommendation 23: The development of assessment for learning practices should be a central feature of a programme for development in Wales of curriculum and assessment.
>
> Recommendation 24: Schools and local authorities should review the range of assessment demands on pupils and teachers to ensure that each form of assessment activity plays its part in an overall pattern of assessments that is well matched to purposes and does not result in pupils being over-assessed and, in particular, over-tested.

[5] This recommendation was never implemented as intended. It was substantially diluted with only the moderation reports to indicate compliance. In addition, where there was significant disagreement, a single re-submission was required but no additional demand subsequent to this. "Accreditation" is not discussed further in this report.

> Recommendation 25: For the purpose of comparing standards in Wales in key aspects of learning with standards in other countries Wales should participate in the PISA survey of 15-year-old pupils from 2006.

The Daugherty vision underpinned the Welsh Assembly Government's decision to replace external tests by a system of moderated teacher assessments. In announcing on 13th July 2004 that:

> "Statutory teacher assessments at the end of key stages 2 and 3 should remain, but should be strengthened by moderation and accreditation arrangements. ACCAC will be remitted to design systems and checks to ensure that teacher assessments are robust and consistent. At the end of key stage 3, this will include secondary schools being awarded accredited centre status. Over the next three years, the current tests will be gradually removed. However, ACCAC will continue to provide optional test and assessment material to support teacher assessments until the new arrangements are in place. In key stage 2, a new style of diagnostic test, which focuses on skills, will be developed and introduced in year 5. The information derived from these tests will help teachers to identify individual strengths and weaknesses, which can be developed or addressed in the final year of primary school and which will provide valuable information for secondary schools to work with when pupils change phases."

The then Minister for Education and Lifelong Learning, Jane Davidson, reasoned that:

> "… both Professor Daugherty and ACCAC acknowledge that the current statutory tests that form part of those arrangements put teachers under pressure to teach to the tests, do not help the transfer from primary to secondary school, narrow the scope of the curriculum, particularly during the final year of the primary phase—as Estyn has also regularly reported—and, subsequently, have a negative effect on teaching and learning. There is clear evidence, therefore, that change is needed if we are to get the best from our pupils, the curriculum and our teachers. I propose, therefore, to move away, during the next four years, from the current testing regime to a system which is more geared to the pupil, focuses more on skills, and puts teacher assessment at its heart."
> (*http://wales.gov.uk/about/cabinet/cabinetstatements/2004/130704JDACCAC?lang=en*)

In a later article, Professor Daugherty explained that:

> "National curriculum assessment (NCA) in Wales has evolved from common foundations into a system that is now distinct from that in England. The influence of the political and social milieu of Wales can be seen both in the distinctive features that have been in place from the outset and in the more radical changes introduced since 2002." (National curriculum assessment in Wales: adaptations and divergence, *Educational Research, Vol. 51. No. 2, June 2009,* p. 247)

> "Two parallel reviews, one by the government's own advisory body (ACCAC 2004) and the other from an independent ad hoc group (Daugherty 2004), made similar recommendations that would form the basis for the changes to national curriculum assessment that are now being put in place over a five-year period from 2005 to 2010. There are four main strands[6] to those changes (Daugherty 2008):
>
> - 'assessment for learning' as a central element in curriculum and assessment across all key stages;
> - 'skills profiles' for every pupil to be reported and developed from Year 5 onwards [7];
> - 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; and
> - assessment by teachers in all National Curriculum subjects at the end of key stage 3, backed by a national system of secondary school accreditation." (*ibid,* pp. 248–249)

And, perhaps most importantly, asserted that:

> "The current changes to assessment policy and practice in Wales rely on the professional expertise of teachers, both through assessment for learning practices in their classrooms and in teachers' ability to make valid judgments of attainment that are reliable enough for the purposes for which summative data is required." (*ibid,* p. 249)

The Daugherty proposals were adopted and Wales set off on a path which was consciously very different from that of England and which was intended to lead to a statutory National Curriculum assessment system based on moderated teacher assessments which themselves were designed to support and enhance pupils' learning.

The National Curriculum for Wales is specified at

http://wales.gov.uk/topics/educationandskills/schoolshome/curriculuminwales/arevisedcurriculumforwales/nationalcurriculum/?lang=en. The National Curriculum subjects were revised and restructured in 2008. Currently they cover the 'core' subjects of English, Mathematics, Science and Welsh and together with Art and Design, Design and Technology, Geography, History, Information and Communication Technology, Modern Foreign Languages, Music and Physical Education. Whilst not formally

[6] Note that Professor Daugherty omitted the fifth strand – international benchmarking – from this list.

[7] Such skills profiles were piloted but never introduced. Consequently, the impact of introducing assessment for learning and thinking skills strategies was never monitored or evaluated.

part of the National Curriculum, for Wales, there is a National Exemplar Framework for Religious Education for 3- to 19-year-olds in Wales, together with related documents which parallel those of National Curriculum subjects.

3.1.2 Supporting materials for the national curriculum in Wales

A range of support materials was developed by the Curriculum and Assessment 3–14 Division of the Department for Children, Education, Lifelong Learning and Skills of the Welsh Assembly Government (and subsequently the Curriculum Division of the Department for Education and Skills of the Welsh Government), becoming available from 2008 onwards. The support materials are listed in Table 2.

Table 2: Support materials for the teacher assessment programme

English in the National Curriculum for Wales	*DCELLS, 2008*
English: Guidance for Key Stages 2 and 3	*DCELLS, 2008*
Ensuring consistency in teacher assessment: Guidance for Key Stages 2 and 3	*DCELLS, 2008*
Making the most of learning – Implementing the revised curriculum	*DCELLS, 2008*
Mathematics in the National Curriculum for Wales	*DCELLS, 2008*
School Effectiveness Framework	*DCELLS, 2008*
Science in the National Curriculum for Wales	*DCELLS, 2008*
Welsh in the National Curriculum for Wales	*DCELLS, 2008*
Welsh second language: Guidance for Key Stages 2 and 3	*DCELLS, 2008*
Skills and learning: English in the national curriculum for Wales	*DCELLS, undated*
Skills and learning: Mathematics in the national curriculum for Wales	*DCELLS, undated*
Skills and learning: Science in the national curriculum for Wales	*DCELLS, undated*
Skills and learning: Welsh in the national curriculum for Wales	*DCELLS, undated*
Skills framework for 3- to 19-year-olds in Wales	*DCELLS, 2008*
Science: Guidance for Key Stages 2 and 3	*DCELLS, 2009*
Strands in progression from the level descriptions for science Key Stages 2 and 3 – Poster ref: CAD/GM/0054	*DCELLS, undated*
Supporting learners' higher-order literacy skills	*DCELLS, 2009*
Mathematics: Guidance for Key Stages 2 and 3	*DCELLS, 2009*
Developing higher-order literacy skills across the curriculum	*DCELLS, 2010*
Supporting triple literacy: Language learning in Key Stage 2 and Key Stage 3	*DfES, 2011*
Welsh second language: Assessment materials for Key Stage 2 – Learner sheets	*DfES, 2011*
Welsh second language: Assessment materials for Key Stage 2 – Teacher's Handbook	*DfES, 2011*

Similar guidance was produced for the non-core subjects in the National Curriculum for Wales. All guidance documents are clear, well-written and accessible.

Unfortunately, these materials were not available at the point of introduction of the revised programmes of study, let alone their predecessors. This further complicated assessment issues and external moderation procedures. For example, KS3 non-core subjects had the option of submitting evidence for external moderation for either the outgoing or revised programmes of study.

Similarly, support materials for primary schools were produced in 2007 by NFER under the collective title *Getting to grips with assessment: primary*. These comprised leaflets on:

- Starting out in assessment
- Policy into practice
- Assessment for learning
- Self and peer assessment
- Interpreting information from different sources
- Moderation of assessment judgments
- Making the most of assessment data
- Assessment information and different audiences
- Understanding assessment information (a leaflet for parents)
- Understanding tests
- Glossary of key assessment terms
- Resources and useful web links

No matter how clear the guidance or the support materials, their late publication appears to have compromised the shift away from national external testing to moderated teacher assessment. This was compounded by the failure to ensure that appropriate training resources had been provided to schools at an early stage in that shift.

3.1.3 The School Effectiveness Framework

In 2008, the Welsh Assembly Government set out its inclusive approach to "aligning policies and their implementation to secure better outcomes" in Wales in the form of the *School Effectiveness Framework.* In acknowledging the "well established and internationally recognised knowledge relating to school effectiveness and improvement" this *Framework* represented:

> "the commitment of the Welsh Assembly Government and local authorities to apply this knowledge in the particular circumstances of Wales, as a major part of our drive towards improving outcomes for children and young people." (p. 3)

It claimed that:

> "The outcomes of the Programme for International Student Assessment (PISA) 2006 provide an international benchmark for the performance of our school system in Wales. The PISA results confirm that we have made progress across the board in children and young people's achievement. However, in relation to the age, ability, gender and socio-economic circumstances of our children and young people, this progress is uneven and needs to be advanced further. The Framework supports our response to PISA 2006 as part of our drive to ensure that high quality outcomes and

> equity of performance are firmly secured in all our schools.” (*ibid*)

The *Framework* established a tri-level arrangement: the Welsh Assembly Government, the Local Authorities and Schools. Under this tri-level agreement:

> “the Welsh Assembly Government is responsible for setting strategic policy direction. ... The Department for Children, Education, Lifelong Learning and Skills (DCELLS) has a special responsibility to make sure that all of its policies and funding contribute to the aims of the SEF and to the wider aims for children and young people. DCELLS needs to be constantly evaluating its work to make sure that it is benefiting our children and young people. DCELLS must use the data that it collects to ensure that local authorities and schools improve and must also provide the right level of challenge for schools and local authorities to make sure that improvement is sustained. Working with local authorities, DCELLS will develop outcome agreements to set priorities for improvement at the local level.”
>
> (See http://www.sefwales.co.uk/sef-p2-home/sef-p2-about-sef/sef-p2-about-sef-sef-and-wag.htm)
>
> “Local Authorities (LAs) are responsible, by law, for securing the best outcomes for the children and young people in their area. ... LAs will secure the best outcomes for all of the children and young people in their areas through their support and challenge function.”
>
> (See http://www.sefwales.co.uk/sef-p2-home/sef-p2-about-sef/sef-p2-about-sef-sef-and-las.htm)

and

> “Schools are responsible for the quality of the experience offered to children and young people enrolled. Schools also have a shared interest for the outcomes and wellbeing of all the children and young people living within the local area. Schools are responsible for enabling all children and young people to develop their full potential [and for] promoting a culture of social inclusion and respect for diversity by developing wellbeing and personalised learning.” (See http://www.sefwales.co.uk/sef-p2-home/sef-p2-about-sef/sef-p2-about-sef-sef-and-schools.htm)

The *Framework* identifies the key role of pupil assessment in supporting learning and providing accountability:

> “The ability of practitioners to set high but achievable targets for learners, to make rigorous use of formative and summative assessment methods, to provide feedback to them and to employ effective learning and teaching techniques are the keys to school improvement and accountability.” (*ibid*, p. 19)

and asserts that:

> "The Assembly Government has the responsibility for developing and reviewing the national curriculum and its assessment so that it is relevant to and capable of engaging all learners." (p. 22)

Yet there are no specific objectives set for assessment and it is therefore not surprising that assessment has disappeared from the updating paper, entitled *Developing a National System for Education in Wales by Embedding the School Effectiveness Framework*, produced two years later (in 2010).

Despite recognising that:

> "Wales has made progress across the board in children and young people's achievement. However, there continue to be differences in outcomes, within schools (where the greatest variation lies), between schools, between local authorities and between the primary and secondary phases of education. Together, these contribute to Wales' overall position within the PISA results and provide a benchmark for the improvement that we seek from our education system." (*ibid*, p. 6)

There is no specific reference to the development of assessment practice within the National Model for School Improvement, which identifies the importance of having "school improvement and effectiveness practitioners" who will, *inter alia*, "create and support networks of effective practice within, between and across schools focusing on teaching and learning, with a drive to improve outcomes for all." (*ibid*, pp. 7−8)

Nowhere did it recognise, as NFER had consistently understood, that

> "Assessment should become an automatic part of what teachers do; it should be viewed as one of the techniques of teaching, rather than a form filling activity to do at periodic intervals." (NFER, *Proceedings from a Policy and research seminar on Methods for Ensuring Reliability of Teacher Assessments*, June 2009)

that

> "results from informal assessments can be used to provide summative information. However, there are three conditions that must be fulfilled before it can be introduced successfully. Firstly, a major investment has to be made in teacher professional development in order to bring about a shared understanding of criteria. Secondly, a part of this professional development would need to address teachers' and advisers' understanding of the nature and purposes of the four quadrants of assessment, as shown in the table above. Finally, the system would need an element
>
> of external monitoring and accountability that commanded public and professional confidence." (NFER, *Assessment Policy*, 2009, p. 5)

but that

> “Moderation takes time and effort but can be very profitable and can feed into reviewing objectives and improving marking quality. For subject coordinators, moderation activities across a number of year groups can reveal strengths and weaknesses that may give rise to whole-school curriculum or assessment focuses. Wherever possible it is important and valuable to carry out cross-curricular moderation. For example, writing should be assessed across all areas of the curriculum – not just on the basis of written work in literacy.” (NFER, *Getting to grips with assessment: primary – moderation of assessment judgments*, 2007)

NFER presciently warned that:

> “It is frequently the case in education systems that information about performance collected for one purpose is then thought to be useful for additional purposes, for which the assessment was not originally designed. In the NFER submission to the Education Select Committee on Testing and Assessment in England seven different purposes for the end of key stage tests were given. The assessments and the results of these assessments were not originally designed to serve all these purposes, some of which have emerged over time, calling into question the validity of decisions made against these purposes.” (NFER, *Assessment Policy,* 2009, p. 9)

and this can be seen to have impacted directly on the self-confidence with which Wales was approaching Key Stage 2 assessment in particular.

3.1.4 Benchmarking against PISA

The Welsh Government had set itself a number of challenges in pursuing its distinctive path, including that of benchmarking its young people’s performance against other OECD countries through the triennial PISA tests.

The 2009 PISA tests came as a disappointment to Wales, with the mean scores for reading and mathematics both below the OECD average and significantly lower than that in the other parts of the UK. In both tests, Wales had fewer pupils at the highest levels of attainment than the average for OECD countries. Only in science did Welsh pupils perform at the OECD mean level but even this mean score was significantly lower than for the other countries of the UK (encapsulated from Executive Summary to Bradshaw, J., Ager, R., Burge, B. and Wheater, R. (2010). *PISA 2009: Achievement of 15-Year-Olds in Wales.* Slough: NFER, pp. vii–xi).

The Minister for Education and Skills, Leighton Andrews, in February 2011 expressed the Welsh Government’s concern about this performance in the context of its overall educational policies:

> “In Wales over the decade of devolution we have implemented many of the changes the profession wanted to see. We have worked with the profession. We don’t have league tables. We abolished SATS. We introduced the Foundation Phase. We created a skill-based curriculum. We have avoided many of the antagonistic

> competitive features of the English system. We do not have academies. We will not have Michael Gove's so-called free schools. We have maintained faith in the comprehensive model of education. As I said to Michael Gove last year, one of the advantages of devolution is that it allows England to be a laboratory for experiments.
>
> But if we believe in the comprehensive model in Wales, then we have to make sure that it delivers for all our children.
>
> The evidence of PISA is that it is not. Performance has fallen back. Why is PISA important? It is well established and internationally respected. As the Chief Inspector of Training and Education in Wales, Ann Keane, has said to me 'PISA tests the skills that should be at the core of any curriculum. The failure in Wales even to maintain what was a disappointing position in the results of the 2006 assessment raises many questions about our education system'." (*Teaching makes a difference*, Speech on 2 February 2011, Reardon Smith Theatre, Cardiff)

He took aim at a number of targets in the performance of pupils, schools, local authorities and his own department. He addressed the issue of measurement of pupil performance in robust and direct terms:

> "We have kidded ourselves about measurement for too long. While statistics have suggested that we do better than England at Key Stage 2 but then suffer a dip at Key Stage 3, I'm afraid I'm not convinced we are comparing like with like. I hear too many stories from secondary headteachers about the quality of entry from cluster primary schools and the inconsistent quality of teacher assessment. Estyn makes the same point and has reported on this. Then, later in the system, the reports from FE Principals about the quality of basic skills of their entrants from schools are shocking. Simply shocking. Meanwhile work-based learning providers tell me too many stories about their entrants arriving without evidence of the qualifications they are meant to have undertaken. Abandoning SATS was not meant to be a signal for anything goes." (*ibid*)

This theme was picked up explicitly by Lord Paul Bew's Review of KS2 testing, assessment and accountability for the Secretary of State for Education (in England), which included a section specifically devoted to Summative teacher assessment in Wales.

> "In 2005 the Welsh Assembly Government replaced statutory testing with a system of teacher assessment (moderated across local clusters of schools) designed to focus on individual pupils and smooth the transition from primary to secondary education. It offers a valuable case-study of the impact of such a change, which we have considered with great interest." (Lord Bew, *Independent Review of Key Stage 2 testing, assessment and accountability,* Final Report to the Secretary of State for Education, June 2011, pp. 51−52)

Taking their cue in part from Estyn's Evaluation of the arrangements to assure the consistency of teacher assessment in the core subjects at KS2 and KS3, (2010), the Bew Review Panel rejected the option of externally moderated teacher assessment for the National Curriculum at KS2 in England:

> "We believe the evidence from Wales and the current shift away from a reliance on teacher assessment in Wales suggests external testing should play an important role in a statutory assessment system which aims to provide data for accountability purposes." (*ibid*, p. 53)
>
> "We feel that a system entirely based on teacher summative assessment would not be sufficiently reliable for the purpose of providing school accountability data. While we believe ongoing and high quality assessment is crucial to ensuring pupils make good progress, we do not believe that schools should be held accountable through a system wholly based on moderated teacher assessment." (ibid, p.53)

3.1.5 Estyn's reports on teacher assessment

Estyn has been consistently critical of the validity and, especially, the reliability of moderated teacher assessment as a basis for assessing pupils' performance against the National Curriculum specifications at KS2/3. Its 2010 *Evaluation of* the *arrangements to assure the consistency of teacher assessment in the core subjects at key stage 2 and key stage 3* rehearsed the recent history of National Curriculum assessment in Wales and noted that:

> "The Welsh Assembly Government's policy on assessment aims to ensure that there is confidence in the validity of end-of-key-stage teacher assessment outcomes at KS2 and KS3. These arrangements rely on having in place robust and reliable teacher assessment across schools throughout Wales." (*ibid*, para 7)

However,

> "Assessment is one of the weakest areas of work in schools. Estyn's evidence from school inspections across Wales consistently indicates that about a quarter of schools inspected each year have shortcomings in aspects of assessment." (*ibid*, para 11)

and

> "The evidence now available from external moderation and verification has confirmed a mixed picture in terms of quality, ranging from schools where teacher assessment systems are outstanding to schools or individual subject departments which do not demonstrate high quality teacher assessment." (*ibid*, para 61)

The report contained ten recommendations, the effective adoption and implementation of which are germane to the present investigation.

These were addressed to use of level descriptions within the key stages; more effective sampling of teacher assessments; and further encouragement to local authorities to embed processes to secure better accuracy and consistency in teacher assessment), to the local authorities (recommending an active role in supporting the cluster moderation process directly and embedding processes more effectively) and to schools (arranging cluster moderation meetings at least annually; ensuring that teacher representatives from each school attend all cluster meetings held for the purposes of standardisation and moderation; ensuring that all teachers are aware of and use the relevant outcomes of their own cluster standardisation and moderation meetings; including the assessment of Welsh second language in school and cluster standardisation and moderation meetings; and ensuring that DCELLS guidance on assessment is taken fully into account). (*ibid*, p. 7)
Estyn's subsequent evaluation of the *Developing thinking skills and assessment for learning* programme found examples of good practice, "mainly in the primary schools", they reported that:

> "Despite the programme's emphasis on assessment, the quality of teachers' feedback in marking is not consistent enough, particularly in the secondary schools visited. In many of these secondary school departments, pupils' work is marked thoroughly and precise targets for improvement are identified. However, in other departments, teachers do not always mark work regularly and do not inform pupils of what they have done well or how they can improve their work. A whole school marking policy has therefore not been implemented." (Estyn: *The 'Developing thinking skills and assessment for learning' programme*, June 2011, para 12)

They identified leadership in the development of professional learning communities and the management of teachers' time as keys to the success of the programme:

> "The programme has had the greatest effect in schools where the senior management teams support its key principles. These leaders promote a culture of enquiry and action-based research among teachers. The schools that succeed best in delivering the programme have done so in the context of a professional learning community of colleagues who reflect on and discuss ways of improving what they are doing." (*ibid*, para 13)

But the overall picture was one of unacceptable inconsistency:

> "Despite the programme's emphasis on assessment, the quality of teachers' marking is not consistent enough, particularly in the secondary schools involved. In many of these secondary school departments, pupils' work is marked thoroughly and precise targets for improvement are identified. However, in other departments, teachers do not mark work regularly and do not inform pupils of what they have done well and how they can improve their work. There is, therefore, insufficient feedback on progress towards targets. Marking is more consistent in the primary schools visited whereas a whole-school marking policy has not been implemented in the secondary schools visited." (*ibid*, para 52)

This has significant implications for the present investigation, in that it suggests that whereas there may be greater marking consistency within certain primary schools, these may be part of a cluster led by a secondary school which has not developed such consistency within and between its own departments.

The report notes that, despite the care taken initially to get the Developing thinking skills and assessment for learning programme under way, this initial preparation and early support proved insufficient to ensure the programme's success.

> "the programme lacked clear expectations about how to measure the impact on teaching and pupils' standards and wellbeing. No success criteria were agreed at the outset of the pilot or the extension. Lessons were not learnt from the challenges of measuring the impact as identified in the evaluation of the pilot programme by BMG (an external research company) in 2008. The Welsh Assembly Government did not provide clear enough guidance on the need to evaluate the impact of the programme on pupils' standards. As a result, many of the local authorities and just over half of the schools lacked awareness of the need to measure the impact. As a result, schools are unable to demonstrate how their strategies to develop thinking and assessment for learning have improved pupil standards or wellbeing." (*ibid*, para 60)

Equally,

> "Nearly all local authorities monitor the development of classroom practice regularly through a mixture of classroom observations and discussions with senior management teams, staff and pupils. However, there is a lack of consistency in the way schools and local authorities monitor impact." (*ibid*, para 61)

These conclusions from inspection evidence suggest a high level of inconsistency and unreliability in the application of national assessment standards to pupils' performance and thus raise significant doubts about the efficacy of planning for the transition from primary to secondary education.

3.1.6 Transition planning

Introducing the concept of Transition Plans in 2004, the then Minister for Education and Lifelong Learning, Jane Davidson, said that:

> "Making the move from primary to secondary school can be a difficult and traumatic time for even the most confident pupils. The transition from being a "big fish" in Year 6 to a "minnow" in Year 7 can be a daunting experience"

and indicated that

> "The aim is provisionally for Transition Plans to be in place to support pupils moving to Key Stage 3 in September 2008. The timetable for Transition Plans will run alongside and complement changes flowing from the Daugherty Report on statutory assessment at Key Stage 2."(Welsh Assembly Government, Easing the transition from primary to secondary school, 25 November 2004 at

> http://wales.gov.uk/newsroom/educationandskills/2004/4024530/?lang=en)

DCELLS recognised that:

> "many LEAs and schools have good plans and programmes to improve educational transition. The most effective plans identify the different aspects of transition, how improvements will be made and the respective roles and responsibilities for leadership and management of the transition process.
>
> However, while most schools have improved some aspects of transition, only a few primary and secondary schools have joint comprehensive policies to formalise transition arrangements, including agreed approaches to managing transition, sharing information about pupils' achievements and learning needs, teaching and learning methods, assessment, tracking pupils' progress, curriculum organisation and professional development. Most schools monitor and review transition arrangements but do not evaluate the impact of practice on learning and standards.
>
> To address this, the Assembly Government, using powers in section 98 of the Education Act 2002, has introduced a requirement that maintained secondary schools and their maintained feeder primary schools establish plans to facilitate the transition of pupils from primary to secondary school." (National Assembly for Wales Circular No: 30/2006, *Guidance on the preparation of Key Stage 2 to Key Stage 3 Transition Plans*, p. 4)

Recognising also that some primary schools' progression links are with a number of secondary schools, the Guidance required that

> "the establishment of Transition Plans therefore is limited to instances where there is an established and ongoing relationship between a primary school and a particular secondary school founded on the majority of the Year 6 cohort from the primary school transferring to the secondary school." (*ibid*, p. 6)

In other cases, the Department expected that voluntary arrangements which used the "overall format" of the Transition Plan should be made. A further exemption was permitted where the total number of registered pupils at a primary school is fifty or less at the end of the school year.

These were important exceptions to DCELLS' original intention, in that they allowed a number of schools an exemption from the statutory requirement for transition planning, including the formal coordination of assessment arrangements.

All Transition Plans were intended to set out the arrangements for

> "Achieving consistency in assessment and monitoring and tracking pupils' progress against prior attainment. For example, improving opportunities for teachers to work together to assess the work of pupils as they move from Year 6 to Year 7, including

> moderation of Teacher Assessment.” (*ibid*, p. 9)

To support effective Transition Planning, the Department indicated that ring-fenced funding from the Better Schools Fund and the Key Stages 2–3 Transition Grant had been made available for 2006−07 (*ibid*, p. 25) and two additional days for In-Service Education for Teachers (INSET) had been made available in both the 2006−07 and 2007−08 school years which could be “used for training in relation to the preparation and implementation of plans aimed at supporting pupils making the transition from primary to secondary school.”

Estyn was remitted to report on *The impact of transition plans* two years later. It reported that:

> “Overall, the first generation of three-year transition plans meets Welsh Assembly Government requirements. Plans include information on how schools intend to improve arrangements in the five core aspects of transition. Nearly all plans also include arrangements in optional areas, such as pastoral links. Many schools had already met some of the requirements before they became statutory in September 2007. However, transition plans vary significantly in quality. Most are at least satisfactory and a few are very good.” (Estyn, *The impact of transition plans – An evaluation of the use of transition plans by primary-secondary school partnerships to improve the quality of learning and standards*, June 2008, p. 5)

On assessment, it reported that:

> “Most primary and secondary schools have begun to assess and moderate pupils’ work together in one or more of the core subjects. Many clusters have also started to produce portfolios of pupils’ work that exemplify achievement at different levels. The majority of transition plans do not show how this effective practice will be extended in line with the roll out of statutory requirements during 2008−2010.Opportunities for moderation of teacher assessment vary too widely. Few LEAs have strategies to ensure the accuracy and consistency of teacher assessment across their primary schools at the end of key stage 2.” (*ibid*, p. 6)

The report made eight recommendations to schools, the sixth of which was that they should

> “give priority to cluster group assessment and moderation of pupils’ work in line with the roll out of statutory requirements during 2008−2010” (*ibid*, p. 8)

Similarly, the first of its three recommendations to the Welsh Assembly Government was that it should

> “monitor and evaluate the roll out of the statutory requirements for assessment at the end of key stage 2” (*ibid*)

In conclusion, the Inspectorate reported that

> “Although most clusters are committed to improving the quality and reliability of

> teacher assessment at the end of key stage 2, opportunities for moderation between schools vary too widely. Few LEAs have strategies to ensure the accuracy and consistency of teacher assessment across their primary schools." (*ibid*, p. 22)

Furthermore,

> "LEA arrangements to help schools improve transition are not consistently in place across all authorities ... In a few authorities, schools have not had enough support. With a few exceptions, LEAs are not monitoring the process of transition planning or outcomes rigorously enough. Many schools are uncertain about funding for transition work in the short and long term. In many authorities, there is a lack of clarity and transparency about the funding available to schools for transition initiatives." (*ibid*, p. 23)

Estyn followed up this evaluation in 2010, reporting that

> "School inspections in 2008–2009 show that transition planning is now a strong feature of the life and work of most secondary schools and their partner primary schools … National Curriculum teacher assessment results at the end of key stage 2 and key stage 3 show that a higher proportion of pupils in 2009 are achieving the expected level in the core subjects compared to 2006. … However, these arrangements are too variable in quality across school cluster groups to assure consistency in teacher assessment in the core subjects at key stage 2 and 3." (Estyn, *Transition plans and grants − An evaluation of the impact of transition plans and grant on primary and secondary school partnerships at key stage 2 and key stage 3*, March 2010, p. 3)

> "Ensuring that primary and secondary school teachers are fully aware of the standards expected at the respective key stages has been a challenge for schools and local authorities. Bringing teachers together for cluster moderation meetings to standardise and moderate pupil work has helped, but key stage 2 and key stage 3 teachers have only recently needed to work collaboratively to moderate and standardise their teacher assessments." (*ibid*, p. 15)

The Welsh Assembly Government's response to this report included the key paragraph

> "Plans are now in hand for piloting of a similar arrangement for external moderation of end of Key Stage 2 assessments. A central feature of this work will lie in cluster group working to ensure a shared understanding of standards by practitioners in secondary schools and partner primary schools and to facilitate exchange of assessment information. This work will focus on the core subjects." (Estyn Remit Reports − Welsh Assembly Government Response, May 2010, p. 3)

The responsibility for administering that Pilot was given to WJEC under contract.

3.1.7 The External Moderation Pilot, 2010

The original *Model for KS2/3 cluster group moderation pilot 2010 (March 2010)* was built on the assumption that one Chief Moderator and two Deputy Chief Moderators for each of the five core subjects, covering primary and secondary areas (and bilingual requirements for science and mathematics), would be appropriate. Each subject team would then comprise 5 primary and 5 secondary specialists working as Assistant Moderators. For mathematics and science, a minimum of 50% of the team would need to be fully bilingual.

Each Local Authority was asked to identify and confirm one pilot KS2/3 cluster group for each of the core subjects (i.e. 4/5 separate primary/secondary school clusters, depending on whether Welsh first language and/or Welsh second language operates within Local Authority). Each cluster would pilot only the core subject nominated for the cluster. The main objectives of the pilot are shown in Table 3.

Table 3: Main objectives of the external moderation pilot conducted in 2010

1.	Highest rate of moderator productivity which is sustainable	(min/max range and average number of clusters moderated based on pair/ single moderator working)
2.	Sufficient evidence for example learner profiles including cluster commentaries	(for moderators to be able to accurately moderate evidence at top or bottom of level)
3.	Format of evidence	(paper based, electronic)
4.	Manageability of evidence for clusters	(extent to which requirements of model reflect cluster practice/state of preparedness across Wales)
5.	Effective report generation	(high-quality, useful reports with minimum QA workload implications)
6.	Effective training of moderators	(methodology, sample materials, checklists, criteria)
7.	Effective QA procedures	(active monitoring, QA starts from day one of moderation etc. Needs to be efficient in terms of CM/DCM/project managers' time and outcomes)
8.	Effective communications/logistics with cluster groups/Local Authorities	(clear, manageable, sustainable arrangements)
9.	'Customer satisfaction'	(schools, LAs, manageability/value questionnaires, see also 10 below)
10.	Positive and measurable impact on accuracy/reliability of end of key stage TA/national data	(value added, evidence that moderation of two levels within range L3−5 is sufficient to achieve this objective)
11.	Risk management	(to support future roll out)
12.	Confirmation of cost estimates for national implementation and value for money	(to confirm budget requirements – see also 10 above)

The intention was that three cycles of external moderation would be needed to cover "c. 240" clusters by autumn 2013. Thus, "c. 80 clusters per core subject will need to be moderated annually." It was estimated that a team of 18−20 moderators would be required per subject for this national roll-out, based on moderator productivity (pair working) of 10 clusters (three whole days plus part of training day). This would require each pair of moderators to review 4 learner profiles/teacher commentaries and generate a report for each cluster. Each pair of moderators would moderate c. 8−9 clusters during external moderation.

The policy objectives which underpinned this programme were given as:

1. Nationally consistent and reliable teacher assessment for 7−14 year old learners
2. Secure end of Key Stages 2 and 3 national data that Wales-based and external audiences trust
3. Increased teacher confidence in shared understanding of standards within each primary/secondary cluster group
4. Measurable improvements in the effectiveness of assessment systems and reliability of teacher assessment outcomes over the period 2011−2014 (compared with benchmark of Estyn survey in 2009).

(*Securing Teacher Assessment at Key Stages 2 and 3*, 10.03.10 SAF Meeting)

WJEC reported that:

> "Training went according to plan for all subjects with the exception of WSL where they felt in future they need an extra day. All teams managed to complete moderation of all materials; however some teams required additional time: science used an additional day and WSL Chief and Deputy Chief took some materials home to complete. With the exception of maths all teams required additional time to complete the QA reports." (*Securing Key Stage 2 and 3 teacher assessment: Year 3 Interim Report* (September 2010 – January 2011))

Following the 2010 pilot, all contacts from the 21 pilot clusters were invited to complete two questionnaires − the first on the manageability of the selection and despatch of sample

evidence to the external moderators and the second on the value of the external moderators' feedback report and how it would impact on future teacher assessment.

Ninety-three responses were received to the first questionnaire with roughly half of the respondents from primary and half from secondary schools. There were 21 responses from English and Mathematics cluster contacts, 23 from Science, 15 from Welsh First Language and 13 from Welsh Second Language. No further breakdown of responses by school phase was given.

Seventy-four (74) considered the pilot model to be appropriate, whilst 13 disagreed; 82 and 85 respectively found the WJEC guidance "clear" and "useful"; 88 considered the pilot administrative instructions from WJEC to be "clear", whilst 86 reckoned them to be "sufficiently detailed"; 74 considered the arrangement to be manageable for their cluster.

Most contacts (63) were able to secure release during their normal working week to attend cluster moderation meetings; 21 used an INSET day, 32 a twilight session, and 9 made some other arrangement.

In response to a series of more detailed questions, more than 60% agreed that “the pilot for KS2/3 external moderation encouraged all schools within my cluster to attend cluster group moderation meetings”; “the pilot’s requirements fit well with current practice within my cluster group’s moderation”; “collating and sending the learner profiles to the external moderators was manageable for me as the cluster contact”; and “the pilot’s focus on lower end and/or top end of levels was manageable for my cluster” whilst 80% or more agreed that “the pilot’s focus on learner profiles at Levels 4 and 5 at each key stage was manageable for my cluster”; “the pilot’s focus on two learner profiles at each key stage was manageable for my cluster”; and “the pilot has strengthened my cluster group’s understanding and application of ‘best-fit’ judgments to learner profiles”. However, just 54% agreed that “the pilot ensured that all schools in my cluster shared learner profiles for the cluster to moderate”.

What is striking about these responses is that the heart of the cluster-based moderation process (what has been called elsewhere “social moderation”) was so clearly missing for almost half of the schools taking part in the pilot. And this is amongst those schools that had agreed to take part in the pilot. Noticeably, 6 of the 93 contacts strongly disagreed with the statement that “the pilot’s requirements fit well with current practice within my cluster group’s moderation” – the highest incidence of “strong disagreement” in the survey.

Detailed comments from contacts on all five subjects were received and analysed.

The second feedback questionnaire concerned the value of the external moderators’ report. Forty-two questionnaires – significantly fewer than for the manageability questionnaire – were returned, even though the sample group was identical. Twenty-three were from primary school contacts, 15 from secondary school contacts, and 4 others. Once again, no breakdown by school phase was given.

All respondents considered the layout of the external moderation report to be clear and 38 found it sufficiently detailed. Forty found the combination of textual comments and summary codes accessible and helpful. The two who did not would have preferred more text. Thirty-nine (39) found the comments were informative and helpful for future teacher assessment and 35 considered the comments appropriately detailed from the perspective of the respondent’s cluster. Thirty-seven (37) agreed that the comments were informative and helpful for the future selection of sample evidence for the cluster's own future use.

Taken as a whole, 19 found the external moderation report's comments very useful, 17 found them of some use, and 5 of very limited use; 25 fully agreed that their cluster agreed with the external moderation report's comments whilst a further 14 said that there was some agreement. Similarly, 28 agreed that the external moderation report provided comments in a constructive and supportive manner whilst 14 said there was some agreement.

Although there is a sense that this sample was only that of the really committed, their responses to that final question are impressively supportive of the model used during the pilot. The WJEC team collated detailed additional comments made by respondents which were then fed back into the planning process for the 2011−12 exercise.

The External Moderators' reports from the pilots were also of great value in this respect, especially in respect of the technical issues and resource demands of the process.

External moderation in English was hampered by the fact that:

> "Whilst many clusters provided evidence from across the three ATs [Attainment Targets], 62% of the profiles moderated included insufficient evidence for moderators to agree with the cluster's best-fit judgment. Whilst this was sometimes due to a lack of cluster commentary and/or stimulus/resource materials, it was usually due to a lack of evidence for one of more attainment target e.g. only one task for Oracy, Reading and/or Writing; or insufficient range, particularly literary and non-literary evidence for Reading and Writing." (*ibid*, p. 2)

Furthermore,

> "Only one cluster failed to present one of the four learner profiles requested although a number of clusters had clearly not identified whether their profiles were lower/top end of the level as requested."(*ibid*)

This was *despite* the clear guidance on portfolio requirements given by WJEC. Beyond *sufficiency*, however, there were serious concerns about the nature and appropriateness of the materials used as the basis for assessing reading skills:

> "In addition there was an over-reliance, particularly at KS3, for clusters to use past SATs material as evidence for reading. Whilst some had successfully made use of these to reflect more open tasks using the SAT stimulus material, some were still evident of test/examination conditions and therefore could not be considered part of 'normal classroom practice'." (*ibid*)

Successful learner profiles, on the other hand,

> "demonstrated a range of evidence, usually two or three tasks for each attainment target with evidence of group and individual work for Oracy and both literary and non-literary evidence for Reading and Writing." (*ibid*)

Moderators were *unable to agree* with the best-fit level of a majority of learner profiles due to insufficient evidence, but they:

> "only *disagreed* with the best-fit level for 5 out of the 80 profiles [6.25%].In most cases this was due to features of a lower level being more evident in the profile than the cluster had indicated" (*ibid*)

and concluded that:

> “this demonstrates a consistency in the identification and application of level characteristics to learners’ work at both key stages.” (*ibid*, p. 3)

They identified a number of areas of both good practice and areas for development:

“Good practice:

- Detailed commentary related to level descriptions
- Learner profile overview
- Contextual information included
- Resources included
- Good range of evidence covered
- Reference to adjacent levels in cluster commentary
- Interesting range of tasks that reflect NC 2008
- Poetry as evidence of Writing
- Good range of texts used for Reading tasks
- Common cluster skills ladders/grids/commentary framework used by both key stages
- Summary paragraph/overview at the end of each attainment target which makes clear the best-fit level for that attainment target
- Clear indication and rationale for the best-fit level for the profile as a whole.

“Areas for development:

- Wrong levels submitted i.e. Levels 3 and 6
- Focus on evidence from years other than years 6 and 9 (i.e. not end of key stage)
- Commentary didn’t reflect cluster understanding of level description
- Cover sheet not included
- Lack of non-literary evidence
- Use of ‘tests’ e.g. SATs papers (under timed conditions)
- Lack of stimulus materials for Reading to assist moderation
- Tasks submitted too similar, not reflecting range
- Over emphasis on literary tasks, especially at KS3
- Lack of evidence resulting in moderators being unable to agree with cluster judgment
- Quality of evidence making photocopies difficult to read / audio evidence inaudible

- Difficult to assess over-scaffolded tasks
- Group writing tasks difficult to use to assess individual contribution
- Tasks included not referenced to a specific attainment target
- Lack of clear commentary referring to 2 levels (to exemplify best-fit at lower/top end of level)." (ibid)

They indicated that "to build on current best practice at KS2 and KS3, clusters should ensure that:

- they demonstrate evidence of moderating learner profiles rather than standardising individual pieces of work;
- learner profiles include evidence for each attainment target. As noted in the guidance and exemplified during the pilot's information meetings, secondary evidence through teacher commentary for Oracy is appropriate;
- learner profiles for school and cluster moderation include a range of evidence for each of the attainment targets. The evidence should sufficiently demonstrate the level characteristics relevant to the learner profile (lower end/top end or securely within);
- cluster commentary reflects the level characteristics demonstrated in the learner profile and demonstrates the cluster's overall best–fit judgment;
- task setting reflects the Skills and Range of the 2008 English programme of study e.g. a balance of non-literary and literary evidence for both Reading and Writing and evidence of a range of Oracy tasks. This is essential to ensure that assessment reflects the normal classroom practice relevant to the programme of study including the 2008 level descriptions.
- cluster commentary includes contextual information for each task;
- evidence includes resource/stimulus material where relevant." (ibid, p.4)

Finally, the Chief Moderator for English identified six Clusters as exemplars for future training (in respect of their learner profiles and moderator reports): in Anglesey, Cardiff, Conwy, Denbigh, Newport and Wrexham. (ibid, p. 8)

Similar issues were identified in the reports of the pilot Chief Moderators for Welsh First Language and Welsh Second Language.

In Welsh First Language, the Chief Moderator focused on inclusivity within the cluster moderation process and did not report the same difficulties with sufficiency of evidence which had bedevilled both the English and the Welsh Second Language moderation processes.

Key messages to the cluster groups – identifying profiles

> “It appears that many more teachers have been involved in the moderation this year and that the responsibility for collecting profiles, determining the best-fit level and writing commentaries has been jointly carried out. Moderating as a team has been valuable experience by giving cluster teachers a feeling of ownership and has cemented understanding of level requirements.
>
> However, there is room to believe that evidence from just one Primary School was used to represent Key Stage 2 in a few clusters. This does not give a correct picture of cluster members’ understanding of the national standards.
>
> Evidence collection should form a natural part of the procedures of all schools within the cluster throughout the academic year. Work schemes should be structured so that tasks included in a profile meet the requirements of level descriptions. This would ensure that they showed evidence of most strands of the level in question and that the evidence related to the appropriate learners namely years 6 and 9. It is not the amount of evidence which is important but how many Study Programme and level description requirements it reflects.
>
> Arranging regular meetings to moderate profiles can be costly but head teachers’ backing is sorely needed to free teachers to attend all meetings to ensure their understanding of the national standards at cluster level.”(*Key Stages 2/3 Cluster Group External Moderation*, *Pilot 2010, Chief Moderator for Welsh Report*, December 2010, pp. 5−6)

The report did not specifically commend the work of any particular cluster as a potential exemplar.

In Welsh Second Language, the Chief Moderator commented:

> “The main issues were lack of sufficient evidence (due to limited number of tasks / lack of diversity of context and / or situation), and lack of commentary, especially commentary on best fit.”(*Key Stages 2/3 Cluster Group External Moderation, Pilot 2010, Chief Moderator for Welsh Second Language Report*, December 2010, p. 2)

Good and poor practices were identified as follows:

Where the profiles are at their best they include:

- Evidence in various situations and contexts for the three attainment targets.

  Oracy – individual work, role play, pair, group

  Reading – evidence of reading aloud, responding verbally and / or in writing

  Writing – letters, diaries, post cards, articles

- Information about the context of setting and completing tasks

- Stimulus materials / pre-preparation (referenced in the commentary)
- Frameworks and guidance for tasks referenced in the commentary (as applicable)
- Copies of the reading texts including complete texts
- Commentary on the level descriptions highlighted within and across the tasks
- Clear cross-referencing between the evidence and the commentary with clear exemplification of the Level characteristics in the evidence:

  A detailed commentary was provided in the form of grids exemplifying Level characteristics in the tasks.

  Pieces of learners' work were received which were highlighted to demonstrate Level characteristics in the tasks.

  There were examples of clusters using code / number systems to identify examples on reading and written work.
- Summative commentary for the single attainment targets Oracy, Reading and Writing which clearly noted the levels under consideration, coming to a conclusion on the best-fit level, and the position within the level (at the top end / lower end / securely within).
- Rationale for the best-fit judgment (at the top end / lower end / securely within) for the subject level.

The report commended four clusters' profiles and two clusters' approach to methods of presentation.

3.1.8 WJEC's Review of the External Moderation Pilot, 2010

In reviewing the effectiveness of the pilot on 8 December 2010, WJEC observed that:

> "The unanimous feeling of the meeting was that it had [been worthwhile]; by extension, it was agreed that it is work which should be continued, as it provides:
>
> - an opportunity for a national overview of standards;
> - an important impetus for cluster meetings;
> - an opportunity to hold teacher assessment to account;
> - an opportunity to raise performance, both through the process itself, and through the sharing of best-practice at regional meetings;
> - a resulting improvement in task-setting;
> - reinforcement of the statutory requirements which will support the implementation of the new curriculum;
> - strengthening of the transition between primary and secondary school."

> (*Securing Key Stage 2 and 3 teacher assessment: Year 3 Interim Report* (September 2010 – January 2011, p. 2)

Importantly,

> "The pilot highlighted that the 2008 National Curriculum is not being consistently used for assessment in all core subjects. The pilot data shows that there is significant
>
> disagreement in some areas with the overall best-fit levels, which reinforces the requirement for effective external moderation." (*ibid*)

The group was also asked to consider how value may be added to the pilot model in any future roll-out. That discussion was summarised as follows:

- "it is the process of cluster moderation that is important; therefore the focus should be on this rather than the end product, and it should be documented through a log of each cluster's activities;
- the model could be reduced to one requiring only two learner profiles for each cluster, one from each key stage (Level 4 for KS2 and Level 5 for KS3). (However, this would not allow clusters to demonstrate consistency across the key stages);
- the value of requesting work at the top/lower end of the levels was discussed: some teams thought that requesting work that was within a level (i.e. not specifically top or lower) would make the process more manageable; however, the maths team queried the value of this as most clusters seem to be able to best-fit the general level correctly and it is the borderline profiles where further support is needed;
- reports need to include detailed guidance on how clusters can improve;
- regional meetings could provide more subject-specific advice on task-setting and examples of learners' work. It was also suggested that discussing hard copies of materials at such meetings is more valuable to schools than only providing materials electronically." (*ibid*, pp. 2−3)

Finally, the group considered three options addressing the question whether the process could be rolled out across all clusters in Wales for completion by 2012:

- Option 1 − the original plan to roll out over 3 years (2010−2013)
- Option 2 − a five-term plan to complete the process by August 2012
- Option 3 − a six-term plan to complete the process by December 2012

and reported that:

> "The group also considered further the draft five-term time-line (Option 2).There was general agreement that it was not easily workable as it stood; indeed one

> DCELLS officer suggested that ideally a five-year plan would be required. However, to ease the impact of having 2 moderation exercises in the same academic year, it was suggested that timetabling of the second round of moderation in the autumn term of 2012 (Option 3), would go some way to making the model more manageable, especially if the number of profiles required was reduced to two." (*ibid*, p.3)

The group also considered the concern of the Chief Moderator for English that:

> "We would have been stronger on [effective report generation] if issues to do with consistency of code application between Welsh/English had not taken considerable amounts of our time during the final day of moderation where we were undertaking QA with moderator pairs across their reports and cross-referencing between moderator pairs; and during our actual QA day. Inaction on this issue was particularly irritating as it wasn't a subject issue, rather an administration of agreed codes issue." (*Key Stages 2/3 Cluster Group External Moderation, Pilot 2010, English Chief Moderator Report* December 2010, p.6)

and commented:

> "During moderation and QA there was an issue concerning the application of codes for the language subjects. On reading some Welsh First Language reports the English CM noted that there were inconsistencies regarding the application of the codes for attainment targets and to the overall best-fit code. Welsh [moderators] applied the codes to each AT [Attainment Target] then looked separately at the overall best-fit level to see if there was sufficient agreement across the evidence, whereas English used the codes for the ATs to determine the overall best-fit code. After detailed discussion with both teams the inconsistency related more to the reason for the issue. It was apparent that the majority of issues for English were due to insufficiency of evidence whereas in Welsh the issues were predominantly disagreement with the level or the teacher commentary. Welsh [moderators] also noted that with hindsight their requirements for evidence in the guidance was not specific enough whereas English was more detailed, so they were unable to apply the code for insufficient evidence in the same way.
>
> For the future it is recommended that the language teams meet early in the process to agree a more consistent approach. Consideration should also be given to the value of coding both the attainment targets and the overall best-fit levels. It may also be useful to timetable the moderation of the two languages subjects together to enable the teams to consult on issues." (*Securing Key Stage 2 and 3 teacher assessment: Year 3 Interim Report* (September 2010 – January 2011) p. 3)

These agreements then formed the basis for the first full external moderation exercise which would take place towards the end of the 2011−12 school year. WJEC organised regional information meetings in Llandudno, Cardiff and Swansea with Cluster Group

Contacts in June 2011 to brief them on the arrangements for 2011−12. These were well attended and well received.

3.1.9 Arrangements for external moderation in 2011−12

The statutory framework for teacher assessment in 2011−12 reminded schools of the Welsh Government's expectations of teacher assessment:

- Accurate and consistent teacher assessment should be at the heart of each school's policy and practice to ensure a high-quality learning experience and effective assessment for all learners.
- All school leaders and individual teachers should see valid and reliable assessment as fundamental to the efficiency and effectiveness of both whole-school systems and best practice in all learning and teaching environments.
- The statutory requirements for the moderation of teacher assessment, including external moderation, are designed to support reliable teacher assessment and robust school-based assessment systems. (Welsh Government, *Statutory assessment arrangements for the school year 2011/12*, Guidance document No: 054/2011, September 2011, p. 3).

It required that headteachers must:

- Remind teachers of their contractual duty to administer the assessment arrangements
- Identify which learners should be assessed at the end of each key stage
- Ensure that teacher assessment levels are recorded for each attainment target in all subjects with more than one attainment target
- Ensure that overall subject levels are recorded for each subject. (*ibid*, p. 5)

Furthermore, headteachers must:

> "ensure that for English, Welsh first language or Welsh second language, mathematics and science (Key Stages 2 and 3), and for all non-core subjects (Key Stage 3 only)
>
> - robust systems and procedures are in place to support accurate and consistent teacher assessment. These systems and procedures need to be focused on internal standardisation and moderation
> - all teachers understand and apply the concept of best-fit judgments to learners' work, in relation to the national curriculum outcome/level descriptions." (*ibid*, p. 6)

and

> "ensure that for English, Welsh first language or Welsh second language,

> mathematics and science:
>
> - cluster group meetings for Key Stage 2 and 3 transition include robust arrangements for moderation of examples of Year 6 and Year 9 learner profiles selected from within the cluster group's own schools
> - these arrangements should add value to school-based standardisation and moderation by strengthening teacher assessment. They should also ensure that good practice within the cluster is identified, shared and built upon, to set an agenda for improvement that reflects local circumstances and needs." (*ibid*, p. 7)

Further guidance was given within the document. There was thus no doubt whatsoever of the Welsh Government's commitment to moderated teacher assessment of KS2 and KS3 core subjects and of its concern to ensure that all those involved in the assessment process should clearly understand the expectations placed upon them by Statute.

WJEC was again contracted to manage the process of external moderation. It did so against a strict schedule, selecting and appointing Assistant Moderators in September 2011, holding a planning meeting with Chief and Deputy Chief Moderators in February 2012, setting deadlines for the submission of sample evidence for Welsh and Welsh Second Language from clusters by the end of March so that these could be moderated in late April and for English by early May for moderation towards the end of May, Quality Assurance Reviews to take place in May (Welsh and Welsh Second Language) and early June (English) in order to return moderation reports to cluster contacts by the end of June, and 20th July (the end of term) as the deadline for appeals from clusters.

Guidance notes and forms were sent to clusters (and were more generally made available on the WJEC's website) as they had been during the previous (pilot) external moderation exercise.

WJEC notified all schools about the arrangements for Cluster Group Moderation at KS2 and KS3 in a document which clearly set out the arrangements and timescales for this. This document is to be found at http://www.wjec.co.uk/index.php?subject=30&level=111.

The document reminded schools that:

> "Following successful roll out of external moderation support for core and non-core subjects at Key Stage 3, a pilot involving cluster groups from the majority of local authorities was developed to extend external moderation support to Key Stage 2/3 school cluster groups in autumn 2010.
>
> In the 2011/12 school year this work will focus on securing an agreed and consistent approach in primary and secondary schools to the assessment of oracy, reading and writing. Each cluster will therefore be asked to provide evidence for two subjects, English and either Welsh First Language or Welsh Second Language. Clusters will select four learner profiles, Levels 4 and 5 identified from KS2, Year 6 and Levels 4

> and 5 from KS3, Year 9. WJEC will continue to manage the administrative arrangements on behalf of the Welsh Government."

It set out a clear timeline for action on the part of schools/clusters.

All clusters were required to participate in 2011−12 by providing learner profiles for two subjects − English and either Welsh or Welsh Second Language. Clusters were also asked to provide the moderators with details regarding the moderation process within their cluster. This requirement was introduced for the first time in 2011−12.

The guidance indicated that:

> "Each cluster group's learner profiles will be moderated by teams of external moderators. The moderators will focus on how the evidence/learner profiles reflect the clusters' overall best-fit judgments at Levels 4 and 5. The moderators' reports will be designed to provide clear, constructive comments, particularly on any issues highlighted."
>
> "Within each cluster group, all schools should play an active part in:
>
> - identifying learner profiles for the cluster group to use as a source of evidence,
> - attending cluster group meetings to moderate selected learner profiles,
> - agreeing the moderated outcomes and adopting the cluster's moderated learner profiles as benchmarks within their individual schools,
> - applying these benchmarks to future teacher assessment and in particular to end of KS2/3 teacher assessment.
>
> Profiles provided for external moderation need to reflect this practice. Cluster contacts have been asked to act as the point of contact between their own cluster and WJEC, who are responsible for managing the external moderation."

This guidance makes absolutely clear the inclusive nature of decision-making which is expected of schools within clusters and this expectation of *inclusivity* underpins all subsequent guidance within the document. The other key principle within the guidance is that of *sufficiency*.

> "The cluster group should satisfy itself that the evidence (learner's work, commentary, task information and stimulus/context) is 'sufficient' to communicate its shared understanding of national curriculum standards. If this is the case, it should follow that the information will also be sufficient for the external moderators' check and feedback.
>
> Each cluster group's learner profiles should represent the cluster's shared agreement on best-fit attainment at the end of Key Stages 2/3. For external moderation, the profiles will need to be selected from learners in Years 6 and 9

> (from either the 2010/2011 or 2011/2012 academic year).
>
> For external moderation, each cluster group is requested to select four learner profiles representing Levels 4 and 5:
>
> - One Level 4 learner profile identified from KS2, Year 6 and one Level 4 learner profile from KS3, Year 9:
> - One Level 5 learner profile identified from KS2, Year 6 and one Level 5 learner profile from KS3, Year 9."
>
> "As moderation will be feeding back on end of key stage best-fit attainment it is important that each learner profile contains a range of work from an individual learner only and not mixed /composite learners. It will not be possible to moderate profiles that contain the work of more than one learner.
>
> The learner profiles selected by the cluster should be identified from normal classroom practice and/or 'out of class' work undertaken by learners. In no sense is it expected that the selection of learner profiles should require additional work by learners or non-routine teaching or learning experiences.
>
> Each learner profile must also include agreed cluster group commentary identifying a level for each attainment target. Taken as a whole, the commentary should clearly direct teachers within the cluster and the external moderators to how the clusters' agreed best-fit judgment was arrived at. Therefore the cluster should ensure that it also makes appropriate reference to characteristics of adjacent levels shown in the profile as appropriate. These commentaries should be sufficiently detailed to convey the cluster's understanding of the level descriptions and agreed best-fit judgments.
>
> Cluster commentary can take a variety of forms including reference, for example to teachers' original annotations on their learners' work, bulleted points or other succinct notes. An optional proforma for cluster commentaries will be provided for each subject, and will be available on the WJEC website. Comments that already exist on learners' work or linked material do not need to be rewritten on the proforma. It will not be possible to moderate profiles received without cluster commentary or stimulus resource materials."

This guidance is clear and unambiguous. As noted, it is driven by the principles of *inclusivity* and *sufficiency*.

Further helpful subject-specific guidance is given for English, Welsh and Welsh Second Language and definitions of three key terms – "standardisation", "moderation" and "making best-fit judgments" are given, drawn from DCELLS/DfES guidance:

> **Standardisation**
>
> "a process of using samples of the work of the same learner or of different learners to enable teachers to reach agreement on levels of attainment by confirming a

> shared understanding of the characteristics of a level. ... Materials collated for standardisation purposes ... are described as the school, department or cluster standardisation portfolio which is used as a reference source of evidence". (*Ensuring consistency in teacher assessment: Guidance of (sic) Key Stage 2 and 3*)
>
> **Moderation**
>
> "Moderation at the end of a key stage, where a 'best-fit' judgment on an individual learner's level of attainment is made. … exemplified through 'a range of work of an individual learner, a learner profile ... to assist judgments to be made at the end of a key stage, through moderation". (*Ensuring consistency in teacher assessment: Guidance of (sic) Key Stage 2 and 3*)
>
> **Making 'best-fit' judgments**
>
> "... to recognise progress within a key stage, best-fit judgments could use the features of adjacent level descriptions to indicate whether a learner is working at the lower end, securely within, or at the top end of a broad National Curriculum Outcome/Level. Typically, a learner at the lower end of an outcome/level shows mainly characteristics of that outcome/level across a range of work, but may still have some characteristics of the previous outcome/level in some aspects of the work. A learner securely within the outcome/level demonstrates the characteristics of that outcome/level across a range of work. A learner at the top end of an outcome/level demonstrates clearly characteristics of that outcome/level across a range of work with some examples of characteristics of the next outcome/level." (*Making the most of assessment 7−14*)

This guidance was supported throughout 2011−12 by appropriate *proformas* setting out WJEC's requirements for profile submission (e.g. 13656 KS2−3 English Cluster Process Form and 13659 KS2−3 English Optional Commentary Sheet) and their equivalents for Welsh and Welsh Second Language.[8]

In addition, WJEC published a number of other documents on their website during 2011−12 including (for all three languages) the PowerPoint presentation on *Securing teacher assessment at Key Stage 2/3* which had been used at the Information Meetings for Cluster Contacts in Llandudno, Cardiff and Swansea in late June 2011 and the Chief Moderators' Reports from the 2010 pilot. Other documents available from that website included an Exemplar Tracking Sheet, an Optional Commentary Exemplar and a document on levelling in Welsh Second Language assessment, entitled *Strands*. For English assessment, there were documents giving Additional Moderation Guidance, Guidance on Exemplar Profiles, Using Process Drama as a response mode for Reading, Progress through Levels in English (similar to the *Strands* guidance for Welsh Second Language), a Pilot Feedback Grid and a Draft

[8] These were downloadable from WJEC's website until the end of the contract for the administration of the External Moderation process.

Cluster External Moderation Report, drawn from the 2010 pilot.

Taken together, these guidance documents served to put flesh on the National Curriculum policy and guidance documents and set out clearly the expectations which WJEC has of schools and clusters in preparing and presenting learners' profiles for external moderation.

3.1.10 Description of the current system

The following is a brief description of the current system that was examined in this investigation. It was drawn from the DfES publication: *Statutory assessment arrangements for the school year 2011/12*, issued in September 2011 (Guidance document No: 054/2011).

General

- Headteachers must ensure "robust systems and procedures are in place to support accurate and consistent teacher assessment. These systems and procedures need to be focused on internal standardisation & moderation"
- Headteachers must ensure "all teachers understand and apply the concept of best-fit judgments to learners' work, in relation to the national curriculum level descriptions (as defined in *Making the most of assessment 7–14*)
- This should allow teachers, within each subject, to confirm a shared understanding of national curriculum standards, based on an agreed selection of pupils' work and supporting teacher commentary that shows the links to the level descriptions" (p. 6).

Internal Standardisation

Headteachers must ensure that teachers "have in place arrangements by which teachers confirm & maintain a shared understanding of national curriculum standards, using samples of their learners' work to generate a reference set of exemplars" (p. 6) (i.e. the school must have standardisation procedures in place and construct standardisation portfolios to support these). This is required for all core subjects and Welsh Second Language in English medium schools.

Internal Moderation

Headteachers must ensure that teachers "have in place arrangements, using selected learner profiles, so that teachers moderate end of key stage assessments and apply the outcomes from this internal moderation prior to finalising all learners' end of key stage attainment" (p. 6). In other words teachers are required to moderate individual pupils' work (determine best-fit NC levels) for Year 6 (end of KS2) and Year 9 (end of KS3) and construct learner profiles to help in this process. This is required for all core subjects and Welsh Second Language in English medium schools and for all non-core subjects at KS3.

Schools are required to "maintain concise documentary evidence of these systems and procedure, and their annual application, for both internal and external quality assurance purposes" (p. 6).

Schools are required to "undertake annual reviews to ensure ongoing value added to existing arrangements and that procedures reflect best practice and direct ownership by all teachers" (p. 6).

Cluster Moderation

Headteachers must ensure that cluster group meetings for KS2 and KS3 transition include robust arrangements for moderation of examples of Year 6 and Year 9 learner profiles selected from within the cluster group's own schools etc. (p. 7)

In order to comply with these requirements, headteachers should ensure that:

- their own school representatives attend all cluster group moderation meetings
- they allow appropriate time for cluster group moderation meetings etc.
- their own teachers select learner profiles from their classes as evidence for the cluster group's moderation stage
- they support their teacher representatives to share the outcomes of cluster group meetings with other staff
- agreed decisions & outcomes from cluster group meetings are implemented by all relevant staff within their own school prior to end of key stage moderation.

This is required for all core subjects and Welsh Second Language in English medium schools.

External Cluster Moderation

The WJEC conducts an external moderation of teacher assessment for English, Welsh and Welsh Second Language within KS2 and KS3 cluster groups.

3.2 Reliability of the current system

The investigation sought to answer several questions that relate to the reliability of the teacher assessment system. These questions include:

1. What works in terms of securing accurate teacher assessments?
2. Do the assessments demonstrate reliability?
    a. What is the reliability and consistency of judgments made in the KS2 and KS3 assessment and levelling?
    b. How does the KS2 system that is focused on the school cluster level compare to the KS3 system?

This section draws upon the literature review and the responses to the four questionnaires sent to LAs, primary schools, secondary schools and cluster coordinators to answer the questions.

Note: The number of schools in the acquired sample is low (18 secondary schools and 48 primary schools). The percentages in the tables and graphs within Section 3.2 cannot be viewed as representative of all schools. Rather, the results should be taken as indicative only.

3.2.1 What works in terms of securing accurate teacher assessments?

Before making any judgments about reliability of the current teacher assessment system, it is important to understand what constitutes reliability in such systems and what factors contribute to higher levels of reliability.

From the literature review **reliability is generally defined in terms of consistency across assessments or "the extent to which assessment can be trusted to give consistent information"** (Chatterji, 2003; Crisp, 2010; Gipps, 1994; Harlen, 2004; and Mansell, James, & the Assessment Reform Group, 2009). The central idea is that the results of an assessment would be the same or similar if the procedure were to be repeated under equivalent conditions. Reliability is seen as a measure of confidence in the quality of the assessments. Psychometric constructions of reliability require a significant degree of standardisation; that is, that any given assessment is the same for all candidates (students) in form, content and conditions (Rust, 2004). In school-based standards-referenced assessment, reliability is related to consistency of teacher judgments and the comparability of reported results. In this context, then, moderation can been conceived as a system's response to the reliability imperative.

Reliability can be treated as a measurable construct that can be expressed numerically in terms such as correlation coefficients (Salvia & Ysseldyke, 1998). For these processes to occur tests and examinations are standardised not only in terms of the conditions of implementation but also in statistical terms. This type of standardisation is not possible or desirable in school-based assessment where classroom teachers develop and implement a range of different types of assessments in accordance with their particular context(s) and are required to develop and maintain a shared understanding of standards.

Despite this difference the fact remains that that there must be some measure of understanding about consistency of teacher judgments and comparability of reported results if the results of school-based assessment are seen to be reliable in the sense of being dependable. This last point about consistency of teacher judgment is often made with the assumption that when the need for reliability is paramount, such as is in high-stakes circumstances of certification and selection, the obvious or default position is to turn to testing. The statistical approaches common to testing focus specifically on reliability. Such approaches are seen to be more trustworthy than teacher judgment, which is assumed to be inherently unreliable and difficult to improve and sustain (Harlen, 2004). Nevertheless, as Maxwell observes: "Each can deliver different benefits – especially, greater validity from school-based assessments and greater reliability from external assessments – though these benefits are possibilities rather than certainties" (Maxwell, 2006, p.3).

According to O'Brien (1998) there are actually two paradigms operating: one is measuring how much of a certain quality (single underlying dimension) is evidenced in student responses; and the other is judging what the evidence says about what the student has learnt and how well. The psychometric model that "observed score = true score + error" suits notions of reliability and validity for multiple-choice testing. Assumptions of the true-score model do not readily suit notions of reliability and validity for testing in open-ended response modes and do not at all suit notions of validity and reliability for the school-based assessment. Here, assumptions of the true-score model do not hold; in particular, assumptions about infinite populations, about markers, items and tasks being sampled at random from a universe of markers, items and tasks, and about identical and independent Gaussian distributions. In many school settings, split-half reliability estimates are not possible and the practice of inter-rater agreement studies is beyond the resources of most schools. According to Moss (1992, 1994) the "epistemological and ethical purposes served by reliability" can be broadened to include the practice of contextualised judgment. One of her three warrants for reliability is the privileging of contextualised (teacher) judgments. This involves the use of a standards schema, criteria/standards matrix, grid or grading master[9] that teachers apply for making decisions about the standard or level of student work.

A recurring theme in the literature on teacher assessment is the level of expertise required of teacher–assessors. The skill sets that can be identified as necessary for a reliable system are (a) expertise in their disciplines, (b) total immersion in their level statements, (c) the ability to evaluate, and (d) the ability to negotiate when seeking consensus at moderation meetings.

[9] A variant of the traditional criteria/standards matrix, the grading master contains features necessary to support the nature of complex multi-faceted tasks

The system in operation in Wales is a social or consensus moderation scheme, which is, according to Linn (1993), the most common form of moderation for school-based assessment systems. Social or consensus moderation is a quality assurance process that brings teachers together to review and discuss judgments across examples of student work, often in different assessments, and reach some level of agreement about the application of standards to that work. In this scheme, the interpretation and application of the standards needs to be consistent within and across school sites to ensure that work of comparable quality will be awarded the same or similar grades. Matters (2006, p. 2) describes moderation as "a set of processes designed to ensure that standards are applied consistently across teacher–assessors and across schools". Linn (1993) notes the emphasis on "collegial support and the movement toward consensus judgments through social interaction and staff development" (Linn, 1993, p. 99). Moderation should be a continually improving system that reflects changes and developments in curriculum, teaching and system requirements and provides valuable feedback to teachers, learners and schools. It is "purported to promote comparability and equity in application of standards" (Hay & MacDonald, 2008, p. 157).

Moderation is less about precision such as might be expected in a 100-point scale and more about broad generalisations that hold true for a system of grades such as seven grades common in many tertiary courses or the five levels of achievement in many secondary school systems. In simple terms, moderation is about reliability in the sense of being a "guarantor of fairness, safeguarding against the possibilities of subjectivity and bias" (Pitman, O'Brien, & McCollow, 1999, p. 2). In some cases, moderation may also provide feedback about quality of assessments. At its best, it can assure both validity and reliability.

A key component of most moderation systems is the emphasis on shared understanding of standards and evidence in student work. When teachers meet to review and discuss grades awarded they usually do so in reference to specific examples of student work tendered to represent grades awarded. Klenowski (2005) takes this step further and argues that "confidence in teacher judgments require consideration of actual student work and the grades that were awarded to students on the basis of that work" (Klenowski, 2005, p. 3). The centrality of the relationship between grades awarded and evidence in student work is at its clearest in the processes of moderation.

However, even in a mature teacher assessment system like the one in Queensland, Australia, there can be inconsistencies in teacher judgments. In 2008, Hay and MacDonald conducted a research project in Queensland involving semi-structured interviews and participant observations into how teachers of Health & Physical Education (HPE) made judgments about students' levels of achievement. They noted that teachers make progressive judgments that are informed more by their internalised understandings of course standards, understandings that do not necessarily align with the intention of the syllabus. Although this was a small study, its conclusion raises concern about the trustworthiness of the judgments and calls for more research about the extent to which the processes identified for HPE also occur in other subjects (Hay & MacDonald, 2008).

The review of literature conducted for this investigation shows that there are a number of factors that can influence how consistent the judgments are in a teacher assessment system. However, the experience of Queensland, Australia demonstrates that very acceptable levels of reliability can be achieved in such systems. In Queensland a study re-assessed a sample of 546 student folios in English, Chemistry, Mathematics I and Modern History drawn from the 1992 Year 12 cohort. The folios had already been reviewed as part of routine moderation procedures.

For the Queensland study, the folios were initially assessed six times by expert teacher–assessors (teachers who were experienced members of moderation panels). These assessments were made without prior knowledge of the original levels of achievement awarded. The inter-marker reliability of 0.94 (Masters & McBryde, 1994, p. vi) indicates a level of agreement higher than that recorded for independent assessment of external examinations. It would seem, therefore, that when teacher judgment is an element of a fully developed and centrally supported assessment regime, it can provide comparability of reported results at least equivalent to that of other systems.

3.2.2 Do the assessments demonstrate reliability?

This section addresses the reliability and consistency of judgments made in the KS2 and KS3 assessment and levelling and how the KS2 system that is focused on the school cluster level compares to the KS3 system.

It was beyond the scope of this investigation to perform a full-scale measurement and analysis of the reliability of the judgments in the current system, but measures of the confidence of the primary-school headteachers, secondary-school headteachers, local authority advisers, and cluster coordinators were obtained from the extensive questionnaires sent to each sector. Of the total 64 questions, there were four questions that were common to all four questionnaires about their levels of confidence in the accuracy and reliability of (a) existing pupil assessment records, (b) internal standardisation procedures, (c) internal moderation procedures, and (d) cluster moderation procedures. Respondents were asked to state whether they had “no confidence”, “some confidence” or “complete confidence”. In addition to the questionnaires, focus group interviews were conducted with a sample of ten clusters; issues of reliability of teacher judgments emerged in those conversations.

The views of each of the sectors (primary schools, secondary schools, local authorities, and clusters) are now described and followed by a comparison of those views.

3.2.2.1 Views of the primary schools

Forty-eight primary schools of the 219 who received the questionnaire responded to the questionnaire and a summary of responses to the four questions about reliability is presented in Table 4. It should be noted that, because of a low response rate, these views can only be considered indicative of the primary schools as a whole, and not representative.

Table 4: Primary school ratings of reliability of the system

Overall, how much confidence do you have in the accuracy and reliability of your existing …?	No confidence	Some confidence	Complete confidence
Pupil assessment records	–	18 (37%)	30 (63%)
Internal standardisation procedures	–	19 (41%)	27 (59%)
Internal moderation procedures	–	21 (45%)	26 (55%)
Cluster moderation procedures	–	20 (44%)	25 (56%)

Overall, the majority of primary schools that responded to the survey expressed complete confidence in the accuracy and reliability of the assessment records, in the internal standardisation and moderation procedures, and in the cluster moderation procedures, while the remainder had some confidence. No primary school's response was of no confidence in these components; however, it should be noted that those schools that responded to the questionnaire might hold different views from those who chose not to respond.

The interviews with clusters revealed that a wide range of pupil tracking systems were in use although some schools were now adopting the Incerts commercial software and existing users commented favourably on it. Not having common systems across the clusters makes transfer of data from KS2 to KS3 more challenging, although one cluster where there were different systems at primary and secondary schools thought that data transfer was only a small problem.

Some primary schools were confident in their reliability and made comments such as "We are very confident on the thoroughness and reliability of the school's procedures and accuracy in assessing pupils' progress and attainment," but others felt that it was a challenge to achieve desired levels of reliability as explained by one school in the statement, "We need to explore the accuracy of TAs and target setting further, be more effective in analysing data − to ensure we are accurate, yet set high expectations." There was also concern that, while a school might be confident in its own reliability within the school, that reliability between schools was more difficult to achieve, as exemplified in the comment in one school's questionnaire that: "Clearly, over the years, achieving accurate and reliable assessments has been problematic. There has been no consistency between schools at a local, county or national level" In relation to accuracy of internal standardisation, one school commented:

> "Although we as a school are quite confident that we are thorough in our assessments and ensure that results of standardised tests and results of software such as RM maths is also used when levelling work, I am sometimes shocked when talking to other teachers from other schools when they comment that a certain percentage of their class have achieved Level 5. I also believe that there is confusion

> between working at a level and achieving a level."

Several schools commented in the questionnaire on the fact that, although they did not have full confidence in the reliability and accuracy of their internal standardisation procedures, they were taking steps to improve. For example, one school said, "We are in the process of developing our internal standardisation procedures in line with WG year 6 portfolio developed in Welsh Second Language and English in 2011–2012. This is an area that needs to be developed in the school."

Another school reported:

> "It has become evident from answering the questions above that the school needs to review its current procedures in relation to standardisation in order to ensure accuracy and reliability of our assessments. And the school needs to develop its own standardisation portfolios (Now use cluster portfolios) in Oracy, Reading and Welsh."

3.2.2.2 Views of the Secondary Schools

Eighteen of the secondary schools sampled responded to the questionnaire and a summary of responses to those questions about reliability is presented in Table 5.

Table 5: Secondary school ratings of reliability of the system

Overall, how much confidence do you have in the accuracy and reliability of your existing …?	No confidence	Some confidence	Complete confidence
Pupil assessment records	–	9 (50%)	9 (50%)
Internal standardisation procedures	–	10 (56%)	8 (44%)
Internal moderation procedures	–	9 (50%)	9 (50%)
Cluster moderation procedures	–	10 (56%)	8 (44%)

Overall, the majority of secondary schools that responded to the survey have some confidence in the accuracy and reliability of the assessment records; in the internal standardisation and moderation procedures; and in the cluster moderation procedures, with the remainder expressing complete confidence. No secondary school felt that they had no confidence in these components.

Again, there was comment from the secondary schools about the need to have compatible pupil tracking systems at KS2 and KS3 to enable pupil progress to be effectively managed across the cluster.

From the comments made on the questionnaires and in the cluster focus groups, it became evident that a reason that some secondary schools reported that they had some confidence rather than complete confidence in the accuracy and reliability across these factors was because they perceived differences across the departments in the school. One school commented that the:

> “English and Welsh Departments and a number of non-core subjects have greater degree of confidence in the accuracy and reliability. Science Department is developing its systems following a recent staffing change.”

Another school explained it like this:

> “Much more support is available to the Welsh and English languages and therefore less support is given to Mathematics and Science teachers in the core subjects in setting levels and standardisation work. There is no appropriate support provided for non-core subjects.”

This pattern, in which mathematics and science departments were seen to be lagging behind the English and Welsh departments, was observed in the cluster focus groups as well. For example, one school reported that learner profiles and standardisation profiles were available for all subjects apart from science, although the issue was being addressed with the science department as a priority matter. Another cluster said that although cluster moderation had taken place for both languages, it had not occurred in mathematics and science, and there were no cluster portfolios for those subjects, although the issue was being addressed. Another cluster reported that internal standardisation was applied to all core and non-core subjects in the secondary school apart from mathematics, owing to some departmental issues that were receiving attention.

3.2.2.3 Views of the Local Authorities

Eighteen of the 22 LAs responded to the questionnaire and a summary of responses to the four questions on reliability is presented in Table 6.

Table 6: Local authority ratings of reliability of the system

Overall, how much confidence do you have in the accuracy and reliability of your existing …?	No confidence	Some confidence	Complete confidence
Pupil assessment records	–	15 (83%)	3 (17%)
Internal standardisation procedures	–	15 (83%)	3 (17%)
Internal moderation procedures	–	15 (83%)	3 (17%)
Cluster moderation procedures	–	14 (78%)	3 (17%)

Overall, the majority of LAs expressed only some confidence in the accuracy and reliability of the (a) existing pupil assessment records, (b) internal standardisation procedures, (c) internal moderation procedures, and (d) cluster moderation procedures.

Only three of the 18 LAs reported that they have complete confidence in the accuracy and reliability of existing pupil assessment records, whilst the 15 remaining have "some" confidence.

Examples of extended responses to these judgments are given below:

> "Our schools were very successful in comparison with the National picture for English during the 2012 External Moderation Exercise at both key stages. Our KS3 Level 5+ is the highest in Wales in 2012 and based on reliable local practice that has been consistently supported by our advisory team in light of the national guidance noted above. Advisers from [our] team have led within National Assessment practice. During the pilot external moderation for core subjects at KS2−3 in 2010 [our] Science/Pedagogy Adviser and English Adviser were Chief Moderators for their Subjects. The English Adviser was also Chief Moderator for the 2012 Exercise. As a result, much of the national practice for the external moderation exercise was influenced by best practice in leading cluster, school and department moderation…. Likewise the verification process at KS3 was led by [our] Pedagogy Adviser. As a result of this and our work with SMTs at Secondary level and Cluster work schools have been very well informed about the summative and formative assessment processes following the guidance above. Judgments are based on specific evidence from normal classroom practice and involve holistic assessment of learners' skills across attainment targets."

> "The LA has worked in conjunction with its advisory service (ESIS) to deliver training on assessment and used the advisory service to moderate cluster meetings. The advisory service also supported schools when developing the profiles for WJEC. Despite this detailed support it is evident from data and working with schools that assessment processes are not always rigorous and with the LA collection of reading and maths scores this year it is evident that many schools have not used this information to support their assessment and in a number of cases pupils are being assessed at a Level 4 when their Reading scores are below a functional reading age."

> "A good number of the KS2 teachers can accurately assess standards and follow a comprehensive training provided at local level. However, there is much work to be done to develop the learning to ensure that each pupil's work reflects the range and skills curriculum."

> "Without being able to have first-hand knowledge of every learner profile for all the core subjects it would be impossible to have 'complete confidence' but it's fair to say that we have a high degree of confidence in teacher assessment at KS2 and some confidence in teacher assessment at KS3."

All but one of the 18 LAs who responded to the questionnaire agreed that they ensure that schools' standardisation portfolios have appropriate commentaries linking the evidence (i.e. samples of pupils' work) to specific level descriptions but only three had complete confidence in the accuracy and reliability of its existing internal standardisation procedures – the remainder had some confidence. Detailed comments included the following:

> "Our work with clusters on standardising Welsh and English occurred as precursors to moderation. Schools have built on this guidance in their schools as a result and

> standardisation processes have been monitored this year through Summer Term LA Monitoring and through Primary Team Sample Moderation. Once schools and clusters have established their portfolios (we did this before the 2010 external cluster moderation!) we expect them to update them with more recent examples. Our main focus has been on Moderation as this is the process that ensures robust end of key stage judgments. The most useful part of standardisation is the inclusion of evidence from other year groups (not just years 6 and 9) and all staff teaching the subject. Teachers know how to identify level characteristics in pieces of work (i.e. standardisation) our main work with schools and clusters has been to ensure that they move on from this to moderate i.e. to come to a best-fit judgment based on an appropriate body of evidence across the attainment target(s) e.g. Individual and Collaborative Oracy evidence, Literary and Non-Literary Reading Responses, Literary and Non-Literary Writing (English/Welsh)."

> "Teachers move year groups, change role, move on, etc. Some have a tendency to be harsh, others lenient when focus [is] on perhaps one or two individual pupils. Small schools find making comparisons difficult. Dialogue [is] greater within larger schools. Very time consuming for teachers, taking them away from AfL and lesson preparation, especially in small schools."

> "Over time the Local Authority has ensured that schools have compiled standardised portfolios as part of their cluster work. In some cases these may not have been kept up to date. On occasions we find portfolios lack an appropriate commentary."

> "Advisers for the core subjects provide support, guidance and advice for coordinators in the developing of standardisation portfolios but do not check that every school has one. They also support those schools who ask for help in this area. Therefore when advice is given, the points in Q26 and Q27 [evidence and level descriptions] are included. We give advice, rather than ensure, because the standardisation, although a necessary part of the process, is non-statutory."

When asked about how confident they were in the accuracy and reliability of their schools' existing internal moderation procedures, only three had complete confidence, while the remaining 15 had some confidence. LAs gave similar ratings about the reliability of the cluster moderation process. Three LAs had complete confidence in the accuracy and reliability of existing cluster moderation procedures, one LA did not answer the question, and the remaining 14 expressed some confidence.

3.2.2.4 Views of the Clusters

A total of 17 of the 33 cluster coordinators responded to the questionnaire that was sent to them. Eight out of them expressed complete confidence in the accuracy and reliability of their existing moderation procedures, while nine had some confidence in the procedures.

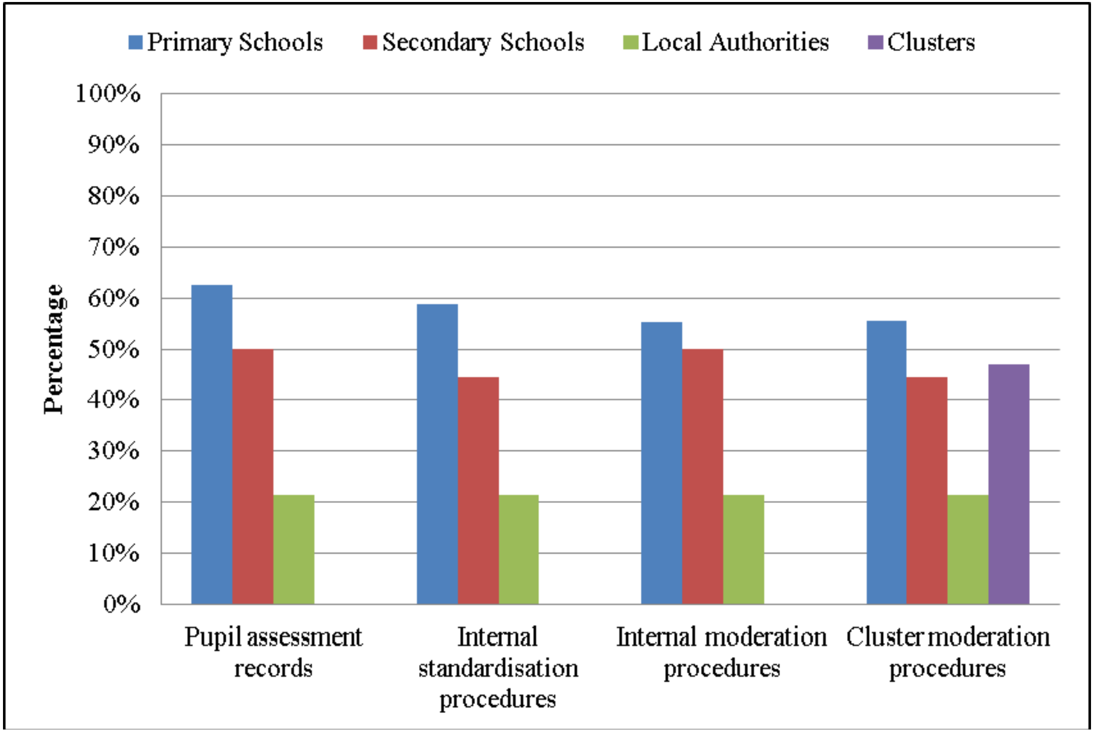
In the cluster focus group interviews, a persistent theme was the reliability of the 'best-fit' decisions. The concern was that the current policy does not define a common agreement on what percentage of a pupil’s work needs to demonstrate a particular level before that level is awarded. In one interview, some participants suggested the work had to be 100% at the level, while others thought that 60–75% of the work at a particular level was sufficient to gain that level. It was felt that, under the current scheme in which there is a wide range within a level, there was a wide interpretation by teachers when assessing pupil outcomes. Several cluster groups thought this was a cause for concern nationally as no guidance had been given on best-fit judgments.

Another concern around best-fit judgments was that schools had standardised assessment results that they used in making best-fit judgments, but the amount of emphasis or weight assigned in this judgment varied considerably. There was also some concern about whether or not teachers fully understood the level characteristics when making best-fit judgments. Overall, several clusters felt that further guidance on best-fit judgments was needed.

3.2.2.5 Comparison of views about reliability across the system

Figure 1 compares the number of primary schools, secondary schools, LAs and cluster coordinators who expressed complete confidence in the accuracy and reliability of (a) existing pupil assessment records, (b) internal standardisation procedures, (c) internal moderation procedures, and (d) cluster moderation procedures.

Figure 1: Percentage of questionnaire respondents with complete confidence in the accuracy and reliability of existing records and procedures



The pattern that emerges is one where the proportion of respondents expressing *complete* confidence in the accuracy and reliability of all four components of the teacher assessment system is highest for primary schools and lowest for local authorities. The proportions of clusters who expressed complete confidence in cluster moderation was higher than the levels expressed by secondary schools, but lower than primary school levels of confidence. The proportion of clusters expressing complete confidence in the accuracy and reliability of the components is the same in all four components. Proportions of completely confident cluster coordinators and secondary schools are similar.

The observed downward trend (from primary schools to secondary schools to local authorities) in level of confidence in all four features of the assessment system that were investigated parallels the levels of experience in assessment of the three categories of respondents. Brown, Lake, and Matters (2011) reported differences between primary and secondary teachers' conception of assessment in New Zealand and Queensland while Borko, Mayfield, Marion, Flexer, and Cumbo (1997) assert that teachers' perceptions of assessment are affected by the ideas teachers have about educational artefacts such as teacher efficacy. Primary teachers' self-efficacy could be inflated due to a lack of prior engagement with more formalised classroom assessment whereas LAs' evaluation of teacher efficacy could be more realistic due to experience with teachers and assessment.

Unfortunately, since no empirical data are available on the actual reliability of the judgments being made in the school year 2011−2012, it is not possible to verify the confidence of any of the sectors.

A part of the current system to ensure reliability of the teacher assessment judgments is the running of external moderation sessions by WJEC. A pilot external moderation programme for English took place in November 2010 and the Chief Moderator's report from that session clearly identified strengths and weaknesses of both the cluster submissions and of the procedures employed within the pilot programme. In that pilot, the range of submissions to which moderators were unable to agree to the indicated NC level varied but comprised around 50% of the 19 cluster submissions. However, some of these disagreements were due to 'technical' issues rather than the assessment of pupils' work per se. The information from the external moderation indicates that the reliability of judgments is low but there is no other empirical data to verify this.

Recommendation 1 – Reliability

The current system does not systematically gather data on the reliability of the teachers' assessments within clusters. External moderation of assessments depends upon the choice of portfolios submitted by clusters. Until methods by which reliability can be measured are available, it is impossible to know how well the system is doing in regards to the consistency of teacher judgment. It is recommended that measures of the actual reliability of judgments be embedded into the teacher assessment systems so that it can be tracked and, if necessary, the implementation of the system can be refined to optimise the reliability.

The following section discusses some of the methods that similar teacher assessment systems have used to establish and monitor reliability.

3.2.3 What do other systems do to establish and monitor reliability?

A comment, which is relevant here but applies in the validity section also, is that in the light of new forms of assessment (e.g., teacher assessment, assessment of higher-order thinking skills, assessment using a variety of instruments, assessment tasks that emulate the kind of process-based tasks thought to represent good practice) a conundrum exists: how to maximise validity and reliability simultaneously. Increasing validity leads to increasing reliability but only to a certain limit where a further increase in validity leads to a plateau followed by a decrease in reliability. Although based on technical considerations, it is a policy decision for systems as to the balance of mechanisms used for establishing reliability and validity.

The description that follows of mechanisms used to establish and monitor reliability in other systems does not assume that these mechanisms or variations thereof are not used or have not been considered in Wales. Nor does it assume that all are appropriate for teacher assessment at KS2 and KS3 given the less-than-high-stakes nature of the assessment programme.

3.2.3.1 Establishing Reliability

Immersion

Immersion is a method for establishing reliability that uses student work as the basis for shared understanding of the standards or levels. It usually involves asking teachers to attend training sessions to become familiar with exemplars of student work at a particular level as a precursor to reaching consensus about grades awarded to selected samples of "real" student work at moderation meetings. A slight but not insignificant twist to this common model of moderation seeks less to reach a consensus about grades awarded on selected samples of student work than it seeks to emphasise teachers' selection and sharing of examples of the range or ways for students to achieve the level. In such an "immersion" process it is the teachers who locate instances of the desirable features of work at a particular level rather than teachers being given examples of student work at that level. An advantage of this approach is that it addresses some of the issues arising from the traditional and comparatively passive approach whereby teachers have to confront exemplars for which trading off has already occurred, which is especially problematic for the middle levels of the marking rubrics where the distinctions between levels can be more subtle and harder to judge.

Marking rubrics

One of the ways to improve reliability in the marking process is to use marking rubrics. Marking rubrics are descriptive marking schemes developed by teachers or other assessors "to guide the analysis of the products or processes of students' efforts" (Brookhart, 1999). Marking rubrics are typically employed when a judgment of quality of student work is required. Marking rubrics are not mere checklists (for on/off decisions as in competencybased assessment or in yes/no non-discountable criteria). They are based on descriptive scales and support the judgment of the extent to which criteria have been met. In some places they are called marking schemes, marking guides, criteria sheets and so on.

Theoretically, marking rubrics can be used for any year level and to assess a broad range of subjects and activities. They have been used in the assessment of extended writing, group activities, extended projects and oral presentations in the USA and elsewhere. In Queensland, marking rubrics have been used in subject-specific assessments at school level and in the assessment of QCS[10] test items (short response and extended writing). Presumably, being upfront about the criteria and their relative importance would stop one marker from weighing heavily on linguistic structure while the other was more interested in the persuasiveness of the argument in the test of written expression. Marking of the writing task proceeds on the basis that good writing would likely have a combination of these and other factors.

The marking in the US and Queensland examples is holistic; that is, the marking rubrics "support broader judgments concerning the quality of the process or product; the criteria are considered together on a single descriptive scale" (Brookhart, 1999), compared with analytic marking rubrics which "allow for the separate assessment of each of several (two or more criteria); each criterion is scored on a different descriptive scale" (Brookhart, 1999).

A perceived problem with holistic (or impression) marking is that markers may differ considerably in their private marking schemes and so different markers could give very different grades to a given piece of student work. If suitably interrogated, an impression marker should be able to give some account of why marks were assigned as they were. S/he will be following some sort of private marking scheme with associated weightings or internalised trade-off rules and priorities. Analytic marking is where some sort of marking scheme is employed that gives guidance to the marker about what features s/he should look for and what weighting should be given to them (or what influence each feature's mark should have on the overall mark).

Choosing an analytic marking rubric does not eliminate the possibility of an holistic factor. An holistic judgment may be built into an analytic scoring rubric as one of the score categories. One difficulty with this approach is that overlap between the criteria set for the holistic judgment and the other properties being assessed cannot be avoided. When one of the purposes of the assessment is to assign a grade, this overlap should be carefully considered and controlled. The teacher–assessor (or syllabus writer) should determine whether the overlap results in certain criteria having more influence than was originally intended. In other words, the teacher–assessor needs to be careful that the student is not unintentionally penalised severely for underperformance on one dimension (or vice versa).

Sometimes it is impossible to separate a judgment into independent properties. When there

[10] Queensland Core Skills, actually a test of the common elements of the senior curriculum

is an overlap between the criteria identified for the assessment, an holistic scoring rubric may be preferable to an analytic scoring rubric. Either way, the challenge of making on-balance judgments is a constant theme in the literature about teacher assessment, not only in assigning a code to piece of student work such as a response to an individual task (e.g. B, 60/100) but also in the process of assigning a summative grade to a collection/folio of student work, say a portfolio of results obtained over a course of study. Grade is often used to mean the single result for reporting (i.e., after combining/aggregating results of several assessments).

The arguments above focus on the process of marking a single piece of work. They can be extrapolated to grading (levelling) (i.e., summative assessment) based on a profile of results.

An Italian study (Canal, Bonini, Micciolo, &Tentori, 2012) that showed that grading of a collection of student work is more complicated than marking a single piece of work provides an insight as to why it is harder to get agreement between teachers in “on-balance” judgments. Grading a collection of student work (say a folio or portfolio for summative assessment) can be more difficult than marking a single piece of work because judgment calls must be made. In theory, five different teacher–assessors could give five different grades to an individual student’s folio. Canal et al. describe the relative ease that teacher–assessors have with making judgments about the performance of typical students compared with the performance of atypical students, thus providing insights into the process of making “on-balance” judgments as well as the application of analytic marking. In this study, teacher–assessors evaluated the relative weights of five assessment dimensions for deciding on an overall grade, and then graded the work of students whose performance on all five dimensions was similar, and of students whose performance on at least one dimension was poor but good or excellent on the others. While it could be shown that teachers had a shared understanding of the relative importance of the five dimensions for grading, they did not make consistent overall judgments for the atypical students (i.e., where trade-offs were required). In fact, their judgments were highly inconsistent. It would seem that choosing an analytic marking rubric does not eliminate the possibility of an holistic factor. Despite having built-in rules for trading off inconsistent performances, teacher–assessors appear to personalise their prioritising of dimensions.

3.2.3.2 Monitoring Reliability

Marker monitoring for tests of written expression

The literature review identified ways that other systems have implemented processes that help to measure the level of reliability of the assessment. Two education researchers, Allen (1987) and Harris (2000) respectively, noted the difficulty that teachers had in marking a writing task and, separately, of a paper made up of short-response items (both open and closed-ended), which led them to implement two rigorous quality control measures – marker monitoring and check marking.

In the case of marking extended writing, the function of quality control is to identify discrepant markers and aspects of the application of marking schemes that need attention.

This marker monitoring model is based on the assumption that even good markers disagree at times for a variety of reasons, such as a momentary loss of concentration or difficulty in judging a borderline response. It follows that an acceptable marker is one whose marking differences are nearly always small or apparently random whereas a "discrepant" marker is one whose marking differences are either not acceptably small or not apparently random. The marker monitoring process involves the comparison of many different pairings of markers on the particular responses they have both marked. This marker monitoring process identifies those folders (random collections of responses of over 30,000 students) that have received significantly different marks from different markers (dissonant markings) and to identify those markers who are involved in a number of dissonant markings or who show other inconsistencies. The discrepant markers are retrained in the application of the marking criteria to get them better calibrated and the responses associated with dissonant markings are re-marked.

Marker monitoring for constructed-response items

Another method for monitoring reliability of teacher judgments in the marking of constructed responses[11], is to check if the differences between the grades assigned to a student's response to an item by a pair of markers are within the tolerance or random. Some constructed-response items encourage a wider variation in responses than do others, so the degree of discrepancy can be expected to vary from item to item, which is reflected in tolerances set for each item, indicating the difference in grades awarded that is acceptable (e.g., for some items the tolerance might be ± one grade). The tolerance level for each item is set with due regards for grade range for the item, the nature of the item, and the marking scheme.

Where the grades awarded by a pair of markers (working independently) differ on a set of items (the collection in a folder) in a way that is either not nearly always within the tolerance or not apparently random, one or both of the markers is/are deemed to be discrepant. Identification of the discrepant marker is made by examining the differences (size and apparent randomness) between a marker's grades and those of other markers of the same responses. The discrepant marker is the one of the pair who, on average, is noticeably more discrepant on these other occasions. (If there is no obvious difference between the two when assessed by this method, then both must be regarded as needing to reflect upon their practices.) Discrepant markers are then check-marked (and possibly recalibrated). As well as being confirmatory when used to follow up on the outcomes of quality control, check marking can be routine; that is, folders that have been marked are checked at random. Check marking is especially useful in determining whether a particular marker is the one who is the more responsible for the dissonant markings of a folder, or whether any disagreement in the marks awarded can be attributed to the open-ended nature of an item rather than to a marker's having a specific difficulty.

[11] Question format in which the test-taker supplies his/her own answer (response) as opposed to selecting from a list of answer choices provided

Inter-marker agreement

Inter-marker agreement, a statistical measure (weighted kappa) of the degree of agreement between two marking judgments, has also been used as tool to monitor reliability in both written expression and short-response. According to Agresti (1990), it is more important to estimate strength of agreement using the weighted kappa statistic but both measures are described below. Concordance is reported as the pairs of markings that have perfect agreement as a percentage of the total number of pairs of markings. Weighted kappa is the probability of a variable being classified in the same category by two different markers, and is zero when the agreement equals that expected by chance and 1.0 when there is perfect agreement. It is implausible for marker agreement to be no better than what is expected by chance. And this is the basis of using the kappa statistic for making decisions about marker quality.

Paired comparisons

In the method of paired comparisons (David, 1988), objects are presented in pairs to one or more judges. This method of monitoring reliability is used primarily in cases when the objects to be compared can be judged only subjectively; that is to say, when it is impossible or impracticable to make relevant measurements in order to decide which of two objects is preferable. The model has been applied in the field of educational assessment in recent years (for example, Greatorex, 2010; Harris, Kelly & Matters, 2004; Pollitt, 2004) by researchers in the UK and Australia. In this approach, judges make multiple pair-wise comparisons of student work. The analysis of these comparisons provides scores that are used to rank order the students after which cut-scores can be set so that students can be assigned to grades if required. Whether or not the method finds a place in educational assessment, it is original (even radical) and attests to the appeal of finding new ways of doing things in assessment.

Post-hoc consistency checks

Random sampling, which has been used in the Queensland system to monitor reliability, is designed to provide feedback and research data about consistency across different district review panels that operate as a part of the Queensland system, occurs at the beginning of the school year following the completion of all processes for certification the year before. Sample folios of student work are selected and distributed to panel members who undertake a review in much the same way as they would as part of the routine quality assurance processes. Results of this research are reported as a percentage agreement and supported by analysis and discussion of implications for schools and the system. One of the stated aims of random sampling is to evaluate the quality of assessment judgments in response to the question: How consistently do teachers apply state-wide [or country-wide] standards in determining students' levels of achievement in various subjects? It is therefore a rare example of long-term research into the consistency of teacher judgments.

An example drawn for the 2009 report on the results of random sampling indicated that there was "… substantial agreement between panels and schools: 84% of the folios were placed in the same level of achievement by both the random sampling panel and the school; 83% differed by no more than one-third of a level of achievement" (Queensland Studies Authority (QSA), 2009: 5). The magnitude of the level of agreement is much the same each year. Such levels of agreements are generally regarded to be high although it should be noted that the placement decisions relate to five broad categories.

3.2.3.3 Summary of what other systems do to establish and monitor reliability

Establishing reliability and monitoring reliability are discussed separately in what follows.

Establishing reliability

There are two common practices for establishing reliability of teacher assessments. One practice involves the use of commonly applied standards for marking/grading through, for example, marking rubrics or criteria sheets. Another practice for establishing reliability is immersion (which has a variety of names across systems). Immersion takes place in training sessions using student work as the basis for shared understanding of the standards or levels, both in their description and their application. Both of these practices are desirable in teacher assessment systems.

Monitoring reliability

The most common practice for monitoring reliability is through measuring inter-marker agreement. This assumes that assessment instruments are double-marked, which is not realistic at the item level in a teacher assessment system. It is however, applicable at the profile or folio level to check best-fit or overall judgments. Based on the same assumption about double marking, random sampling after a cohort of pupils has been assessed provides a way or monitoring reliability of the assessment system over time but not of the assessments currently being made.

Marker monitoring in external standardised tests of written expression and constructed-response items typically looks at agreement between two markers. The assumption here is that when two markers agree the assessment is accurate; it could be however, that both markers are incorrect even though they agree. A more sophisticated method for monitoring reliability looks at the marker rather than the mark. It monitors markers over multiple markings noting unusual marking patterns. Markers thus identified are then re-trained in the use of marking schemes.

The method of paired comparisons described elsewhere in this report, while hardly common in practice in systems across the world, could in its simplest form be used in the training of teachers to make subjective judgments between multiple pairings of folios.

Comments on establishing and monitoring reliability

No mechanism, of itself, can guarantee reliability in teacher assessment. Reliable assessment only occurs after large-scale implementation strategies, or experience over time, or a tacit understanding amongst the practitioners. Only the first-mentioned of these is a transparent mechanism for disseminating standards. It is important, however, to recognise the second of these (experience over time) when making a realistic evaluation of an evolving system. Vital, therefore, are (i) teacher professional development (including in-built professional development of the type described in “immersion” above) and (ii) ensuring that marking rubrics, whatever form they take, provide teachers with a simple structure for assessment, written in such a way that multiple interpretations of the standards are not likely to occur; rather, that the intended standards are the applied standards.

3.3 Validity of the current system

The investigation addressed two questions concerning the validity of the current system:

1. Do the current assessments accurately reflect the actual ability of the learner?
2. What are the threats to validity of the current system?

Note: The number of schools in the acquired sample is low (18 secondary schools and 48 primary schools). The percentages in the tables and graphs within Section 3.3 cannot be viewed as representative of all schools. Rather, the results should be taken as indicative only.

3.3.1 The concept of validity in a teacher assessment system

Before answering the questions posed, it is valuable to understand what validity means in the context of teacher assessment. **Validity is generally understood in terms of the extent to which an assessment can be seen to assess what it is intended to assess** (Crisp, 2010; Harlen, 2004); that is, that the evidence produced in response to the assessment is likely to be a suitable demonstration of the targeted aspect(s) of learning. Teacher assessment allows for higher levels of context validity because it is based upon judgments of student work produced during the learning process and it can encompass a range of student assessment types such as complex problem-solving, investigations, inquiries, authentic and performance assessments. This expansion of assessment types is often taken to be at the expense of reliability (Freebody, 2005).

Therefore it is also valuable to be aware of recent and continuing discussions about the shifting relationship between validity and reliability that have been brought on by the introduction of new forms of assessment (Johnson & Burdett, 2008) such as school-based, performance-based, and authentic assessment (Broadfoot & Black, 2004; Chatterji, 2003; 898; Harlen, 2005; Moss, Girard, & Haniford, 2006; Pitman et al, 1999; and Rust, 2007). While validity (and reliability) have been extensively defined and explored in the context of testing, there has been recent additional work about what might constitute quality in these alternative forms of assessment.

What is being highlighted here is the complexity of the validity concept and the necessity to demand high levels of validity in teacher assessment without necessarily expecting that reliability can be sustained at a level commensurate with that for standardised external tests.

According to Sireci (2009) there are some fundamental notions about establishing the appropriateness of assessments; namely, that assessments should (a) measure what they purport to measure, (b) demonstrate predicted relationships with other measures of the intended constructs, (c) contain content consistent with their intended uses and (d) be put to purposes that are consistent with their design and are supported by evidence.

In summary, it is widely accepted that validity is an important concept and it is increasingly accepted that validity is as important as reliability, especially in a system that privileges teacher judgment but there continues to be debate about how it can be measured and established.

3.3.2 Measurement of validity in the current system

As O'Brien (1998) alludes to, validity measures are not within the technical repertoire and resources of schools. However, variations of methods for establishing validity (see later) are applied in Wales, but often without a formal structure.

The importance of trust

Inevitably in a discussion of teacher assessment the issue of trust arises, and while it is not strictly a part of validity, it is worthy of discussion here. Understanding the nature of trust is urged by Parkes and Maughan (2009) for whom the purpose of moderation is "on the one hand to ensure that the judgments are as reliable as possible, but on the other hand, and perhaps more importantly in some cases, to ensure that stakeholders have trust in the teacher judgments". Brookhart (2012) warns that "for the more politically charged judgments of school accountability, standardised tests will garner more trust [in the USA]." Also in the USA, Meisels, Bickel, Nicholson, Xue, and Atkins-Burnett (2001) have concerns about trustworthiness and consistency. For example, they query the subjectivity inherent in teacher assessment. Sadler's (1986) view on this phenomenon is:

> "Subjectivity/objectivity does not reside in the format but in item/test construction and scoring. At every stage in the design and administration of any objective test constructed by a teacher, subjective judgments are involved. The teacher has to decide on the subject matter to include, the behaviours to sample, the complexity and difficulty of proposed asks, the item format, and the wording and mode of presentation. The process is objective only at the very last stage, which is deciding on the correctness of an answer. So-called objective assessment consists of a chain of subjective decisions, with one final objective link. Unfortunately the essential objectivity of the end point and the fact that the outcome of the final step is often expressed in numerical form (which, to many people, is the hallmark of objectivity) obscures the subjectivity inherent in all the steps leading up to it" (Sadler, 1986).

The current state of research reflects the small number of instances where moderated teacher assessment regimes (or systems) have been systematically implemented and supported over time. In most cases the system is new (the result of a reform process, for example) or evolving. The recurring theme in the literature about teacher assessment is that there is great potential for teachers' professional judgments to be used in the full range of roles attributed to other forms of assessment but that this will not be realised without "well-designed research on the technical characteristics of teacher assessment under different system conditions. Without this, a drive towards teacher assessment could well be a leap of faith, in the dark" (Oates, 2008, p. 39).

The onus is on systems where teacher assessments are reported or certified to develop more trust in teacher assessment, to develop a healthy assessment culture by carefully managing the quality of those assessments.

3.3.3 Do the current assessments accurately reflect the actual ability of the learner?

The scope of the investigation did not allow us to provide a definitive answer to this question, but it is worthwhile considering the small amount of evidence that addressed it because doing so shows that there were differing views across the sectors. A question that was common to all three questionnaires for the LAs, clusters, and the secondary and primary schools asked whether they considered the current assessments to accurately reflect the actual ability of the learners. Table 7 shows the ratings of the different sectors.

Table 7: Comparison of the ratings of the accuracy of the system

Overall, do you consider that your current (year 6 or year 9) assessments accurately reflect the actual ability of learners?	No		Yes	
	N	%	n	%
Primary schools	3	7%	45	93%
Secondary schools	4	25%	12	75%
Local Authorities	9	50%	9	50%
Clusters	2	13%	13	87%

As in the pattern seen in the responses from LAs and schools when they were asked about their confidence in the reliability of the assessment system, the primary schools had most confidence in the accuracy with 93% feeling that the assessments were an accurate reflection of student ability. This drops to 75% confidence in the secondary schools and to 50% at the LA level. Cluster coordinators had 87% confidence, which is somewhere between the levels of confidence in the primary and secondary schools – perhaps appropriately in view of the equally mixed composition of this group (between primary and secondary).

To obtain further evidence for this aspect of validity, a correlational study that compared students' performance on other independent assessments, such as standardised assessments, to their ranking using the teacher assessment system could be carried out.

A discussion in one of the cluster focus groups revealed an example of the problem of ensuring that teacher assessments do reflect pupils' actual ability. A secondary school in the cluster was critical of the fact that pupils' functional literacy on entry to the school at KS2, did not always correspond to the NC level they were awarded, particularly in relation to English and Science. This reflects the fact that, although secondary schools may work with the primary schools that feed pupils to them with regard to summative assessment, there is very little collaboration regarding formative assessment. In a separate issue, there was discussion in the cluster focus group about a situation in which 40% of a recent cohort of pupils entering Year 7 was assessed at being below functional literacy level. Both the secondary school and the primaries within the cluster stated that this was due to a drop in the proportion of pupils statemented and reassessed as not needing support on leaving Year 6 and entering Year 7. According to the secondary school and the cluster primaries this had become a very noticeable feature of the past few years. This reflects the fact that the policies and procedures for categorising special educational needs (SEN) pupils (including statementing) vary across authorities. In some authorities, the SEN criteria might deny additional support to those pupils with low reading scores.

3.3.4 What are the threats to validity of the current system?

An emerging threat to validity was identified by some secondary schools in their questionnaire responses. One school put it like this:

> "The assessments are gradually improving pupils' achievement levels BUT there is a danger that schools are under increasing pressure to raise levels in order to avoid being placed in the lower quartiles. KS3 national assessments in my opinion is beginning to lose credibility."

Similar comments were made in two of the ten cluster focus groups. One cluster focus group participant said that:

> "Pressure to raise standards, and the publication of performance data which is used as the basis of value added by the LA, Welsh Govt and Estyn. This is forcing the Level 4/5 boundary at KS3 in particular, and confidence in the national picture is eroding. This is manifested in the annual increases in % achieving Level 5 or higher in the Core Data Sets."

Another focus group stated that:

> "There is increasing pressure from the Government to raise standards. This message is passed on to Heads and Heads of Departments and teachers. If a school or department set consistent levels from year to year rather than increasing levels the school / department will fall into the 3rd and 4th quartile. As a result, the school and the relevant departments will be seen as failing. All schools are under pressure to inflate levels. If you don't do so your rank position in the Core Family Data sets[12] will decrease and Estyn are likely to view this negatively."

A secondary headteacher in that cluster said,

> "KS3 teacher assessments are losing credibility because of the increasing pressure to raise standards. If your school does not perform well in the Family Data the

[12] [12] The Core Family Data sets have been compiled by DfES to group schools in terms of their socio-economic characteristics, taking into account the percentage of pupils at each school who are eligible for Free School Meals; living in areas classed in the most 20% deprived areas in Wales; with special educational needs and whose first language is not English or Welsh. There are 20 such Families.

> school/department comes immediately under the spotlight and is interpreted as failing… It is unrealistic and nigh on impossible to raise standards each year as the school is an inclusive community and accepts cohorts of variable quality. Some years there are cohorts with higher percentages with additional learning needs and/or Special Educational Needs."

A similar concern was expressed by a primary-school headteacher in a different cluster, who was very concerned regarding the lack of verification and failure to sample TA/NC results. She felt that the current system was punitive (FSM [free-school-meals] benchmarking/banding/inspection etc.) and this encouraged teachers and schools to inflate their results. Consequently, she had little trust in the national data and core data sets that contribute to these.

Another threat to validity identified in two of the cluster focus groups was the standard of evidence used in making judgments about oracy, reading and writing. Participants questioned the current use of taped oral evidence and other "less robust evidence of pupil attainment", such as providing a video or PowerPoint presentation, which was felt by the secondary teachers to disadvantage pupils by not preparing them for KS4 where pupils undertake externally assessed examinations based on writing and comprehension only.

What other teacher assessment systems do to establish validity

As was the case for monitoring reliability (section 3.2.9) the description that follows of mechanisms used in other systems to establish validity does not assume that these mechanisms or variations thereof are not used or have not been considered in Wales. Nor does it assume that all are appropriate for teacher assessment at KS2 and KS3, given the less-than-high-stakes nature of the assessment programme. Also, as has already been mentioned, there is the necessity for system design to include trade-offs (Freebody, 2005; Nunnally & Bernstein, 1994) between validity and reliability. Again, although based on technical considerations, it is policy decision for systems as to the balance of mechanisms used for establishing reliability and validity.

The simplest definition of validity is that the assessments measure what they are purported to measure. In the case of teacher assessments this means that the assessments should measure achievement in the content and skills of the curriculum.

Internal moderation

Teacher assessments have the property of context validity because they occur close to where learning occurs – in the classroom or school – and the assessed curriculum is most likely to be the intended and presumably taught curriculum. Within a school, however, different teachers might be assessing different classes (pupils). This raises the question of intra-school comparability of standards. Some other systems in which teacher assessments "count" towards a final result have introduced internal moderation whereby teachers of the same subject in the same schools meet to share and discuss student work and provide feedback to each other about the way standards have been applied. This is a form of peer review.

Alignment of assessment with curriculum and pedagogy

Valid assessments are aligned with curriculum and pedagogy. Where, for example, students undertake extended writing or rich tasks or inquiry-based learning in the classroom while the assessments are in multiple-choice format, the alignment of the three “message systems” (Bernstein, 1990) cannot be accomplished. Most teacher assessment systems establish validity by using assessment techniques that reflect classroom experiences, not only in assessment format but also by allowing unlimited time (within reason), computer-generated text as well as or instead of handwriting, and so on.

Construction matrix

From Cronbach’s (1988) “operationist perspective” on validity, the notion of range and balance is vital in the construction of assessment instruments, whether the instrument is composed of items on a test or a collection of tasks making up a portfolio. Range and balance can be represented by a construction matrix or grid in which characteristics of the assessment instruments are tabulated, characteristics such as perceived difficulty, estimated time for completion of instrument, curriculum element(s) or objective(s) being assessed, and nature of the text that dominates in the instrument (e.g., verbal, numerical, spatial) to name a few. As assessment is ultimately a process of selecting only some of the content and skills that pupils are expected to learn, it is essential to cover the curriculum in a way that parallels the frequency and depth of elements in the curriculum. The construction matrix is one possible representation of the curriculum elements or syllabus outcomes that are sampled in any instance of the assessment and mapped on to syllabus content.

Face validity

Face validity, while based on opinion rather than facts, is particularly important in new assessment systems. The opinions of parents, the general public and government cannot be overestimated for the ultimate success of an initiative. What new systems often do in such circumstances is to administer questionnaires to a sample of stakeholders. A novel approach is to invite members of the public to undertake the assessments and/or to study the work of pupils.

Panelling

Panelling (or reviewing) is primarily a validation exercise. Systems producing standardised tests include panelling as an essential part of the test development cycle. Convening panels in the development of teacher assessments is probably a luxury but, nevertheless, some other systems do establish procedures whereby assessment instruments are reviewed by internal and external panels. Experts work collaboratively in small groups, both at the item level and the test level. Many features of a test or an item can affect validity, including equity issues and presentation, so different panels with different kinds of expertise should be employed. Types of panels that exist in systems include the following: in-house, subject expert, equity, editorial, and scrutiny (in terms of difficulty and time allowed). Each panel has its own special charter.

Accreditation of assessment tasks

Where teachers are not experienced in designing assessments the system has a special responsibility to pupils. If the assessments are deemed invalid (say in content or difficulty level) after the assessment has taken place (say at moderation meetings) there becomes a moral issue about the level that should be awarded to the pupil. On the basis of evidence produced the pupil's work is probably at a lower level than it might have been if good assessment had occurred. On the other hand, the basis for assessment decisions has to be what a pupil has accomplished rather than what s/he might have accomplished. One way of reducing the chances of having to make a hard decision at the end of the assessment programme is to have a process whereby the assessment instruments to be used (or at least blue-prints of them) are submitted to an external panel for approval.

Statistical evidence

With regard to statistical evidence of validity after the assessment has occurred, a technical panel is convened to study and evaluate data relating to instances of possible bias against sub-groups of the population (differential item functioning). Key concepts here are the Hoover-Welch (HW3) statistic, which is based on testing hypotheses between means of sub-groups defined by gender (or other indicator of interest) and ability level, and the Mantel-Haenzel statistic, which is based on constructing a three-way contingency table, dimensions being gender (or other indicator of interest), ability, and response.

Recommendation 2 – Validity

There is evidence that the face validity of the current system is already under threat because some schools have lost trust in the judgments made in the teacher assessment process. While face validity is not the only form of validity, it is important that those who need to operate the teacher assessment system should believe that its outcomes are valid. Without that confidence in the system, it will falter. To restore confidence Recommendation 3 on the impact of the teacher assessment system and Recommendation 4 on the operability of the system that are contained in this report should be implemented.

3.4 Impact of the teacher assessment system

The investigation examined the impact that the implementation of the teacher assessment system is having. Specific questions addressed in this part of the investigation include:

1. Is the implementation and delivery of the moderation programme in line with best practice?
    a. How do the systems at KS2 and KS3 compare to alternative systems that have been shown to work effectively in other education systems around the world?
2. Are the assessment and levelling procedures being implemented as planned?
    a. Are schools actually following the procedures as designed? If not, why not?
3. How does the current cluster moderation programme for English, Welsh and Welsh Second Language work to improve assessment?

Note: The number of schools in the acquired sample is low (18 secondary schools and 48 primary schools). The percentages in the tables and graphs within Section 3.4 cannot be viewed as representative of all schools. Rather, the results should be taken as indicative only.

3.4.1 Quality of the implementation

The investigation addressed the question, "Is the implementation and delivery of the moderation programme in line with best practice?" and the subsidiary question of, "How do the systems at KS2 and KS3 compare to alternative systems that have been shown to work effectively in other education systems around the world?"

In order to answer the question about whether the implementation and delivery of the moderation programme in Wales is in line with best practice, a review of the literature on school-based assessment regimes across the world was conducted. The review revealed that there are many ways of executing teacher assessment and moderation, and that each has its own special way of operating. However, a more detailed analysis of the literature reveals enormous similarities (at least, in statements of purpose, policy and practice) between systems internationally, similarities that quite possibly arose from the export and import of moderated teacher assessment models from jurisdiction to jurisdiction over the past decade (at least) since external examinations and tests have been relegated a lesser role than in the past.

3.4.1.1 Components of Quality Teacher Assessment Systems

Three particular pieces of work (Allen, 1993; Meiers, Ozolins, & McKenzie, 2007; and Gipps, 2002) are useful in developing a unifying construct for viewing teacher assessment systems, and each is discussed below.

Allen (2003) distilled the five key *elements of an effective assessment system* that underlie the best of a wide range of assessment policies and practices permeating the educational research literature. These five elements are shown in the first column of Table 8 below in general terms with some exemplification in square brackets after them. Corresponding entries in the second column refer to the particular situation in Wales. As seen in the table, the current system contains all five of the elements.

Table 8: Comparison of the elements of the current Welsh teacher assessment system to the five key elements identified by Allen (2003)

The five key elements of an effective assessment system (Allen, 2003)	Current Welsh teacher assessment system
1. There are guidelines that teachers/schools must use in planning [syllabus].	1. As shown in Table 1 in this report, there are numerous documents that provide guidance and support to teachers implementing the curriculum and assessment system.
2. There are formal plans for student learning and achievement that teachers/schools must make [work programme].	2. Schools set out plans for covering the content and skills required by the National Curriculum.
3. Evidence of student achievement must be produced [folio of student work].	3. Schools are required to produce folios of student work as part of the moderation process.
4. This evidence must be assessed against the guidelines and plans [teacher judgment based on pre-set standards].	4. Teachers are required to judge student work against a set of pre-set standards.
5. There is a process for validating teacher judgments of student achievement [social moderation].	5. Moderation sessions are held to validate the judgments that teachers use.

Meiers et al. (2007) used a parallel set of elements in their review of a research base that was structured around the themes of assessment in education, standards-referenced assessment, moderation of teacher assessments, and teacher professional learning as a key *strategy for improving the consistency of teacher judgments.* Table 9 summarises the five elements identified by Meiers in the first column and in the second column is a comparison of the current Welsh teacher assessment system against those components. Again, the Welsh teacher assessment system contains the elements identified by Meiers et al.

Table 9: Comparison of the elements of the current Welsh teacher assessment system to the five key elements identified by Meiers et al. (2007)

The five key elements of an effective assessment system (Meiers et al., 2007)	Current Welsh teacher assessment system
1. Teachers' knowledge and understanding of the standards	1. Teachers do know about the standards and have some understanding of them.
2. Curriculum planning; opportunities for students to demonstrate achievement standards	2. Students are given the opportunity to demonstrate the achievement standards.
3. Evidence of student achievement	3. Evidence of student achievement is produced for assessment by teachers.
4. Assessment of the evidence against the standards	4. Teachers assess the student work against the standards.
5. Validation of teachers' judgments.	5. The moderation sessions are designed to validate the teachers' judgments.

In another relevant review, Gipps (2002) lists five *elements in the process of* assessment, and Matters (2005) later added a sixth entry to the list. The combined list of six elements are shown in the first column of Table 10 below, and the second column shows how well the current Welsh system matches the six criteria.

Table 10: Comparison of the elements of the current Welsh teacher assessment system to the six key elements identified by Gipps (2002) and Matters (2005)

The six key elements of an effective assessment system (Gipps, 2002 and Matters, 2005)	Current Welsh teacher assessment system
1. Assessment task (derived from the curriculum)	1. Student work is generated from assessments that are aligned with the national curriculum.
2. Student performance (which is not always written)	2. There are student performances, and they are not all written.
3. Judgment of the performance with reference to a standard	3. Performances are judged against standards.
4. Feedback to the learner and the teacher/curriculum	4. There is no feedback to the learner although there is feedback to the teacher.
5. Moderation	5. Moderation is used to validate teacher judgments.
6. Teachers and administrators are assessment-literate in order for them to take advantage of the information that data have to offer.	6. Administrators and teachers in the Welsh system are working on their assessment literacy.

The Welsh teacher assessment system contains all of the six elements in the Gipps/Matters list with a couple of qualifications. The qualifications are associated with element (4) in there not being direct feedback to the learner, and with element (6) in there not necessarily being high levels of assessment literacy – a common observation about assessment systems world-wide.

So, the Welsh teacher assessment system contains the components that have been identified in the research literature as being the essential elements of effective teacher assessment systems as identified by Allen (2003), Meiers et al. (2007), and Gipps (2002). However, when considering the list of elements in the process of assessment, based on evidence collected in other parts of this investigation, the implementation of the Welsh system is not fully meeting best practice.

3.4.1.2 School-based assessment regimes internationally

The detailed literature review conducted for this investigation provides a thorough account of where school-based assessment systems have been implemented around the world and what their characteristics are. In many educational systems, such as those of Australia, Canada, the UK and Finland, school-based assessment is used extensively or exclusively in providing information about student achievement. In Hong Kong, school-based assessment has been a part of the public examinations system since 1978 when it was first introduced into the Chemistry examination so that there could be an assessment of laboratory work. By 2006, school-based assessment had been implemented in 13 "A" Level subjects and 13 Certificate of Education subjects, including English language. School-based assessment became a core component of the Hong Kong Certificate of Education Examination in English Language in 2005–2007, and was then revised and extended for the Hong Kong Diploma of Secondary Education. School-based assessment is to be progressively incorporated into all 24 subjects.

School-based assessment is policy-supported practice in an increasing number of educational systems around the world, including those of Australia, New Zealand, Canada, and the UK. It is increasingly being adopted as national educational policy in Asia as well as in some developing countries, including Ghana and Zambia. It is also actively promoted in the USA, although always overshadowed by national testing programs.

Appendix 4 presents an overview of the status of school-based assessment in 12 jurisdictions. It should be noted that the cycle for change in assessment systems is quite short; the currency of entries was that at the time of drafting the report.

The literature review did not produce a definitive statement on the relationship between type of assessment regime and performance on international tests because it is not a simple matter to categorise assessment regimes as internal or external as there are many variations within each (internal and external) as well as between them and also because the nature of the assessment regime in a particular country can change rapidly. Furthermore, causal links are difficult to establish.

The literature review revealed that, unsurprisingly, there is variation in the moderation model applied: the moderation model may be different at different stages of schooling (compulsory versus non-compulsory years). The moderation model may vary according to the mode of assessment (written expression versus constructed response). The moderation model may be different for high-stakes and low-stakes assessment. The moderation model may differ depending on where control of the assessments lies (say, with teachers or with a central authority). The international experience indicates that there is no single factor that will be sufficient on its own to ensure that teacher assessment is good or to ensure that moderation is good. For assessment in the primary and middle years of schooling, however, cluster moderation meetings appear to be the most appropriate for the level of control required and the stakes of the assessment decisions. Wales currently applies this kind of cluster moderation at KS2.

Models of assessment that exist internationally for socially moderated teacher assessment in Years 3–9 belong to one of two categories – emergent upwards or derivative downwards. The former category includes systems moving from a position where teacher assessment had been considered suitable only for formative assessment, to a position where teacher judgments had come to be privileged for summative assessment, typically in the primary and middle school years. What these "emergent upwards" systems have in common is that they try to make sense of practices undertaken elsewhere in order to distil a distinctive model for their own circumstances. The "derivative downwards" category includes systems incorporating a model that has been established in the senior years of schooling to the junior years of schooling. What these derivative downwards systems have in common is that they need to develop an understanding of the principles upon which the senior model is based and then decide how to adapt that model to suit lower-stakes purposes.

3.4.1.3 Research on Queensland's system of externally moderated assessment

Queensland in Australia has one of the longest-standing and most fully developed systems of school-based standard-referenced assessment, and is a case study worthy of examination. It operates at the senior schooling level, Years 11 and 12, and leads to certification at the end of schooling. There is also a less developed system operating for Years 1−9, which has been in place since 2008.

Many of the papers cited in the education literature about externally moderated school-based assessment are those written by Queensland academics (e.g., Beasley, McMeniman, Sadler) in the 1980s and key figures in the implementation of criteria-based assessment (e.g., Pitman, O'Brien, Allen) in the period 1980 to 2000. Beyond this, the Queensland model has been used as an exemplar in international scholarly articles (e.g., Shavelson, Black, Wiliam, & Coffey, 2004; Elwood, 2006; Myford, 1999; Stobart, 2003; and Strachan, 2002). New Zealand work, too, is often cited (see for example, Smaill, 2012; Brown, 2011).

The Queensland senior assessment system is not so much underpinned by theory as having been and continuing to be a theory-building exercise in itself. As it has matured, Queensland has struggled with implementing the system when teachers were unprepared for radical change. It has done this by implementing the recommendations for change along the way and successive governments allowed the model to evolve, a process which continues today, 40 years on from when external examinations were abolished in Queensland (Wyatt-Smith & Matters, 2009). The system's strengths stemmed from its dual reliance on the agency of the teacher and the partnerships forged between schools and the curriculum–assessment authority.

Discussion of alignment with best practice

In answer to the question as to whether the implementation and delivery of the current moderation programme is in line with best practice, the Welsh system has all of the components identified as key elements of effective assessment systems. As demonstrated in the literature, having all of the requisite components is important, but not sufficient to ensure the success of the system in practice.

The review of literature demonstrates that introducing such teacher assessment systems requires considerable change across schools and local authorities because it has implications not only for the assessment, but also for pedagogy in that it brings the focus of teaching back to promoting learning by careful judgment by the teacher coupled with feedback to the student and changes to instruction as needed.

One of the most challenging features of the systems is ensuring the quality and consistency of teacher assessments of student work. The literature is full of exhortations about consistency of teacher judgments, the use of exemplars, and the need for professional development of teachers. All of this indicates that there is a high level of agreement among educationists, policy makers and researchers about what constitutes good teacher assessment and moderation, but it is evident from the experience of countries that are further along the path than Wales that it takes years and several iterations of the process to develop the levels of teacher skills and experience to achieve the desired levels of reliability of judgments.

Even in the system that has been running in Queensland for 20 years, which achieved high degrees of reliability of scoring and widespread acceptance by schools and universities, there is a culture of continuous theory-building. In comparison, the Welsh system is still in its infancy, having been in existence since 2005, but only since 2010 with the external moderation component. The evidence from this investigation is that there are certainly areas in the system where improvements can be made, but given the history of the development of other systems, this is to be expected. The Queensland case study shows that high reliability is ultimately achievable with sustained training of teachers and continued refinement of processes.

3.4.2 Fidelity of implementation

The second evaluation question regarding impact asks if the assessment and levelling procedures are being implemented as planned. In addition, it asks whether schools are following the procedures as designed and, if not, why not. Essentially, these are questions about the fidelity of implementation of the current teacher assessment system. It is important to address this question first, because if the system is not being implemented as designed and intended, then it is hard to answer the subsequent questions relating to impact.

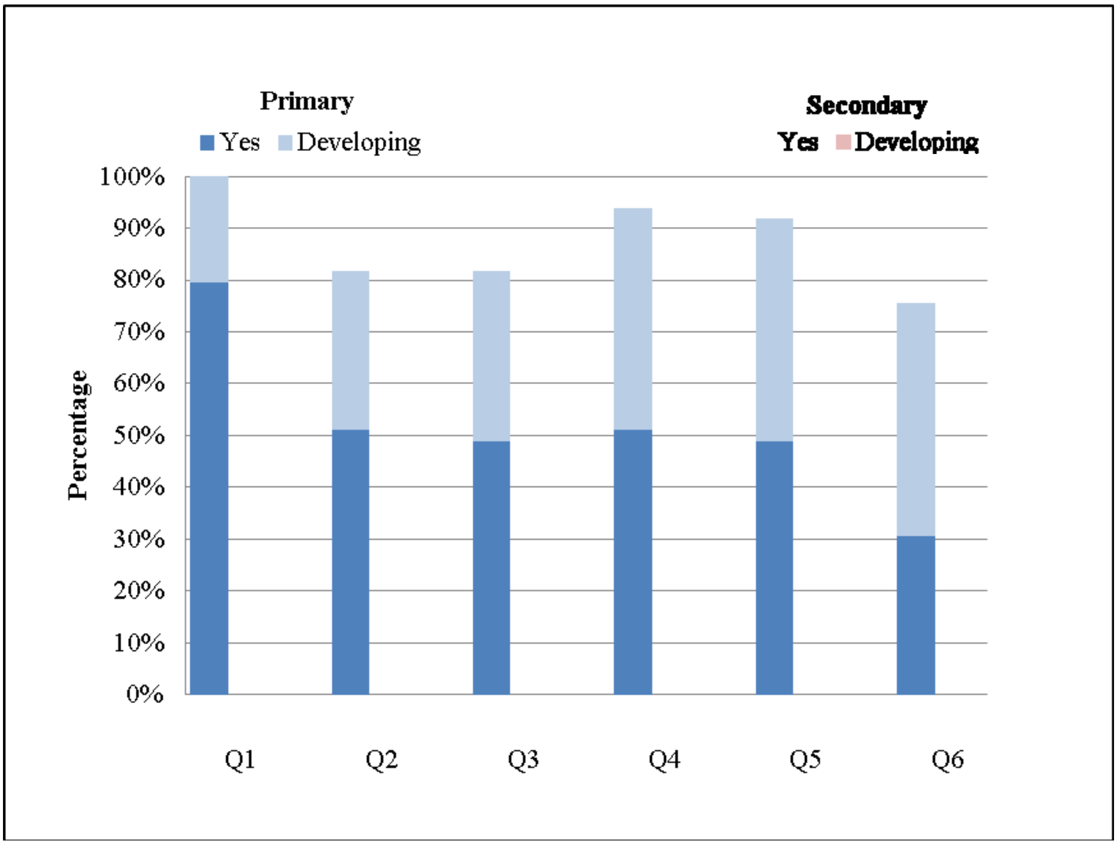
As described in section 3.1, which explains the origins of the current system, the current teacher assessment scheme has a number of components that rely upon teachers, schools, local authorities and the WJEC for its implementation. For the system to work smoothly and effectively, each of these players has to understand the system and their role in it, as well as following the necessary procedures.

The investigation examined, through the use of questionnaires and interviews, the extent to which stakeholders in the system understood the system and the fidelity with which they were able to implement the intended procedures. The following sections report on the findings for the different sectors (the primary schools, the secondary schools, the local authority advisors, and the clusters).

3.4.2.1 Schools' understanding of policy

Figure 2 shows a graph and table that present the primary and secondary schools' responses to six questions that addressed school assessment policy. The first question (Q1) asked schools to judge how clearly their assessment policy distinguished between formative and summative assessment. Ninety-four percent of secondary and 80% of primary schools thought that their policies made that distinction clear.

Figure 2: Schools' views of their assessment policy



		Primary schools		Secondary schools	
		n	%	n	%
Q1: Does the policy clearly distinguish between formative and summative assessment?	No	0	0%	0	0%
	Yes	39	80%	17	94%
	Developing	10	20%	1	6%
Q2: Does the policy clearly define standardisation and moderation?	No	9	18%	1	6%
	Yes	25	51%	12	67%
	Developing	15	31%	5	28%
Q3: Does the policy clearly distinguish between standardisation and moderation?	No	9	18%	1	6%
	Yes	24	49%	10	56%
	Developing	16	33%	7	39%
Q4: Does the policy clearly describe arrangements for internal standardisation?	No	3	6%	2	11%
	Yes	25	51%	12	67%
	Developing	21	43%	4	22%
Q5: Does the policy clearly describe arrangements for internal moderation?	No	4	8%	2	11%
	Yes	24	49%	10	56%
	Developing	21	43%	6	33%
Q6: Does the policy clearly describe arrangements for cluster standardisation and moderation?	No	12	25%	2	11%
	Yes	15	31%	9	50%
	Developing	22	45%	7	39%

Schools were less confident, however, in how well their policies defined standardisation and moderation (Q2) and distinguished between them (Q3). Only about half of primary schools felt that their policy did these things, while about a third said they were still developing, and the remainder answered that their policy did not do this. Of the secondary schools, 67% were confident that their policies clearly defined standardisation and moderation, and 56% thought it distinguished between them.

When asked how well their school policy clearly described arrangements for internal standardisation (Q4) about two-thirds of the secondary schools thought they did, while 22% said they were still developing and 11% said that it did not. Just over half of the primary schools thought their policy did and 43% said they were still developing, while 6% said they did not.

When asked how well their school policy clearly described arrangements for internal moderation (Q5), 56% of the secondary schools thought they did while 33% said they were still developing and 11% said that it did not. About half of the primary schools thought their policy did and 43% said they were still developing, while 8% said they did not.

When asked how well their school policy clearly described arrangements for cluster standardisation and moderation (Q6), half the secondary schools thought they did while 39% said they were still developing and 11% said that it did not. Only 30% of the primary schools thought their policy did and 45% said they were still developing, while one-quarter said they did not.

Discussion of schools' views of assessment policy

Overall, the secondary schools tend to have higher levels of confidence that their assessment policy makes clear definitions and distinctions between the key components of the teacher assessment program, although there is an obvious need to have schools ensure that policies meet the required levels of clarity. This is echoed in the data gathered in the cluster focus group sessions, which revealed that not all schools had a complete understanding of the assessment policy that they should have in place. One primary school's assessment policy still lacked a number of the required elements and two others in the same cluster indicated the same. In that cluster it was noted that there was confusion about what needed to be in the school assessment policy and about the difference between internal and external standardisation and moderation procedures. In another cluster the secondary school had a good understanding, but one primary school's assessment policy failed to define and distinguish between standardisation and moderation, while another primary school's policy did not describe arrangements for either internal or cluster standardisation and moderation. In another cluster two primary schools stated that their policy does not define or differentiate clearly between standardisation and moderation or for the arrangements for those in cluster moderation sessions. Similar situations existed in other schools across four other clusters, indicating that there seems to be a weakness in policy documentation in a proportion of the primary schools, which is of concern given that it is four years since the statutory introduction of the policy documents.

3.4.2.2 Transition plans

In the questionnaires schools were asked two questions about their transition plans. Over 85% of the primary and secondary schools responded that their transition plans had been renewed or updated in September 2010, about 10% of them said they were still developing in this respect, and the remainder said they were not in development. When asked if cluster moderation arrangements were the same as described in the current cluster transition plan, 72% of primary schools and 82% of secondary schools reported that they were, with 15% of primaries and 18% of secondaries saying that they were still in development. Thirteen per cent of primary schools said they were not.

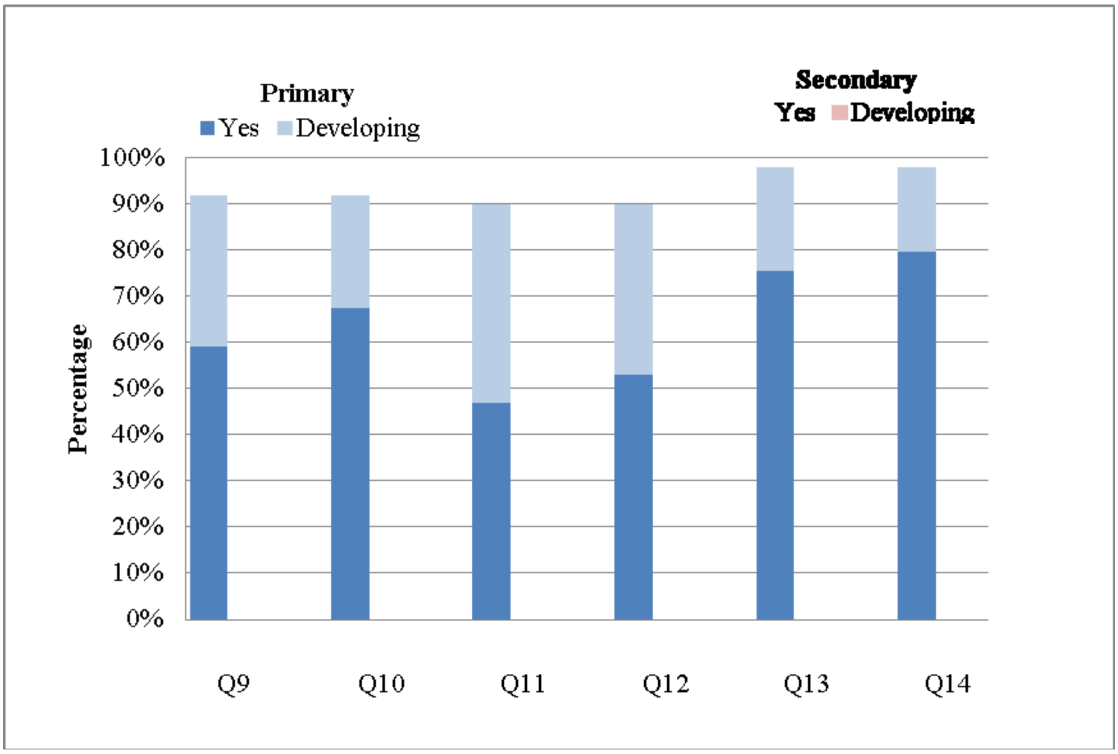
Discussion of school transition plans

School transition plans have, on the whole, been renewed or updated, but there is some difference between the cluster moderation arrangements described in the transition plans and what happens in practice and it would be good to make sure that these are aligned.

3.4.2.3 Teachers' understanding of the assessment system

Six questions on the schools' questionnaire addressed teachers' familiarity with, and understanding of, elements of the teacher assessment scheme. As can be seen from the graph and table in Figure 3, teachers' understanding of assessment procedures was weaker than it should be.

Figure 3: Teachers' understanding of assessment procedures



		Primary schools		Secondary schools	
		n	%	n	%
Q9: Are all teachers in your school familiar with the contents of Making the most of learning (DfES, 2008)?	No	4	8%	5	28%
	Yes	29	59%	7	39%
	Developing	16	33%	6	33%
Q10: Are all teachers familiar with the contents of Ensuring consistency in teacher assessment: Guidance for Key Stages 2	No	4	8%	2	11%
	Yes	33	67%	8	44%
	Developing	12	25%	8	44%
Q11: Are all teachers familiar with the contents of Making the most of assessment	No	5	10%	3	17%
	Yes	23	47%	6	33%

(DfES, 2010)?	Developing	21	43%	9	50%
Q12: Are all teachers familiar with the contents of the annual DfES guidance Statutory assessment arrangements for the school year 2011-12?	No	5	10%	3	17%
	Yes	26	53%	11	61%
	Developing	18	37%	4	22%
Q13: Do all teachers clearly understand the difference between standardisation and moderation?	No	1	2%	0	0%
	Yes	37	76%	9	50%
	Developing	11	22%	9	50%
Q14: Do all teachers clearly understand the difference between standardisation portfolios and learner profiles?	No	1	2%	1	6%
	Yes	39	80%	8	44%
	Developing	9	18%	9	50%

When asked if all teachers in the school were familiar with the contents of the document *Making the Most of Learning*, published by DfES in 2008 (Q9), 59% of primary schools agreed, but only 39% of secondary schools said yes, with about a third of primary and secondary teachers still regarded as developing in this, while 8% of primary and 28% of secondary said their teachers were not familiar with the publication.

Similarly, when asked if all teachers in the school were familiar with the contents of *Ensuring Consistency in Teacher Assessment: Guidance for Key Stages 2* (Q10), 67% of primary schools agreed, but only 44% of secondary schools said yes, with about a quarter of primary and 44% of secondary teachers still regarded as developing in this. Eight per cent of primary and 11% of secondary said their teachers were not familiar with the publication.

There was even less familiarity reported with *Making the Most of Assessment* published by DfES in 2010 (Q11), with less than half of the primary schools and a third of the secondary schools reporting that teachers were familiar with its contents.

Similar low levels of familiarity with the contents of the annual DfES guidance *Statutory Assessment Arrangements for the School Year 2011−12* (Q12) were reported, with 53% of primary schools and 61% of secondary schools agreeing that their teachers were familiar with the guidance.

Three-quarters of primary schools thought that their teachers understood the difference between standardisation and moderation (Q13), whereas only half the secondary schools thought so.

Similarly, 80% of primary schools thought that their teachers understood the difference between standardisation portfolios and learner profiles (Q14), whereas only 44% of the secondary schools thought so.

Discussion of the understanding of the assessment system

Overall, primary schools rated their teachers' familiarity and understanding of the assessment procedures higher than the secondary schools rated their teachers' understanding. However, the levels of understanding in both school sectors are less than ideal. In particular, schools reported low levels of familiarity with DfES publications. The cluster focus groups also revealed that there were still gaps in teachers' understanding of the system. Similarly, in another cluster some schools indicated that teachers were still unsure of the difference between standardisation and moderation.

The cluster focus groups indicated that a possible explanation for schools saying that not all of the teachers understand the assessment procedures is that there is inconsistency in dissemination of information and related training within schools. In five of the ten focus groups, comments were made about the inadequacy of the current scheme of training that assumed that once some teachers are trained, they will go back to their schools and 'cascade' what they had learned to other teachers. In reality, this was not implemented evenly. Several schools reported that there was insufficient time to do this owing to other pressures of work. In four of the cluster focus groups it was suggested that a better scheme would be to have a national training programme that covered all aspects of the assessment system. It was felt that if such a programme was rolled out to all teachers, then accuracy and consistency of teacher assessment judgments would improve and restore confidence in the system which was increasingly being viewed by teachers as lacking rigour, accuracy and credibility.

3.4.2.4 Pupil tracking systems

Of those schools that responded to the questionnaire, all of the them except for one primary school, said that they have a system in place for recording and tracking pupil attainment. The tracking systems in use are fairly comprehensive and track pupil attainment across most subjects – oracy, reading and writing (Welsh and English, but less so in Welsh); mathematics; science; and non-core subjects in secondary schools. The types and mixtures of tracking systems varies from school to school and between the primary and secondary sector. School-based systems predominate in the secondary schools (comprising 61% of those reported compared with 29% for primary schools), whereas in primary schools commercial packages (31%) and multiple tracking systems (35%) were more common (compared with 22% and 17%, respectively, for secondary schools). An instance of an LA-based tracking system was reported for a primary school.

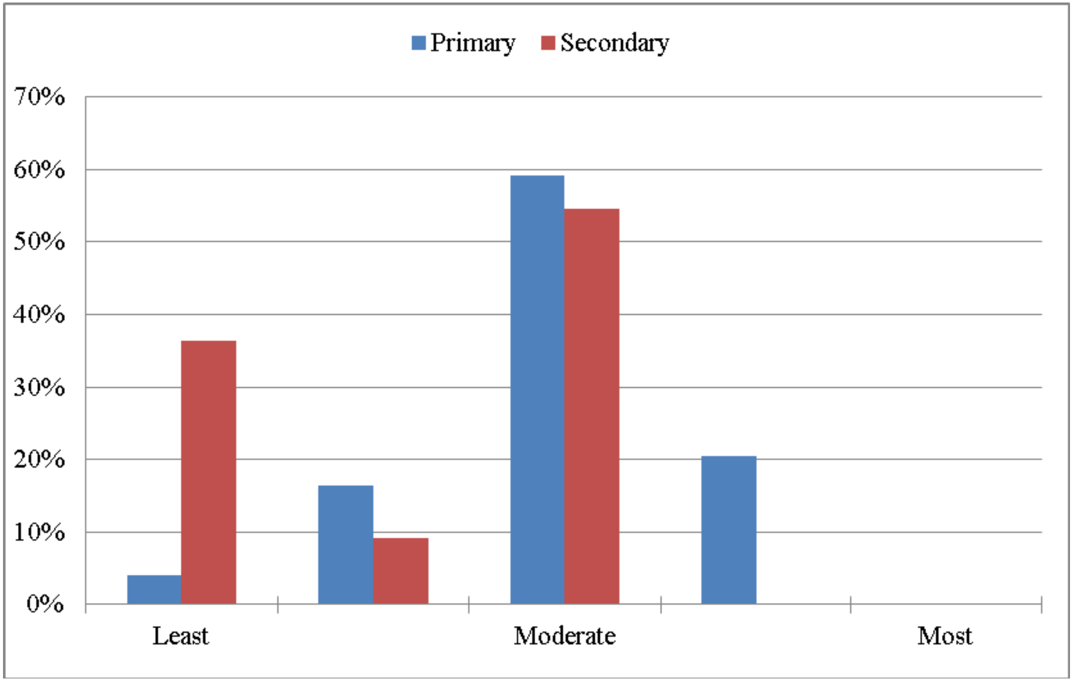
Discussion of the tracking of pupil attainment

Generally, the tracking of pupil attainment data across subject areas seems adequate, but there is a range of different pupil tracking systems in use across the primary and secondary schools, which can lead to lack of compatibility that hampers transfer of data across schools. As reported earlier, there was comment from the secondary schools about the need to have compatible pupil tracking systems at KS2 and KS3 to enable pupil progress to be effectively managed across the cluster.

3.4.2.5 Standardised tests

The schools' questionnaires asked if they used standardised tests to assess pupils at KS2 and KS3. Nearly all of the primary schools (95%) reported that they used standardised tests at KS2, whereas only 59% of secondary schools used them at KS3. Schools were then asked to indicate, on a scale of 1 to 5, the weighting that they gave to standardised test outcomes in contributing to pupils' best-fit NC levels and the results are presented in the graph and table in Figure 4.

Figure 4: Schools' weighting of standardised test outcomes



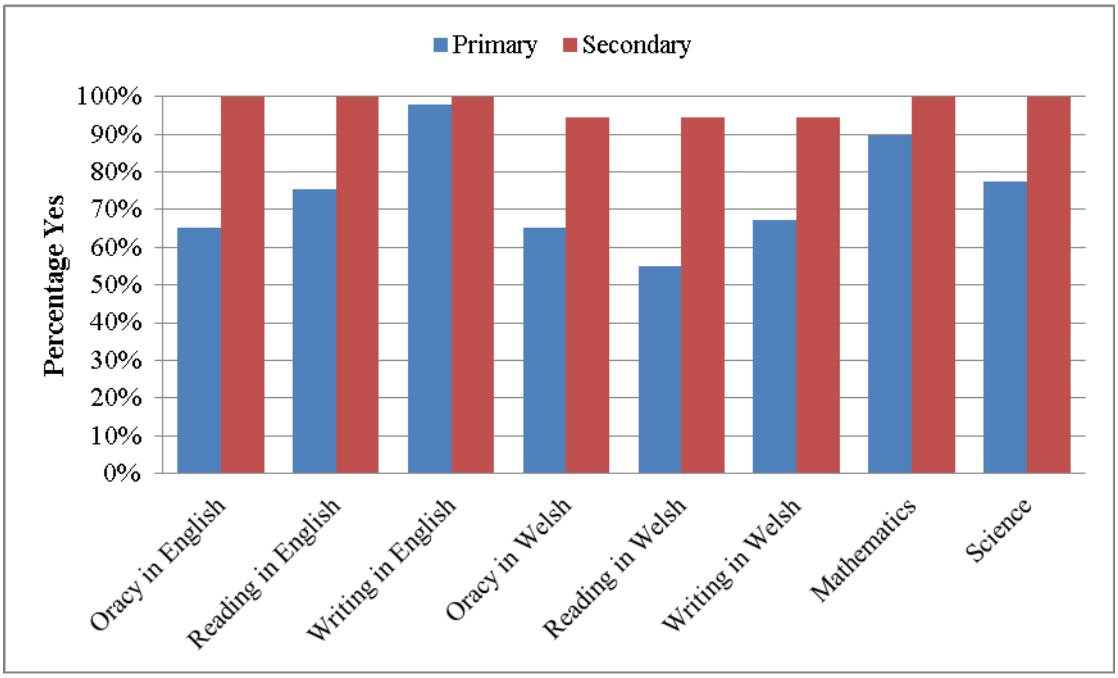
	Primary schools		Secondary schools	
	n	%	n	%
1 (Least weight)	2	4%	4	36%
2 (Little weight)	8	16%	1	9%
3 (Moderate weight)	29	59%	6	55%
4 (Substantial weight)	10	20%	0	0%

Figure 4 shows that both primary and secondary schools predominantly give moderate weighting when considering standardised test outcomes for best-fit decisions. They differ in that the second most common weighting given for primary schools was the 'substantial weight' category whereas the second most common category for secondary schools was 'least weight'. Looking at the distribution of weightings numerically, the mean weight for primary schools is 2.96 (SD .735) compared to a mean weight for secondary schools of 2.18 (SD .982). The fact that, overall, primary schools give more weight to standardised test data when they make best-fit judgments on NC levels may be reflective of the fact that more primary schools give standardised tests than do secondary schools. This could be a function of different levels of confidence in devising assessment by primary and secondary schools.

3.4.2.6 Internal standardisation

As shown in Figure 5, the majority of secondary schools that responded to the questionnaire reported that they were implementing internal standardisation across all subjects – oracy, reading and writing (Welsh and English); mathematics; science; and non-core subjects. However, fewer primary schools reported that they conducted internal standardisation. The numbers for Welsh oracy, reading and writing are lower because they do not apply to all schools, but even for other subjects, the range is from 65% for oracy in English to a high of 98% for writing in English.

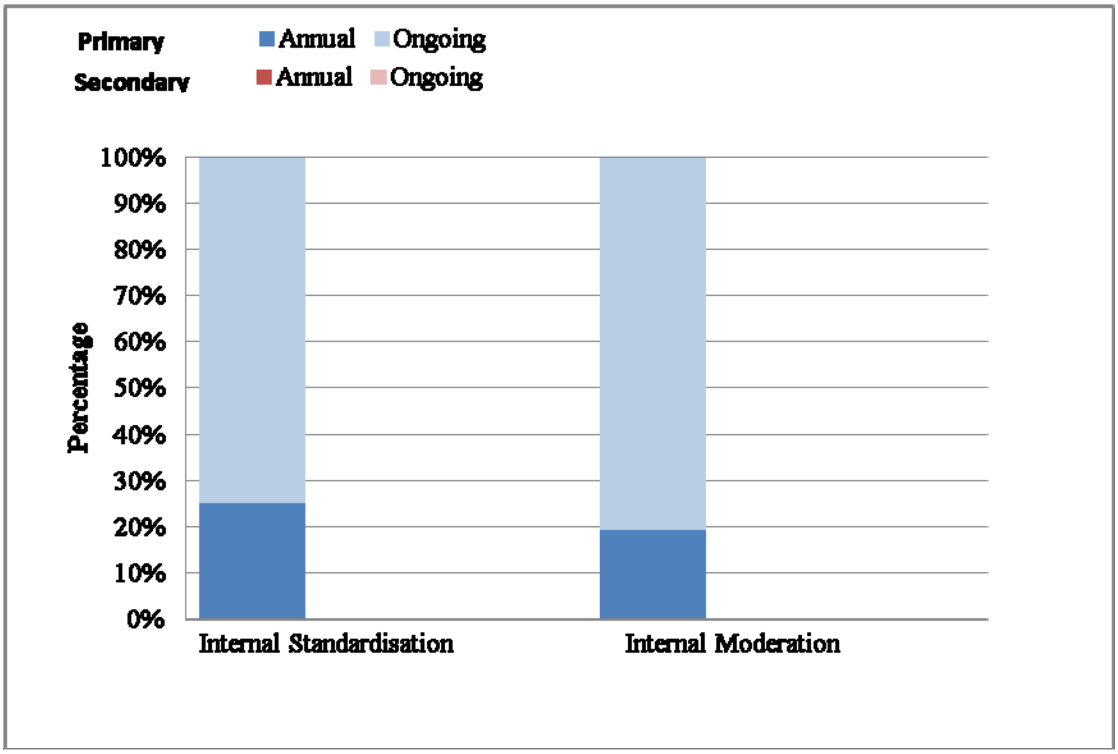
Figure 5: Percentage of schools that conduct internal standardisation across subject areas



Does internal standardisation take place in the school for …	Primary schools		Secondary schools	
	n	%	n	%
Oracy in English	32	65%	18	100%
Reading in English	37	76%	18	100%
Writing in English	48	98%	18	100%
Oracy in Welsh	32	65%	17	94%
Reading in Welsh	27	55%	17	94%
Writing in Welsh	33	67%	17	94%
Mathematics	44	90%	18	100%
Science	38	78%	18	100%

Schools were asked about the frequency with which internal standardisation and moderation meetings were held. Figure 6 shows that internal standardisation was an ongoing event for 75% of the primary schools, the remainder doing it annually, while all secondary schools conducted standardisation in an ongoing manner. For 81% of primary schools and 94% of secondary schools internal moderation was an ongoing event.

Figure 6: Frequency of internal standardisation and moderation



		Primary schools		Secondary schools	
		n	%	n	%
Q21: Is internal standardisation an ongoing or annual event?	Annual	12	25%	0	0%
	Ongoing	36	75%	17	100%
Q32: Is internal moderation an ongoing or annual event?	Annual	9	19%	1	6%
	Ongoing	38	81%	17	94%

3.4.2.7 Existence of standardisation portfolios

Table 11 shows that among the schools that responded to the questionnaires, the implementation of standardisation portfolios across the subject areas is much higher in secondary schools than in primary schools. All of the secondary schools reported that they had standardisation portfolios in place for English oracy, reading and writing, and in mathematics, which contrasts with primary schools where less than half of the respondents reported having portfolios for English oracy and reading. Overall, primary schools' implementation of standardisation portfolios was low.

Table 11: Extent of implementation of standardisation portfolios across subject areas

Does the school have standardisation portfolios in place for …	Primary schools		Secondary schools	
	n	%	n	%
Oracy in English	22	45%	18	100%
Reading in English	22	45%	18	100%
Writing in English	44	92%	18	100%
Oracy in Welsh	19	39%	16	89%
Reading in Welsh	17	35%	16	89%
Writing in Welsh	36	76%	16	89%
Mathematics	33	69%	18	100%
Science	29	61%	17	94%

3.4.2.8 Attendees at external training in standardisation for teaching staff

Of the teachers who responded to the questionnaire, approximately 50% of the secondary school teachers compared with just over 60% of the primary school teachers reported that they had attended external training in standardisation during the past year. For external moderation training, the figures were 59% and 71%, respectively. In this instance, as in many others described in this report, the attitudes and behaviours of primary and secondary teachers appear to differ. For cluster standardisation/moderation training the attendance rates were almost the same for secondary and primary schools (77% and 75%, respectively).

3.4.2.9 Evidence from standardisation portfolios

In the questionnaire, schools were asked three questions about the range of evidence in the standardisation portfolios and the results of these are presented in Table 12.

Table 12: Evidence from standardisation portfolios

		Primary schools		Secondary schools	
Q26: Do all standardisation portfolios have an appropriate range of evidence covering NC levels 1 to 5/levels 3 to 7?	No	2	4%	0	0%
	Yes	25	52%	14	78%
	Developing	21	44%	4	22%
Q27: In your opinion, do standardisation portfolios have appropriate commentaries linking the evidence (i.e. samples of pupil's work) to specific level descriptions?	No	1	6%	1	6%
	Yes	28	58%	9	50%
	Developing	17	35%	8	44%
Q28: Does the evidence and	No	1	2%	1	6%

judgments within standardisation portfolios reflect the shared understanding of NC level descriptions of all KS2/KS3 subject teachers?	Yes	32	69%	12	67%
	Developing	13	29%	5	28%

On the question of whether standardisation portfolios have an appropriate range of evidence covering NC levels (Q26), primary and secondary schools differed in their responses. Only about half the primary schools thought they provided a range of evidence for NC levels and 44% thought this was still developing, whereas 78% of secondary schools thought that they provided sufficient range of evidence for NC levels and 22% judged that this was still developing.

When asked if standardisation portfolios have appropriate commentaries linking the evidence (i.e. samples of pupil's work) to specific level descriptions (Q27), primary and secondary schools responded similarly. Fifty-eight per cent of primary schools and 50% of secondary schools said they did have appropriate commentaries, and 35% of primary schools and 44% of secondary schools thought this was still developing. In the cluster focus groups, mention was made that in some cases there was insufficient commentary in standardisation portfolios, supporting the questionnaire data.

In answer to the question whether the evidence and judgments within standardisation portfolios reflect the shared understanding of NC level descriptions of all KS2/KS3 subject teachers (Q28), primary and secondary schools responded almost the same, with just over two-thirds of the schools agreeing that they did. Almost 30% of the schools thought that this was still developing.

Discussion of the implementation of internal standardisation

Among the schools that responded to the questionnaires, implementation of internal standardisation is strong in the secondary schools but lagging among primary schools. While the majority of secondary schools have standardisation portfolios in place across most subjects, primary schools have much lower levels of standardisation portfolios across subject areas.

Focus-group interviews in the cluster moderation sessions indicated that not all teachers have a robust understanding of the level descriptors, of the range within a level and of the process of standardisation and moderation.

The questionnaire results also indicate that primary schools felt that even where there were standardisation portfolios, they did not always represent an appropriate range of evidence. The lack of standardisation portfolios in primary schools and their poor quality will make it harder for primary teachers to make reliable judgments about pupil work. Although some schools indicated a developing awareness of these aspects, this is some five years since their introduction as statutory elements of assessment. Attendance at training for standardisation could be raised beyond the current level in both primary and secondary schools.

3.4.2.10 Internal moderation

As shown in Table 13, nearly all secondary schools that responded to the questionnaire reported that they were implementing internal moderation across all subjects – oracy, reading and writing (Welsh and English); mathematics; science and non-core subjects. However, primary schools reported that, apart from writing in English and Mathematics, their implementation of internal moderation was low.

Table 13: Percentage of schools that conduct internal moderation across subject areas

Does internal moderation take place in the school for …	Primary schools		Secondary schools	
	n	%	n	%
Oracy in English	24	49%	18	100%
Reading in English	28	59%	18	100%
Writing in English	46	96%	18	100%
Oracy in Welsh	20	41%	17	94%
Reading in Welsh	24	49%	17	94%
Writing in Welsh	31	65%	17	94%
Mathematics	40	84%	18	100%
Science	31	65%	18	100%
Non-core subjects	–	–	16	89%

3.4.2.11 Learner profiles

As shown in Table 14, around 80 to 90 per cent of the responding secondary schools had specific learner profiles in place for all subjects (with science and mathematics being the lowest) and, again, primary schools' implementation of learner profiles lagged behind that for secondary schools apart from in the subject of writing in English.

Table 14: Percentage of schools with specific learner profiles in place

Does the school have specific learner profiles in place for …?	Primary schools		Secondary schools	
	n	%	n	%
Oracy in English	25	53%	17	94%
Reading in English	30	63%	17	94%
Writing in English	44	92%	17	94%
Oracy in Welsh	23	47%	15	83%
Reading in Welsh	24	49%	15	83%
Writing in Welsh	31	65%	15	83%
Mathematics	27	57%	14	78%
Science	26	55%	12	67%

In the questionnaire, schools were asked three questions about the range of evidence in the learner profiles. The results are presented in Table 15.

Table 15: Evidence from learner profiles

		Primary schools		Secondary schools	
		n	%	n	%
Q39: Do learner profiles have appropriate commentaries linking the evidence (i.e. named pupil's work) to specific NC levels?	No	3	6%	0	0%
	Yes	35	73%	11	61%
	Developing	10	21%	7	39%
Q40: Do teachers use the learner profiles to help moderate their overall pupil assessments at the end of KS2/KS3?	No	1	2%	0	0%
	Yes	29	60%	10	56%
	Developing	18	38%	8	44%
Q41: In your opinion does the evidence and judgments within learner profiles reflect the shared understanding of NC levels and standards of all KS2/KS3 teachers?	No	1	2%	0	0%
	Yes	29	60%	12	67%
	Developing	18	38%	6	33%

On the question of whether learner profiles have appropriate commentaries linking the evidence (i.e. samples of pupil's work) to specific level descriptions (Q39), primary and secondary schools responded differently. Seventy-three per cent of primary schools and 61%

of secondary schools said they did have appropriate commentaries, and 21% of primary schools and 39% of secondary schools thought this was still developing.

In answer to the question whether teachers use the learner profiles to help moderate their overall pupil assessments at the end of KS2/KS3 (Q40), primary and secondary schools responded similarly. Sixty per cent of primary schools and 56% of secondary schools agreed that they did, but 38% of the primary and 44% of the secondary schools thought that this was still developing.

When asked if the evidence and judgments within learner profiles reflect the shared understanding of NC levels and standards of all KS2/KS3 teachers (Q41), primary and secondary schools responded similarly. Sixty per cent of primary schools and 67% of secondary schools agreed that they did, while 38% of the primary and 33% of the secondary schools thought that this was still developing.

Attendance at training in moderation

Seventy-one per cent of primary schools and 59% of secondary schools reported that teachers had undergone training in internal moderation within the past year.

Discussion of the implementation of internal moderation

From the responses to the questionnaires, it is clear that secondary schools are implementing internal moderation but implementation in primary schools is very low in comparison. Similarly, secondary schools have learner profiles in place across most subjects, but primary schools have much lower levels of learner profiles in place. Interviews in the cluster focus groups revealed that small primary schools do not have internal moderation procedures in place for mathematics and science. This is due mainly to time constraints. The process does not wholly comply with statutory requirements. These requirements state:

"Headteachers must ensure that for English, Welsh first language or Welsh second language, mathematics and science (Key Stages 2 and 3), and for all non-core subjects (Key Stage 3 only): robust systems and procedures are in place to support accurate and consistent teacher assessment. These systems and procedures need to be focused on internal standardisation and moderation."

Page 6: Statutory assessment arrangements for the school year 2011/12. Key Stages 2 and 3 Guidance document No: 054/2011September 2011

Also the range of evidence in learner profiles could be improved. If primary schools are not engaging in internal moderation, the quality of teacher judgments about pupil work will be inconsistent and lead to a lack of reliability in the system. Attendance at training for internal moderation could be raised beyond the current level in both primary and secondary schools.

Views of the Local Authority Advisors

The responses to the questionnaires by LAs reveal a mixed level of monitoring of their schools assessment policies and procedures. Only one LA of the 18 that responded reported having electronic or paper-based copies of assessment policies from all of the authority's schools. Nine (50%) did not, with the remainder indicating that this was work in progress. Four LAs (22%) reported that they monitor schools' assessment policies to ensure compliance with statutory requirements, whilst five (28%) did not and another four said that progress on this was still developing. Five did not answer that question. On the other hand, 11 of the 18 LAs reported that they provide guidance to schools in developing their assessment policies and practice. All but one of the 18 said that this guidance helps schools to distinguish clearly between formative and summative assessment and between standardisation and moderation.

LAs awareness of their schools' standardisation or moderation arrangements was low. Only three responses out of 18 said that the LA was informed of arrangements for internal standardisation within all of its schools, whilst four (22%) were informed of arrangements for internal moderation. Six LAs (33%) were not informed of these arrangements for either standardisation or moderation.

Ten (56%) reported that they were informed of arrangements for cluster standardisation and moderation within all of their KS2/3 clusters. One admitted that it was not. The remainder were on this journey.

Monitoring of schools' Transition Plans was also limited with nine (50%) of LAs reporting that they have copies of all of their Clusters' Transition Plans, whilst two (11%) do not. Similarly, five (28%) stated that they had ensured that Transition Plans were renewed or updated in September 2010 whilst a further five (28%) said that this was still in train and two (11%) said that they had not done so. The failure of most LAs in South East Wales[13] to respond to the survey (due to recent reorganisation into Consortia) suggests that this key intelligence function may be even more attenuated than these statistics suggest. It should be noted, however, that one response included the views of the five local authorities that make up the Consortium De Ddwyrain Cymru (Southeast Wales EAS) − Blaenau Gwent, Caerphilly, Monmouthshire, Newport and Torfaen.

Only four (22%) LA respondents reported that they ensured that cluster moderation arrangements were the same as described in the current cluster Transition Plan (i.e. as operable from September 2010). A further five (28%) said that they were in the process of doing this, whilst two (11%) admitted to not having done so.

Standardisation and moderation

All LA respondents reported that they were familiar with the contents of the annual DfES guidance *Statutory assessment arrangements for the school year 2011-12*, whilst all but one (94%) said that all LA officers/advisers are familiar with the contents of *Making the most of learning* (DfES, 2008), *Ensuring consistency in teacher assessment: Guidance for Key Stages 2*

[13] Blaenau Gwent, Bridgend, Caerphilly, Cardiff, Merthyr Tydfil, Monmouthshire, Newport and Vale of Glamorgan. Four of these responded in the form of the EAS Consortia questionnaire

and 3 (DfES, 2008) and *Making the most of assessment* (DfES, 2010).The remaining LA reported some awareness of these.

Although all LAs claimed that all of their colleagues understand the difference between standardisation and moderation one was less sure about their understanding of the difference between standardisation portfolios and learner profiles.

Most LAs (88%) have provided some training regarding standardisation during the last year. Two (11%) have not. Fourteen (78%) have organised cluster-based or whole authority meetings for schools in relation to standardisation. Four (22%) have not.

For example,

> "Meetings in 2012/13 focused on end of key stage moderation not standardisation in 2011/12 in English and Welsh. The SAA Guidance for 2011/12 notes that Moderation not Standardisation is expected for Cluster Moderation. Standardisation was left to schools' internal assessment procedures. Considerable training and guidance was given on this between 2006 and 2010 across all core subjects and to support the KS3 Non-Core 'Moderation' Exercise".

> "Most meetings were held after school and training days were also used. For some schools improving assessment procedures was a priority in their SIP and, as such, they had allocated some funding from their own budgets to cover it".

Between half and three-quarters ensure that schools have standardisation portfolios in place for core subjects. Only 3 (17%) reported having such portfolios in place for non-core subjects. All recommend that teachers record pupil attainment on the basis of best fit.

In addition to teacher assessments, 11 (61%) of the LAs who responded also use standardised tests to assess pupils at KS2 and 3. Just over one third of respondents (38%) do not. All Wales Reading Tests for English and Welsh were mentioned by ten LAs. NFER tests are used by three, whilst GL Assessment tests (including the Cognitive Ability Test and the Suffolk Reading Scale) are used by eight LAs. Other (often Local Authority designed) tests are used by four LAs.

All LAs have provided some training around moderation and most LAs (83%) have organised cluster or whole-authority meetings for schools on moderation. These meetings were funded in various ways, with the LA taking a more active role in some than in others. For example,

> "[One LA] used SEG monies to support advisory staff attending cluster moderation meetings – this is under discussion at present for 2012/13 – not organised specifically by LA – the clusters informed the advisory service of dates for their meetings and advisors attend – written reports provided to clusters and LA."

> whilst in another LA,

> "Clusters agreed that each school, primary and secondary, would contribute a percentage of their SEG funding to a 'consortium pot' to enable teachers to be released to moderate learner profiles within the cluster."
>
> in a third LA,
>
> "Clusters are required to set dates for moderation meetings and a LA officer attends the meeting to monitor procedures. Schools fund the meetings."
>
> and a fourth LA,
>
> "Most meetings were held after school and training days were also used. For some schools improving assessment procedures was a priority in their SIP and, as such, they had allocated some funding from their own budgets to cover it."

Most LAs (all but two) ensure that schools have an appropriate range of learner profiles in place for English and Welsh (where this is all but one). Their provision for other core subjects is much lower than this (64%) and even more so for non-core subjects (21%).

Eighty-eight (88) per cent ensure that schools' learner profiles have appropriate commentaries linking the evidence (i.e. named pupil's work) to specific NC levels.

However, only three have complete confidence in the accuracy and reliability of their schools' existing internal moderation procedures. For example,

> "We have many examples of excellent practice in English/Welsh following on from this year's External Moderation Exercise. Gwynedd KS2 and KS3, Anglesey KS3 - Most of our learner profiles were agreed with [majority agreed with fully for English and many for Welsh].These reflect our clusters' best practice. Where practice needs to be further developed e.g. where there have been changes in staffing, particularly at a strategic level or with year 6 teachers this best practice will now influence these developments in individual schools. Target schools have also been identified for further support with moderation at KS2.This reflects the national KS2-3 Moderation findings for English/Welsh where KS3 proved stronger than KS2."
>
> "This has been done through the advisory service and I cannot fully vouch for the robustness of the portfolios – I would like to say yes but in all honesty would not want to be as definite as that."
>
> "Stress is occasionally omitted on the applications skills aspects. This does impact significantly on the levels; often pupils can work mechanically within the range, but do not apply knowledge. If the range is used in isolation to determine an overall best-fit level, it skews the data. If not accurate, then schools are required to re-visit the process."

Training and support

All but one LA responding to the survey have provided training in KS2/3 moderation during the last year. All but one have ensured that cluster moderation takes place for all core subjects and Welsh second language in English medium schools. Almost three-quarters (72%) organise cluster or whole authority meeting for schools in relation to cluster moderation. Those which did not organise such meetings saw this as "a cluster's responsibility – ours is to monitor and hold schools/clusters accountable; also to support clusters on request" or "Headteachers within the clusters organise the meetings. LA officers monitor this." Over three quarters (83%) reported that all cluster schools are appropriately represented at cluster meetings.

Discussion of schools' understanding of National Curriculum assessment

There is clearly considerable variation between LA Advisers in respect of their knowledge and understanding both of the requirements of National Curriculum assessment in Wales and of the way in which this is being implemented in schools. Some believe it to be barely functioning within certain clusters whereas others believe that the experience of external moderation of cluster-based assessment is now so well-embedded for English and Welsh that the focus must move on to other National Curriculum core subjects and higher levels of differentiation.

What is very clear is that LA responses to the questions about teacher assessment do not vary systematically. There is no evidence that all LAs within a particular Consortium (as they are now grouped) will share a common understanding of, still less a common view about, the assessment process as it is operating within their area.

From the observations of the external moderation process conducted as part of this investigation, it became apparent that the role of the local authority/consortium in the process did not seem sufficiently clear to local authorities and was not prioritised enough to ensure accuracy and reliability of the levels awarded at school and cluster levels. This was due to time constraints and other pressures in what is a very demanding role for school improvement/link officers. In addition, local authorities have few subject specific advisors who can input their expertise into improving the work of the DfES in the area of support for teacher assessment or in providing comments and expertise to the WJEC in improving the moderation process for the subjects of the National Curriculum and religious education. The expectation seems to be that this work is now carried out by the chief moderator.

Thirty-nine per cent (39%) of respondents did not have confidence in the National Curriculum level descriptors.

Two questions were asked about resources to support teacher assessment. Two thirds of LA respondents believed that current teacher assessment procedures are an effective use of resources whilst just over half (54%) did not consider the current KS2/3 external moderation programme to be an effective use of resources.

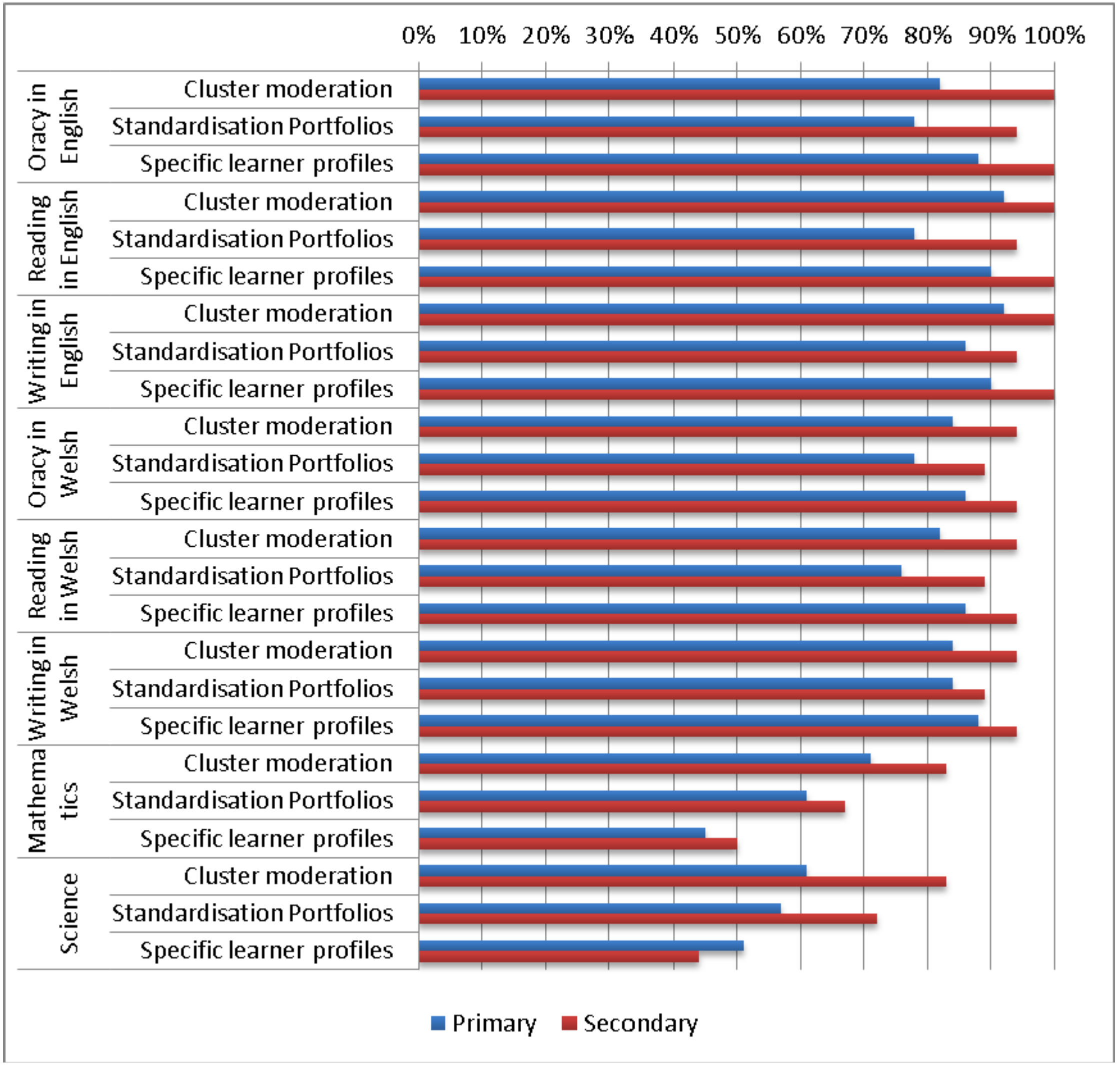
3.4.3 How does the current cluster moderation programme for English, Welsh first language and Welsh second language, work to improve assessment?

To answer the evaluation question that asked "How does the current cluster moderation programme for English, Welsh and Welsh Second Language, work to improve assessment?" the investigation included questions for schools in their questionnaires about the cluster moderation process, a separate questionnaire for cluster coordinators and focus group interviews at a sample of cluster moderation sessions.

3.4.3.1 Views of the schools

As shown in Figure 7, all secondary schools that responded to the questionnaire reported that cluster moderation took place across English oracy, reading and writing with implementation in Welsh oracy, reading and writing at over 90%, and in mathematics and science it was above 80%. In contrast, implementation in primary schools was typically 10–15% lower than in secondary schools. In both sectors, implementation of cluster moderation was lowest in mathematics and science.

Figure 7: Frequency of the implementation by primary and secondary schools of cluster moderation, standardisation portfolios and specific cluster learner profiles across subject areas



As shown in Figure 7, over 90% of secondary schools had cluster standardisation portfolios in place for English oracy, reading and writing, and almost 90% in Welsh oracy, reading and writing. However, in mathematics and science only 67% and 73% of schools had cluster standardisation portfolios in place. Again, primary school implementation trailed behind that in secondary schools by about 10−15% in most subjects. For 93% of the primary schools and 82% of secondary, cluster moderation was an ongoing event, and the remainder did it annually.

Also as shown in Figure 7, all secondary schools that responded to the questionnaire reported that their cluster had implemented specific learner profiles in English oracy, reading and writing, and over 90% said that they had implemented them in Welsh oracy, reading and writing. However, the implementation in mathematics and science was very much lower, with 50% and 44% respectively. Again, implementation of cluster-specific learner profiles in primary schools was about 5−10% lower than in the secondary schools.

Cluster standardisation portfolios

In the questionnaire, schools were asked three questions about the range of evidence in the cluster standardisation portfolios, and the results are presented in Table 16.

Table 16: Cluster standardisation portfolios

		Primary schools		Secondary schools	
		n	%	n	%
Q52: Do standardisation portfolios have appropriate commentaries linking the evidence (i.e. samples of pupil's work) to specific level descriptions?	No	0	0%	0	0%
	Yes	40	84%	13	72%
	Developing	8	16%	5	28%
Q53: Does the evidence and judgments within cluster standardisation portfolios reflect the shared understanding of NC level descriptions of all KS2 and 3 teachers attending the cluster meetings?	No	0	0%	0	0%
	Yes	40	84%	13	72%
	Developing	8	16%	5	28%
Q54: Do cluster representatives take and consider a range of learner profiles to determine the 'best fit' NC levels?	No	1	2%	0	0%
	Yes	40	84%	15	82%
	Developing	7	14%	3	18%

On the question of whether standardisation portfolios have appropriate commentaries linking the evidence (i.e. samples of pupil's work) to specific level descriptions (Q52), primary and secondary schools responded differently. Eighty-four per cent of primary schools and 72% of secondary schools said they did have appropriate commentaries, and 16% of primary schools and 28% of secondary schools thought this was still developing.

In answer to the question whether the evidence and judgments within cluster standardisation portfolios reflect the shared understanding of NC level descriptions of all KS2 and KS3 teachers attending the cluster meetings (Q53), primary and secondary schools responded differently. Eighty-four per cent of primary schools and 72% of secondary schools said they did, and 16% of primary schools and 28% of secondary schools thought this was still developing.

When asked if cluster representatives take and consider a range of learner profiles to determine the 'best fit' NC levels (Q54), primary and secondary schools responded similarly. Eighty-four per cent of primary schools and 82% of secondary schools agreed that they did, while 14% of the primary and 18% of the secondary schools thought that this was still developing.

About three-quarters of primary schools and secondary schools both reported that teachers had attended external training regarding cluster standardisation/moderation during the past year.

Discussion of cluster moderation

The majority of the secondary schools responding to the questionnaire participated in cluster moderation, but for primary schools the participation was lower. Overall, clusters tended to have standardisation portfolios and learner profiles in place, except in mathematics and science. The majority of schools thought that the standardisation portfolios had appropriate commentaries and that the evidence in them reflected shared understanding of the NC levels.

3.4.3.2 Views of the Clusters

Seventeen cluster coordinators completed a questionnaire containing 25 questions. All report that cluster moderation takes place annually and 14 (82%) of the cluster coordinators reported that some cluster staff had attended external training regarding cluster standardisation and moderation during the previous year.

All respondents indicated that the cluster had standardisation portfolios in place for oracy, reading and writing (Welsh and English), but 8 cluster coordinators did not respond when asked about mathematics and 7 did not reply about science. This is probably because, as revealed in comments made on the primary and secondary school questionnaires and in the cluster focus groups, implementation in mathematics and science is lagging behind that of English and Welsh.

Fifteen (88%) of the cluster coordinators felt that these standardisation portfolios had appropriate commentaries linking the evidence (i.e., samples of student work) to specific level descriptions and one other indicated that this was still developing and another did not respond to that question.

Fifteen (88%) of the cluster coordinators confirmed that cluster representatives take and consider a range of learner profiles to determine the best-fit NC levels and one other indicated that this was still developing and another did not respond to that question. All cluster coordinators reported that they had specific learner profiles in place for NC levels 4 and 5 for oracy, reading and writing (Welsh and English), but for mathematics only 6 did and only 8 had them in place for science.

Fourteen (82%) of cluster coordinators reported that, for English, Welsh and Welsh 2nd language in English medium schools, overall subject levels took account of aggregation of separate assessment task outcomes in oracy, reading and writing, and two were still developing that, while one said this was not done.

The cluster focus group interviews showed that the secondary and primary schools sample a range of pupils' work to come to a joint agreement on levels using the best-fit method. This has had benefits in improving transition arrangements and teacher expectation in KS2 and KS3.

Feeder primaries have to attend cluster moderation meetings only if 50% or more of their cohort progress to the secondary school. Where the 50% threshold is not met schools can remain outside of the moderation requirements. Also attendance at cluster moderation meetings is not monitored closely enough to ensure all who have to attend do so. These represent weaknesses in the current system.

3.4.4.3 Views of the Local Authority Advisors

The time allocation and organisation given to cluster moderation during 2011–2012 varied both within and between LAs, as the following comments show:

> "This varied between clusters depending on their maturity with the process and the subject focus. All will have met for at least the equivalent of a day. Many met in excess of this, particularly to ensure that Welsh Oracy evidence (video/audio) was moderated as required for the external moderation for WJEC. Where clusters looked at all 4 subjects up to 4 days were allocated – this is a substantial time investment for our small primary schools where Heads are teaching heads. In some instances afternoon or day sessions were held and then shorter after school meetings to complete the work. Where clusters met several times and we supported these meetings. This demonstrated a significant time investment from the Advisory Team."

> "Each cluster moderation was attended by officers where possible. Clusters organised these sessions using SEG to fund teachers' release. Sessions were moderated by officers in 2012."

> "Varies from cluster to cluster, the initial brief was one day for English and half day each for other subjects but clusters have variations on this practice."

> "Meetings are arranged in Cluster Groups of schools. Usually in autumn/spring term. Half a day session. This has developed over the last 4 years – school becoming more proficient at this now."
>
> "Each cluster spent on average half a day on the moderation of learner profiles in each of the core subjects. An appropriate LA officer attended the meeting."
>
> "Held in the spring term over a period of several weeks. After school meetings and use was made of the two additional training days either side of the Easter holiday."

Just over half (8) of the LAs reported that they ensure that cluster standardisation portfolios have appropriate commentaries linking the evidence (i.e. samples of pupil's work) to specific level descriptions.

Roughly three-quarters said that they ensure that each cluster has specific learner profiles in place for National Curriculum Levels 4 and 5 in English and Welsh. Far fewer (50% and 43%) claimed to ensure the same for Mathematics and Science. This suggests that LAs' commitment to oversight of their schools' assessment practices may vary with the extent of external moderation more generally.

One LA considered that it did not ensure that cluster learner profiles have appropriate commentaries linking the evidence (i.e. a named pupil's work) to specific National Curriculum levels and another did not answer this question.

Recommendation 3 – Impact of the teacher assessment system

The teacher assessment programme in Wales has in place all of the main components that high-quality teacher assessment systems across the world have, but there is room for improvement across the system in the implementation and delivery of the programme. There is a lack of consistency in how it is being implemented and so it is not operating fully as planned. As such the operation of the system is not fully in line with best practice. The areas of operation that require attention are:

3.1 – Schools' understanding of policy

There is a need to ensure that all schools' policy documents on assessment clearly define standardisation and moderation, and distinguish between them as well as describing arrangements for internal standardisation and moderation. Similarly, it should be ensured that school policy clearly describes the arrangements for cluster standardisation. Without the policies having this level of clarity, there is an insufficient base upon which to continue to operate the teacher assessment system and to train school staff in how it should be implemented.

3.2 – Transition plans

Transition plans and actual practice in the system need to be matched, and that match needs to be monitored more effectively.

3.3 – Teachers' understanding of the system

After four years of operation, teachers' understanding of the system and how it should work is still low. National training of teachers across both the secondary and primary sectors is needed to improve understanding, which will set the basis for more full implementation. Training in particular needs to cover the difference between standardisation and moderation, and on making best-fit judgments.

3.4 – Pupil tracking systems

Although standardisation of the pupil tracking systems may not be possible, steps should be taken to ensure that data can be transferred between systems, particularly between KS2 and KS3.

3.5 – Internal standardisation

Steps need to be taken to improve the implementation of internal standardisation within primary schools and to ensure that they have standardisation portfolios in place that represent an appropriate range of evidence.

3.6 – Internal moderation

Steps need to be taken to improve the implementation of internal moderation among primary schools and to ensure that they have learner profiles in place that represent an appropriate range of evidence.

3.7 – Cluster moderation

More primary-school teachers should be participating in cluster moderation meetings. Changes should be made to ensure that no feeder primary schools end up outside of the cluster moderation process as some do under the current 50% threshold rule. Also, attendance at cluster moderation meetings should be monitored closely enough to ensure that all who have to attend do so. Effort needs to be made to ensure that clusters have standardisation portfolios and learner profiles in place for all subjects, especially in mathematics and science that have been ignored until now.

3.5 Operability

The investigation addressed the following questions about the operation of the teacher assessment system:

1. *Do the assessments and the moderation programme represent an effective use of resources?*
2. *How useful are the national curriculum level descriptors?*
3. *What are the barriers to reliable and consistent teacher assessment?*
    a. *What is working well and what is not working as well as desired?*
4. *Is the moderation programme fit for the intended purpose?*
5. *Is the moderation programme assuring reliable and consistent teacher assessment outcomes nationally?*

Effective use of resources

In the questionnaires answered by schools, LAs and cluster coordinators a common question asked whether they considered the current teacher assessment procedures to be an effective use of resources. Seventy-one per cent of primary schools, 78% of secondary schools, 67% of LAs and 77% of clusters considered that the current teacher assessment procedures are an effective use of resources. In answer to a more specific question about whether they considered the current KS2/3 external moderation programme to be an effective use of resources, 60% of primary schools, 61% of secondary schools, 46% of LAs, and 59% of clusters considered the current external moderation programme to be an effective use of resources.

Discussion of effective use of resources

Amongst the schools and LAs that responded to the questionnaires, there is general agreement that the current teacher assessment procedures are an effective use of resources, but about one-quarter of respondents across the education system do not feel that. Over one-third of schools, LAs and cluster coordinators do not view external moderation to be an effective use of resources.

Usefulness of the National Curriculum levels

A question that was posed in the schools, LA and cluster questionnaires asked if they considered the National Curriculum level descriptions useful for assessment.

Overall, the schools agreed that the NC levels descriptions were useful for assessment, with 93% of primary schools and 83% of secondary schools having that view. The LAs and cluster coordinators were less positive, with 62% of LAs and 71% of cluster coordinators having the view that the NC level descriptions were useful for assessment.

Discussion of National Curriculum level descriptors

Overall, the schools that responded to the questionnaires considered the National Curriculum level descriptors to be useful for assessment, although less than two-thirds of the LAs had this view. External moderators who were interviewed felt that the NC level descriptions were useful for assessment and for moderation. However, some descriptions were imprecise and all would benefit from review. Level descriptions would be better represented as *skill ladders* as used for English moderation.

3.5.1 Barriers to reliable and consistent teacher assessment

Each of the questionnaires for schools, cluster coordinators and LAs asked for comments on what respondents considered to be barriers to ensuring reliable and consistent teacher assessment and also asked what they think might improve the quality of pupil assessments. There were several consistent themes among the responses from all sectors, which have identified things that they consider not to be working well and made suggestions for improvement, and these are presented below.

Reliability

A major theme was about the difficulty of achieving consistent teacher judgments across all schools. One primary school said,

> "… the unreliability is understandable in regard of the varying experience of different clusters across Wales. i.e. different interpretations between authorities, within authorities, within clusters, within schools themselves. With all those differing interpretations of what constitutes a particular level, it is little wonder that data are unreliable."

Schools highlighted again the lack of guidance about best-fit judgments and what percentage of work at a level would constitute a fit to that level, and there were comments about the lack of detail in some descriptors. A secondary school commented:

> "There is an uncertainty with some staff as to how much of a particular level must be achieved to secure that level. Even though pupils may be showing characteristics of working at a particular level some are reluctant to award that level …"

At the cluster level, similar comments were made, like this one:

> "Schools have different ideas on what fits a descriptor. There is no parity across schools or LAs. Some are told to give a level when the pupils are only 'dipping their toes in.'"

This is also echoed at the LA level as illustrated by the following comments:

> "Mixed messages that have come from different LAs re best fit, the heavy reliance within the inspection framework on data which can persuade HTs to be generous as

> data will then 'look better'"

> "There are currently too many variables and too much that is open to interpretation. The level descriptors are too broad and general. In standardisation and moderation meetings teachers are still discussing the meaning of phrases within the level descriptors with which they have been working for the last 4 years!"

Another LA felt that teachers were still not familiar with documents already issued, as illustrated in this comment:

> "The level descriptions are too broad. Schools have not fully taken Assessment for Learning on board."

A different LA commented that:

> "Curriculum documentation is not sufficiently detailed or clear. It is possible to interpret some of the descriptions in various ways and no national guidance has been given on the extent of the descriptions within the level needed to achieve a particular level."

Inflation of results

A problem of great concern to several schools was inflation of the results being reported by some schools, and the depth of concern is shown in the detailed comment made by one primary school.

> "… when a school makes its own judgments about the standards attained by its pupils it is possible that in some cases, a degree of 'adjustment' is made. It is somewhat understandable that results are adjusted as schools are faced with annual judgments of their own performance based for the most part on 'raw' data. It is easily forgotten that assessments are not only for schools they are primarily for the pupils; to support their progress and achievement through their education. But this 'dishonesty' makes the process of cluster moderation worthless, a needless expense (and of valuable teacher time) and can lead to cynicism amongst the profession, especially amongst those who feel they have remained honest about the assessments made. Primarily it is an injustice to our pupils."

This is echoed in the comment from another primary school:

> "The competition to be in the upper quartiles is great and the sanctions imposed at all levels is even greater. A punitive accountability system which relies on honest reporting of teacher assessment which is not checked in any way, is also an extremely naive system. I am often surprised by the number of schools who get 100%, this implies that they have no special needs children, odd when the average for a school is considered to be around 20%."

Inflation of teacher judgments was also emphasised by secondary schools. One school felt that there were unintended consequences in the system.

> "The publication of KS3 core data sets may have inadvertently contributed to NC level inflation. Focus for some schools may now be about matching family, local and national averages rather than ensuring accuracy."

Another school commented,

> "Pressure to raise standards, and the publication of performance data which is used as the basis of value added by the LA, Welsh Govt and Estyn. This is forcing the Level 4/5 boundary at KS3 in particular, and confidence in the national picture is eroding. This is manifested in the annual increases in % achieving Level 5 or higher in the Core Data Sets."

Lack of external checking

The absence of any system to check the reliability and validity of teachers' assessment judgments was identified as a problem. One primary school said:

> "The fact that teacher assessment is not externally verified is hugely irritating to say the least. I and many of my colleagues are sceptical about the teacher assessment results in each of our FSM benchmarking groups. We are working so hard at school level to continually improve the quality of teaching and learning in our school, our trends are genuinely improving, however, not always at the 'reported' rate of schools in our FSM benchmarking group. It simply is not a level playing field, we are playing catch up with reported increased percentages of level 4's and 5's that are not verified and make the quartiles an absolute farce. I spend many hours analysing our data in order to identify areas of concern for school improvement purposes and increasingly rely on our internal data rather than comparisons with benchmarking groups – the production of core data is in danger of being a very expensive waste of time. So many colleagues have lost confidence in this 'honesty box' teacher assessment system – time to start paying at the check-out and leaving with a receipt!"

Many respondents identified the need for some form of external checking of teacher judgments to improve reliability and reduce the perceived inflation of judgments at present. As one secondary school remarked:

> "There is no external check / verification of KS3 results and we feel that this is necessary if we are to move to a more robust method where trust and confidence can be restored in KS3 results. A lot of work was done on developing assessment strategies / techniques in 2008−10 e.g. developing tracking systems, target setting, WJEC external moderation, developing standardised tasks and subsequently moderation."

One method suggested is to have external verification visits as explained in this comment by a primary school:

> "If schools knew that there was a strong possibility that they may receive a verification visit then they would feel compelled to be honest (the stakes should be high if they are not!), quartile benchmark thresholds would be in the realm of realism (not fantasy land). In that case, improvements through the quartiles would be reliable and could be celebrated, the schools that genuinely remaining (sic) in the lower quartiles could be identified on the grounds that they really need support to improve rather than the grounds that they are more honest than some of the schools in their FSM group."

A similar comment from a cluster respondent said:

> "External procedures made cluster moderation more thorough. Maybe an external verifier to check levels given would make end of key stage levels more reliable."

There were also calls for some form of standardised testing to provide the check on the convergent reliability of the teacher assessments as expressed in this comment by a secondary school:

> "Standardised tests sat by the whole cohort should, in theory, lead to greater consistency, though the issue of marking would need to be addressed. Maths and Science departments would favour a return to SATs, though the English and Welsh departments did have issues with the quality and reliability of SATS marking."

One LA cautioned that:

> "Numeracy and literacy tests will not or should not be a substitute for teacher assessment but greater control of the moderation process needs to be undertaken to help ensure greater accuracy and confidence in teacher assessment."

There was dismay that the external moderation that does currently exist is to be discontinued. One LA said:

> "It is a shame that this [investigation] is taking place when the rich KS2−3 Moderation process established by WJEC has come to an end. The 18 months spent on the Language External Moderation Exercise at school, LA and National Level has made a significant difference to teacher understanding of assessment, planning and differentiating of tasks (e.g. Level 5 allowed for a focus on More Able Learners at KS2), and demonstrating cluster judgments through commentary, learners' work and contextual information such as resources. A wealth of expertise in assessment at KS2−3 is being lost."

Lack of time

The need for more time to devote to training and implementation was frequently mentioned by respondents. As one cluster put it:

> "The main issue is that time and budget constraints make the process a difficult one. Also, Welsh Assembly Government has now reduced the number of INSET days by two which has also impacted on this statutory area of assessment. SEG budget needs to be increased or a specific budget put in place to allow for standardisation and moderation between colleagues to occur. It is vital that support and guidance from LEA and examining bodies is offered to schools."

Many schools felt that they had insufficient time to devote to the processes necessary to fully implement the teacher assessment system. As one school put it, "It is incredibly time consuming for schools and we're having to recreate profiles every couple of years." Primary schools were concerned by the loss of the two days that had previously been allocated as they felt this would make the task even harder. A secondary school commented on the, "Increasing workload for staff and finding quality time to effectively moderate pupils work."

Staff training and inclusion

Staff turnover was problematic for some schools and schools felt that more training would be beneficial, as well as having the funds to do training. One LA commented:

> "When dealing with a number of schools and changes in staffing, it is inevitable that there will be divergences in terms of understanding of levels."

One school in a rural area that was a member of a small cluster felt that meant that their range of training examples was limited and asked for "more national exemplars of standardisation portfolios." The problems of operating the teacher assessment system in smaller schools was echoed in a comment made by an LA:

> "The model for standardisation and moderation practice both within schools and in clusters is difficult to implement fully when working with many small schools."

A couple of clusters also felt that the fact that not all teachers participated in cluster meetings was a barrier to building their understanding and judgment.

With regard to improving accuracy of teacher judgments, it was suggested that a first step is to ensure that teachers know their subjects well and are used to working as part of a team to share their expertise with colleagues. One respondent felt that this task should begin in teacher training programmes and another identified a need for increased in-service training on assessment, although schools felt that more money and time needed to be devoted to make this happen. Several respondents commented that training should be done on a national basis.

One LA said that what was needed was:

> "Further training based on the national moderation training. Understanding the difference between standardisation and moderation on a national level – this is still unclear from training our national Moderation team in English. Training is also needed on assessing reading – planning appropriate tasks and making best use of the resources available, particularly non-literary reading in English/Welsh."

Respondents expressed a desire for improved training materials. In particular, there were calls for clearer guidelines on what constitutes best fit when making judgments and for there to be moderation portfolios that exemplify best fit in all subjects.

Simplification of the system

Several respondents request a simplification of the current system where there was one key document that specified how the teacher assessment system should operate, rather than several different documents as at present. As one cluster response put it:

> "A national assessment system should be simple – online (with) tick box or comment banks available. This programme should be linked to ready-made assessment material/tasks that are easy to administer and pinpoint pupils' area/s of difficulty and suggest strategies to support their learning to improve. Assessment material should also provide next step learning opportunities and this should link to ready-made resources for teachers to download."

One LA commented on the need for:

> "Clearer guidance within the levels – not so wordy, almost tick list, clear expectation that schools use standardised tests as part of assessment procedures."

Another LA thought that the system could be simplified by not creating profiles annually as explained in this comment:

> "One complaint heard consistently is that schools creating profiles annually is an inefficient use of time. Rather than sending cluster samples to be nationally validated annually, would it not be an idea that national assessors will visit a sample of schools each year to explore their profiles, portfolios and standardised assessments? This would allow direct joint working with schools to ensure consistency. For information process moderated by the Authority as rigorously for CS and it has for KS2/3." (West Wales LA)

Allowing time for systems to "bed down"

One respondent identified a need to, "Keep to one process; too much change too quickly." The downside to a constantly changing system was identified in this comment by one primary school response that said:

> "I have found the amount of effort, time and work that has gone into producing

> cluster portfolios disproportionate to their impact upon teaching and learning, as well as standards. With the introduction of "testing" in Wales, it has also rendered them somewhat dysfunctional. Their development has been of some use in terms of sharing expertise/expectations BUT I undertook this work as a class teacher. It appears to be a repetition of this-I fear the energy it took to complete them as been wasted as the requirements of Welsh Assembly Government seem to be constantly shifting."

Tracking systems

Several respondents commented on the need to improve the tracking of students in the assessment system. One primary school said:

> "Within my school we have a very good pupil tracking system and we regularly review and assess pupil progress (formative and summative). We use a range of evidence to make 'best-fit' judgment annually in relation to NC levels. Since becoming headteacher (3 years ago) staff knowledge of assessment, moderation and standardisation has developed quite considerably. Previously only the Yr 6 teacher had a 'handle' on assessment and cluster moderation."

A cluster respondent said that:

> "The assessment process should be part of pupil record-keeping so that there is no duplication. The system should provide pupil tracking and target setting opportunities so that data can be generated. Assessment system should also generate a simple report for teacher, pupil and parent. At present keeping the annotated assessments for pupils is too time-consuming for staff."

Discussion of what is working and what is not working

The rich and incisive comments made by respondents to the school, LA and cluster questionnaires indicate that stakeholders across the system think deeply about what is working and not working in the current teacher assessment system. They have highlighted the reliability of teacher judgments; inflation of results; lack of external checking; lack of time; staff training and inclusion; simplification of the system; allowing time for the system to bed down; and pupil tracking systems as areas where changes need to be made to improve the effectiveness of the current system.

3.6 Suitability of the moderation programme

The final question addressed in the investigation was, "Is the moderation programme fit for the intended purpose?" In particular, the question, "Is the moderation programme assuring reliable and consistent teacher assessment outcomes nationally?" was considered.

The moderation programme comprises a set of processes that are designed to fulfil the statutory requirements to support reliable teacher assessment and robust school-based assessment systems. In addition to the school based and cluster moderation processes already addressed in this report, DfES supports an external moderation of teacher assessment for English, Welsh and Welsh second language within KS2 and KS3 cluster groups. It is this external moderation process that is considered in answering the questions above.

The external moderation process

Another aspect of the teacher assessment up until this year has been the external moderation process undertaken by WJEC, and part of the investigation involved observations of the external moderation process. The observations during the external moderation conducted by WJEC during April to May 2012 revealed that it was well run as evidenced by the fact that the generic training materials were of good quality and gave appropriate information regarding the cluster moderation process. Also, the documentation contained precise (national) definitions of standardisation, moderation and "best fit" as applied to KS2/3 teacher assessment in Wales. Feedback from participants was generally positive particularly with regard to the exemplification materials used on the day. WJEC provided a wealth of documentation to schools that describes moderation requirements succinctly and clearly guides schools in providing the evidence required for the external process.

An additional requirement for 2011–12 was that submissions should include details of the moderation process within the cluster including a list of contributing schools. WJEC personnel indicated that the primary schools within moderation clusters did not always correspond to local authority designated 'catchment' schools. Some primary schools appeared in more than one moderation cluster while some may not have been listed at all. This information might be usefully employed in identifying those schools *not* engaged in cluster moderation. In addition, it might indicate variation in cluster procedures and correlation between these and external moderation outcomes. However, neither the moderation team nor WJEC personnel have any remit to collate and evaluate this information.

The moderation process for English

The external moderation process for English was well led, well organised and effective in confirming (or otherwise) the NC levels awarded to the learner profiles within the cluster submissions. A whole-day's training preceded actual moderation. Training materials were very good and provided clear guidance for moderators. Training resources included a range of exemplar reports, a report-writing guide and associated comment banks.

There was appropriate reference to the use of NC level descriptions. Usefully, these were presented as *skill ladders* – allowing moderators to identify specific literacy strands to help gauge their judgments. Moderators were instructed to identify examples of good practice within the submissions. Most submitted work appeared to derive from normal classroom practice and a significant proportion of submissions contained aspects of very good practice.

The Chief Moderator and Deputy Chief Moderators provided appropriate support and guidance throughout the process. There were effective quality assurance procedures in place to ensure reliability and consistency of judgments / agreements. These included an end-of-session plenary which provided opportunity for feedback and discussion of issues arising during the day.

Participating moderators were exposed to a range of good practice and became highly skilled in assessing pupils' work. However, there are no current arrangements for employing this pool of expertise at the end of the moderation programme, although there are proposed arrangements to disseminate exemplar work to schools. However, this will depend upon the WJEC's continued involvement in the process which is not secure.

The Chief Moderator and Deputy Chief Moderators indicated that most school clusters had submitted appropriate and sufficient evidence and that documentation was usually well completed with appropriate information regarding the stimulus materials. They noted that teacher commentaries for reading and writing submissions were good but weaker for oracy. Also, primary evidence was usually good but secondary evidence weaker.

While evidence was received from every cluster, not every (primary) school was represented in the submissions. Further, different cluster primaries might submit evidence for different core subjects. Consequently, it is impossible to determine the validity of assessments across clusters, local authorities and Wales as a whole. In addition, while moderators noted variation in the quality of submissions from different areas, it is difficult to substantiate whether the quality of assessment varies on a geographical basis.

While the submitted evidence represented the joint views of all contributory cluster schools, the (very) small sample of evidence provided was insufficient to validate the accuracy and reliability of assessment within all contributory schools. Submissions included information on schools involved in cluster moderation and a description of moderation procedures. The

WJEC currently has no remit to analyse and evaluate this valuable data, which could prove useful at a number of levels.

The moderation process for Welsh First and Second Languages

The external moderation processes for Welsh first and second language were observed to be robust and rigorous. Quality was assured by the requirement that moderators had to be current practitioners with relevant experience in the subjects being moderated. Moderators then underwent a one-day training provided by the chief and deputy chief moderators on the day prior to the start of the moderation process. This training was highly effective and valued greatly by the recipients, who saw it as invaluable professional development that allowed them to undertake the moderation work effectively. Teachers reported that they also made effective use of their improved skills, knowledge and understanding of the assessment and moderation process when back in their schools and clusters to improve the accuracy and reliability of teacher assessments. In Welsh First Language, 16 out of the 24 assistant moderators were undertaking the task for the first time this year.

Great care was taken during the moderation to ensure accuracy in the verification of the levels. Moderators had to make one of three judgments on each piece of work submitted from a cluster – agree with the level; agree but with issues raised; and can't agree. These judgments were accompanied by explanatory comments on tracking sheets as to why there is agreement. Where there was agreement but with issues, guidance was given on what clusters need to improve on in their next year's submission. Comment banks helped moderators in this task. Where agreement on a level could not be reached a detailed commentary was provided explaining why the level cannot be verified. The moderator's comments were returned to the local authority school improvement /subject specialist where the expectation was that the information would be shared and cascaded back to each cluster. Some authorities do this better than others (for example in Welsh Second Language good practice in cluster moderation exists in Torfaen, Blaenau Gwent, Newport, Monmouthshire and Wrexham local authorities). This is as result of these local authorities' cluster moderation exercises being led by the chief and deputy chief moderators for Welsh Second Language.

Cluster moderation of Welsh First Language was undertaken more effectively throughout most of the local authorities as it has been in place for a number of years compared with the more recent Welsh Second Language moderation system. However, without effective national coordination of the whole moderation process, consistency of practice at school/cluster and local authority/consortia level cannot be assured through the present cascade system.

There were rigorous quality assurance processes throughout the moderation. The assistant moderators' work was sampled by the Chief and Deputy Chief moderators on a daily basis to ensure accuracy of levelling and quality of the supporting text. A further quality assurance process of cross-moderation was undertaken by the chief moderator of the work of the deputies and to help resolve any difficult cluster outcomes at the end of each verification exercise. These exercises were very demanding of time and concentration. Time constraints meant that frequently the deputy chief moderators could not moderate and provide support for the chief moderator's own work.

One of the strengths of the moderation system was the pairing of primary with secondary assistant moderators to undertake the verification tasks. It led to informed debate and shared understanding of what constituted reliable and robust teacher assessment and what was a borderline level or a secure level as well as the kind of evidence and comments required when a level cannot be verified.

Moderators/teachers at Key Stage 2 and Key Stage 3 found it very useful and good professional development when working together to verify pupil profiles. Secondary subject teachers often benefited from a greater understanding of the methodology of teaching and learning in primary schools, what motivation or experiences pupils have received in the primary sector especially in regards to developing and reinforcing their literacy and numeracy skills in subjects across the curriculum or when undertaking project work. This helped subject specialists develop their competence and raised their expectations in promoting pupils' key skills of literacy and numeracy through their subject.

However, each moderation exercise is undertaken by a primary and secondary teacher with the expectation that they will complete two cluster reports each day. This productivity expectation is initially challenging for those assistant moderators undertaking the moderation exercise for the first time, though as expertise and familiarity increases during the moderation exercise most are able to complete the two clusters allocated on a daily basis.

Impact of the external moderation

The external moderation produces quantitative data on the levels of agreement between moderators and the teacher judgments, which is useful to highlight areas where the teacher assessment is inconsistent. For example, in the 2012 external moderation of Welsh Second Language, there was only 50% agreement on average across KS2 and KS3 with the teacher assessments, although in Welsh First Language, there was 76% agreement on average with the teacher assessments. These data are the only empirical measures of reliability of teacher judgments provided in the existing scheme.

Interviews with external moderators revealed that they felt that the external moderation programme had impacted positively on teacher assessment. Teachers involved had developed greater expertise and confidence during cluster moderation programmes and in the preparation of the external submissions. Non-literary assessments provided opportunities to develop literacy skills across the curriculum. Also, interviewees observed that teacher assessment was impacting positively on teaching and learning. The learner profiles (as submitted) allowed teachers to identify pupils' strengths and weaknesses and to address the latter in future planning. Teacher assessment had also encouraged the development of engaging, open-ended and cross-curricular tasks. There was an intention to disseminate examples of good practice on completion of external moderation. However, there is currently no mechanism in place to investigate and verify this assertion.

Although the external moderation is already functioning well, there were some weaknesses observed in the moderation process and there are some improvements that could be made to bolster the system.

One improvement is in the timing of the submission of pupil profiles. The requirement for a relatively early submission of pupil profiles from the cluster groups (the end of spring term, March 30th) excluded the whole of year 6 and year 8 pupils' summer term work - arguably the term when pupils have consolidated their learning in the key stage and are able to apply their skills and demonstrate secure performance in the expected levels and crucially at above the expected level – level 5, end of KS2 in primary, and level 6, end of KS3 in secondary.

Also, the assessments were based on a narrow range of pupils' work and generated an attainment level based on one or two pieces of work. This does not necessarily reflect the pupils' true ability as it does not take into account the full range of pupils' work which is required when judging their standards of achievement (as for example during a section 28 Estyn inspection).

Another area for improvement concerns the fact that the guidance from DfES does not stipulate the number of pieces of work required for submission from clusters of schools. Some provided abundant evidence, others the bare minimum. For example, for Welsh language moderation (both Welsh First and Welsh Second Language), schools were required to provide the WJEC with evidence to demonstrate levels attained in three strands; namely, creative writing, factual writing and expressing opinions. Some schools provided separate samples for each of the three areas whilst others submitted two samples where the evidence for all three strands were interwoven. Schools needed to improve their cluster arrangements as to who submits which pieces of work to ensure full coverage of the range of work necessary to achieve a particular level. At present the moderator's task is made more difficult as each strand is sometimes not explicit enough in pupils' work. In such cases, assistant moderators were frequently unable to agree the level awarded - not because of schools'/clusters' lack of understanding of the level descriptors but because the language used in the sample of pupils' work was too narrowly focused and did not reflect the whole range of the level descriptor. In addition, in the weaker samples submitted for moderation at WJEC level by some clusters, the relationship between the learner profile and the commentary provided to justify the level awarded was tenuous. For example, a generic comment such as "there is sufficient evidence to show that the learner demonstrates the ability to provide an opinion" was all the evidence provided and was not explicit enough where and in what context this occurred in the sample provided.

The samples submitted by school clusters for assessment were of variable quality and were not always annotated to evidence clearly why they had awarded a particular level. In several cases observed during the moderation exercise, the school merely quoted the level descriptor without referencing or personalising aspects of the level to the individual pupil's work.

It was also observed that no uniform agreement existed amongst schools and amongst clusters of schools as what was required to comply with a best-fit notion for awarding a particular level. Some schools accepted the bare minimum as the threshold whilst others insisted on awarding a level only when the evidence was secure enough. For example, some schools and clusters awarded a level based on the best-fit model when the sample of work contained some 60% of the level descriptor, whilst others insisted on 70% and some edged towards 90%+ before awarding the level.

There was no requirement for schools/clusters to resubmit their samples/portfolios when issues were raised, or crucially when moderators could not verify the level awarded by the cluster. Schools were free to suggest levels based on their “professional interpretation” which, in some cases, may be wide of the mark. Some schools did this in the knowledge that there was no comeback, retribution or further scrutiny of the quality of their teacher assessment process and outcomes. The most they would receive was feedback in their cluster meeting from the local authority link officer about what needed to improve for next year’s cluster submission. Some Estyn section 28 inspections had alluded to ”over-generous” teacher assessments when inspectors found a considerable mismatch between the standards of pupils’ attainment from the previous cohort with the standards of achievement as evidenced in the current work in pupils’ books.

It was noted that there was minimal impact arising from the local authority cascading back the moderators’ comments about improvements looked for in next year’s submission. The process was not monitored effectively enough to ensure consistency and accuracy of levels awarded and schools could continue to over- or under-level pupils’ work. The process had no teeth and hence no real impact. The level which schools and clusters awarded for the sample initially is the level that appeared in the Fynnon data.

The WJEC required the chief moderators for Welsh First and Welsh Second Language to identify and record case studies of good practice in school/cluster moderation. However, these were not published at present but merely retained by the WJEC as a resource should they be asked to produce good practice/exemplification of standards guidance in future. This was in line with the WJEC confidentiality agreement with schools, that they will use the samples provided to demonstrate issues in awarding levels to pupils’ work for training purposes only and not for publication.

Another potential improvement is related to the fact that chief moderators and their deputies are seconded from their substantive posts for the period of moderation only. The part-time nature of these roles leads to intense pressure on them and the deputy and assistant moderators to complete the moderation tasks within the relatively short time scale provided. Such pressures and time constraints can lead at times to errors in the process. In addition moderators for Welsh Second Language produce materials in a bilingual format for training and support and when communicating with schools. It is felt that the time allocated for this work is insufficient at present.

Discussion of fitness for purpose

Our evidence suggests that the external moderation programme is fit for the original purpose for which it was designed, that is to monitor the quality of judgments in the teacher assessment system. The external moderation produces quantitative measures of agreement between moderators and the teacher judgments so that areas of weakness can be identified and remedied in future.

The processes are of high quality and serving a valuable direct purpose in monitoring the quality of teacher judgments and an equally important secondary purpose of developing teachers' expertise in those judgments. This is in line with findings from the literature review, which concluded that moderation meetings are seen to be powerful professional development and a networking opportunity for those involved although the primary purpose of moderation is to ensure comparability of results. In particular, the process develops a pool of individuals (moderators) with significant skill and expertise in assessment. However, there are no national or regional plans/models in place to implement and employ moderators' skills at this time and there is a danger that this developed and practised expertise will disappear as new and different pressures on teachers and their schools come into play.

Recommendation 4 – Operability

Given that there is a need to increase the reliability of the teacher assessment system it would be wise to retain the external moderation scheme and extend it to other core subjects, as planned. It is a mechanism for ensuring that teacher judgments across the system are consistent and one of the few places where data are being gathered on reliability of those judgments at present.

While the phenomenon of teachers being unable to envision student work that matches a level description is not unusual in standards-based assessment, a shared understanding of standards in writing and in application is a cornerstone of moderated teacher assessment. There is a need to revise the level descriptions using a combination of curriculum experts, assessment experts, pedagogues, and editors with skills in instructional language and clear expression.

While national in-service programmes for teachers can be costly in both financial and human terms, teachers must not be subjected to conflicting messages from within the system and/or between schools and clusters. The costs involved in nation-wide professional development would be recovered in the future because less time and energy would be expended in dealing with anomalies of interpretation and/or the consequences of colleagues inadvertently exchanging inaccurate second-hand information.

4. Conclusions and recommendations

The investigation used a mixture of quantitative and qualitative research methods to examine the current system of teacher assessment in Wales. It addressed a range of research questions in four areas of concern that focused on the reliability of the assessments, their validity, the impact that the system is having, and the operation of the system. The questions posed are addressed below.

Reliability

1. *What works in terms of securing accurate teacher assessments?*

Reliability is generally defined in terms of consistency across assessments or "the extent to which assessment can be trusted to give consistent information," and in the context of school-based standards-referenced assessment, reliability is related to consistency of teacher judgments and the comparability of reported results.

There must be some measure of understanding about consistency of teacher judgments and comparability of reported results if the results of school-based assessment are seen to be reliable in the sense of being dependable. Obtaining reliability in teacher assessment systems involves the use of a standards schema, criteria/standards matrix, grid or grading master that teachers apply for making decisions about the standard or level of student work.

The system in operation in Wales is a social or consensus moderation scheme, which is the most common form of moderation for school-based assessment systems. Social or consensus moderation is a quality assurance process that brings teachers together to review and discuss judgments across examples of student work, often in different assessments, and reach some level of agreement about the application of standards to that work. Such systems can produce levels of reliability comparable to other forms of assessment, but mechanisms need to be in place to ensure consistency of teacher judgments.

A recurring theme in the literature on teacher assessment is the level of expertise required of teacher–assessors. The skill sets that can be identified as necessary for a reliable system are (a) expertise in their disciplines, (b) total immersion in their level statements, (c) the ability to evaluate, and (d) the ability to negotiate when seeking consensus at moderation meetings.

2. *Do the assessments demonstrate reliability?*

   a. *What is the reliability and consistency of judgments made in the KS2 and KS3 assessment and levelling?*

   b. *How does the KS2 system that is focused on the school cluster level compare to the KS3 system?*

As it was beyond the scope of this investigation to perform a full-scale measurement and analysis of the reliability of the judgments in the current system—(a) existing pupil assessment records, (b) internal standardisation procedures, (c) internal moderation procedures, and (d) cluster moderation procedures—measures of the confidence of the primary-school headteachers, secondary-school headteachers, Local Authority advisers and cluster coordinators were obtained from extensive questionnaires sent to each sector.

Confidence in the accuracy and reliability of all four components of the teacher assessment system is not very high. Primary schools expressed the most confidence in the reliability but even then, only 50−60% expressed this view. In contrast, about 80% of local authorities expressed some confidence in the accuracy and reliability of the system. These results point to a problem with the reliability of teacher judgments in the assessment system. Possible causes for any reduction in reliability and ways that they might be addressed are described in other parts of the investigation.

The current system does not systematically gather data on the reliability of the judgments made by teachers. Until methods by which reliability is measured are available, it is impossible to know how well the system is doing in regards to the consistency of teacher judgment. It is recommended that measures of the actual reliability of judgments be embedded into the teacher assessment systems so that it can be tracked and, if necessary, the implementation of the system can be refined to optimise the reliability.

Consideration should be given to the following methods that are employed in other teacher assessment systems around the world to determine which might be suitable to introduce in Wales to enhance the reliability. Methods include:

For establishing reliability

Immersion

In an “immersion” process teachers study samples of student work to locate instances of the desirable features of work at a particular level rather than being given examples of student work at that level. An advantage of this approach is that it addresses some of the issues arising from the traditional and comparatively passive approach whereby teachers have to confront exemplars for which trading off has already occurred, which is especially problematic for the middle levels.

Marking rubrics

One of the ways to improve reliability in the marking process is to use marking rubrics. Marking rubrics are descriptive marking schemes developed by teachers or other assessors to guide the analysis of the products or processes of students’ efforts. The descriptors should not to be too wordy, and they must convey meaning with clarity and precision.

For monitoring reliability

Marker monitoring for tests of written expression

The marker monitoring process involves the comparison of many different pairings of markers on the particular responses they have both marked in order to identify markers who are discrepant with other markers, with the purpose of re-calibrating them to enhance their accuracy.

Marker monitoring for constructed-response items

Another method for improving reliability of teacher judgments in the marking of constructed responses is to check if the differences between the grades assigned to a student's response to an item by a pair of markers are within the tolerance or random. The tolerance level for each item is set with due regards for grade range for the item, the nature of the item, and the marking scheme.

Inter-marker agreement

Inter-marker agreement, as represented by two measures, concordance and weighted kappa, has also been used as a tool in the quality control procedures in both written expression and short-response.

Paired comparisons

In the method of paired comparisons (David, 1988), objects are presented in pairs to one or more judges. The method is used primarily in cases when the objects to be compared can be judged only subjectively; that is to say, when it is impossible or impracticable to make relevant measurements in order to decide which of two objects is preferable.

Post-hoc consistency checks

Random sampling is designed to provide feedback and research data about consistency across different marking sessions. Sample folios of student work are selected and distributed to panel members who undertake a review in much the same way as they would as part of the routine quality assurance processes.

Comments on establishing and monitoring reliability

No mechanism, of itself, can guarantee reliability in teacher assessment. Reliable assessment only occurs after large-scale implementation strategies, or experience over time, or a tacit understanding amongst the practitioners. Only the first-mentioned of these is a transparent mechanism for disseminating standards. It is important, however, to recognise the second of these (experience over time) when making a realistic evaluation of an evolving system. Vital, therefore, are (i) teacher professional development (including in-built professional development of the type described in "immersion" above) and (ii) ensuring that marking rubrics, whatever form they take, provide teachers with a simple structure for assessment, written in such a way that multiple interpretations of the standards are not likely to occur; rather, that the intended standards are the applied standards.

Recommendation 1: Reliability

The current system does not systematically gather data on the reliability of the teachers' assessments within clusters. External moderation of assessments depends upon the choice of portfolios submitted by clusters. Until methods by which reliability is measured are available, it is impossible to know how well the system is doing in regards to the consistency of teacher judgment. It is recommended that measures of the actual reliability of judgments be embedded into the teacher assessment systems so that it can be tracked and, if necessary, the implementation of the system can be refined to optimise the reliability.

Validity

1. *Do the current assessments accurately reflect the actual ability of the learner?*

Responses to a question about whether LAs, cluster coordinators, and the secondary and primary schools considered the current assessments to accurately reflect the actual ability of the learners revealed different views from each sector. Primary schools had most confidence in the accuracy with 93% feeling that the assessments were an accurate reflection of student ability. This drops to 75% confidence in the secondary schools and to 50% at the LA level. Cluster coordinators had 87% confidence, which is somewhere between the levels of confidence in the primary and secondary schools. In addition, there were complaints from high schools that pupils' functional literacy on entry to these secondary schools in KS3 did not always correspond to the NC level they were awarded, particularly in relation to English and Science. This erodes the validity of the assessment because it casts doubt on the judgments being made by teachers.

2. *What are the threats to validity of the current system?*

Although there are many views on validity of assessment systems, they are all underpinned by fundamental notions about establishing the appropriateness of assessments; namely, that assessments should (a) measure what they purport to measure, (b) demonstrate predicted relationships with other measures of the intended constructs, (c) contain content consistent with their intended uses and (d) be put to purposes that are consistent with their design and are supported by evidence.

A threat to the validity of the system was emerging from the fact that, with schools being compared to one another, there was pressure on them to raise their pupils' performance. This pressure seems to be leading to an inflation of the teacher judgments of pupils' work and this, in turn, results in a lack of confidence that a judgment is accurate.

Another threat to validity identified was the use of taped oral evidence and other "less robust evidence of pupil attainment", which secondary teachers felt disadvantaged pupils by not preparing them for KS4 where pupils undertake externally assessed examinations based on writing and comprehension only.

Other teacher assessment systems around the world employ different strategies to enhance the validity of the system, and consideration should be given to which of them might be suitable for introduction to the Welsh system. The methods include:

Internal moderation

Internal moderation is a form of peer review whereby teachers of the same subject in the same schools meet to share and discuss student work and provide feedback to each other about the way standards have been applied. Internal moderation is already implemented in the Welsh system although, as reported elsewhere in this investigation, it is not being implemented consistently.

Alignment of assessment with curriculum and pedagogy

Valid assessments are aligned with curriculum and pedagogy. Most teacher assessment systems establish validity by using assessment techniques that reflect classroom experiences, not only in assessment format but also by allowing unlimited time (within reason), computer-generated text as well as or instead of handwriting, and so on. Wales does allow for diverse forms of evidence in the assessment process.

Construction matrix

A construction matrix is intended to ensure a range and balance of items and tasks across the portfolio of pupil work that is assessed. Range and balance can be represented by a construction matrix or grid in which characteristics of the assessment instruments are tabulated, characteristics such as perceived difficulty, estimated time for completion of instrument, curriculum element(s) or objective(s) being assessed, and nature of the text that dominates in the instrument (e.g., verbal, numerical, spatial) to name a few.

Face validity

Face validity, while based on opinion rather than facts, is particularly important in new assessment systems. The opinions of parents, the general public and government cannot be overestimated for the ultimate success of an initiative. As identified above and elsewhere in this report, the face validity of the teacher assessment system in Wales is under threat because of perceptions of components that are seen to not be working.

Panelling

Panelling (or reviewing) is primarily a validation exercise often used in a test development cycle. Experts work collaboratively in small groups, both at the item level and the test level, to review the features of a test that can affect validity. Although panelling for teacher assessment is not widespread, getting a range of opinions from different types of expert can add validity to the assessment.

Accreditation of assessment tasks

Assessment instruments (or at least blue-prints of them) that are designed by teachers are submitted to an external panel for approval before administration to pupils in order to alleviate the need for special consideration at the end of the assessment programme should the assessments be then deemed invalid (say in content or difficulty). The purpose is to ensure that pupils are not disadvantaged because the assessments (possibly set by inexperienced teacher−assessors) are not capable of bringing forth evidence of learning.

Statistical evidence

Some teacher assessment systems convene a technical panel to study and evaluate data relating to instances of possible bias against sub-groups of the population (differential item functioning). The purpose is to measure whether different sub-groups defined by gender (or other indicator of interest) and ability level differ systematically in their performance on the assessment. If they do, the assessment may be biased in favour or against a particular sub-group.

Recommendation 2: Validity

There is evidence that the face validity of the current system is already under threat because some schools have lost trust in the judgments made in the teacher assessment process. While face validity is not the only form of validity, it is important that those who need to operate the teacher assessment believe that its outcomes are valid. Without that confidence in the system, it will falter. To restore confidence Recommendation 3 on the impact of the teacher assessment system and Recommendation 4 on the operability of the system that are contained in this report should be implemented.

Impact

1. *Is the implementation and delivery of the moderation programme in line with best practice?*

    a. *How do the systems at KS2 and KS3 compare to alternative systems that have been shown to work effectively in other education systems around the world?*

The Welsh teacher assessment system contains the five components that have been identified in the research literature as being the essential elements of effective teacher assessment systems as identified by Allen (2003). However, as demonstrated in the literature, having all of the requisite components is important, but not sufficient to ensure the success of the system in practice. When considering the list of necessary elements in the process of assessment (Meiers et al. 2007 and Gipps, 2002), the implementation of the

Welsh system is not yet meeting best practice, based upon the evidence gathered in other parts of this investigation.

This situation, however, is not surprising since the review of literature demonstrates that introducing such teacher assessment systems requires considerable change across schools and local authorities because it has implications not only for the assessment, but also for pedagogy.

As is being experienced in the Welsh system, one of the most challenging features is ensuring the quality and consistency of teacher assessments of student work. It is evident from the experience of countries that are further along the path than Wales that it takes years and several iterations of the process to develop the levels of teacher skills and experience to achieve the desired levels of reliability.

Even in the system that has been running in Queensland for 20 years, which achieved high degrees of reliability of scoring and widespread acceptance by schools and universities, there is a culture of continuous theory-building. In comparison, the Welsh system is still in its infancy, having been in existence since 2000, but only since 2004 with the external moderation component. The evidence from this investigation is that there are certainly areas in the system where improvements can be made, but given the history of the development of other systems this is to be expected. The Queensland case study shows that high reliability is ultimately achievable with sustained training of teachers and continued refinement of processes.

2. *Are the assessment and levelling procedures being implemented as planned?*

    a. *Are schools actually following the procedures as designed? If not, why not?*

Deficiencies in the implementation of the system are the source of the lack of reliability in teacher judgments and of the threats to validity within the system. Most areas of implementation within the system could be improved upon, although some more than others. Although the sample sizes of responses to the questionnaires that were used to gather evidence in the study are lower than desired, the evidence for the things identified as needing attention came from across the questionnaires for all sectors and from focus groups and interviews as well.

Overall, implementation at the secondary school level seemed to be more in line with the planned scheme, while primary schools lagged in their implementation. In particular, there seems to be a weakness in policy documentation in a proportion of the primary schools, which is of concern given that it is four years since the statutory introduction of the policy documents.

The levels of understanding of the teacher assessment system in both school sectors is less than ideal, with some schools indicating that teachers were still unsure of the difference between standardisation and moderation. In particular, schools reported low levels of familiarity with DfES publications. There are problems with the current scheme of training that assumes that once some teachers are trained, they will go back to their schools and 'cascade' what they had learned to other teachers. In reality, this was not implemented evenly.

Generally, the tracking of pupil attainment data across subject areas seems adequate, but there is a range of different pupil tracking systems in use across the primary and secondary schools, which can lead to lack of compatibility that hampers transfer of data across schools.

Primary schools gave more weight than secondary schools to standardised test data when they make best-fit judgments on NC levels.

Although internal standardisation seemed strong in the secondary schools, it was lagging among primary schools. While the majority of secondary schools have standardisation portfolios in place across most subjects, primary schools have much lower levels of standardisation portfolios across subject areas.

Not all teachers have a robust understanding of the level descriptors, of the range within a level and of the process of standardisation and moderation. Primary schools felt that even where there were standardisation portfolios, they did not always represent an appropriate range of evidence. The lack of standardisation portfolios in primary schools and their poor quality will make it harder for primary teachers to make reliable judgments about pupil work. Although some schools indicated a developing awareness of these aspects, this is some five years since their introduction as statutory elements of assessment.

Secondary schools are implementing internal moderation but implementation in primary schools is very low in comparison. Similarly, secondary schools have learner profiles in place across most subjects, but primary schools have much lower levels of learner profiles in place. Interviews in the cluster focus groups revealed that small primary schools do not have internal moderation procedures in place for mathematics and science. This is due mainly to time constraints. The process does not wholly comply with statutory requirements.

Also the range of evidence in learner profiles could be improved. If primary schools are not engaging in internal moderation, the quality of teacher judgments about pupil work will be inconsistent and lead to a lack of reliability in the system.

From the observations of the external moderation process conducted as part of this investigation, it became apparent that the role of the local authority/consortium in the process did not seem sufficiently clear to the local authorities and was not prioritised enough to ensure accuracy and reliability of the levels awarded at school and cluster levels.

3. *How does the current cluster moderation programme for English, Welsh and Welsh Second Language work to improve assessment?*

The majority of the secondary schools participate in cluster moderation, but for primary schools the participation was lower. Overall, clusters tended to have standardisation portfolios and learner profiles in place, except in mathematics and science. The majority of schools thought that the standardisation portfolios had appropriate commentaries and that the evidence in them reflected shared understanding of the NC levels.

Although the external moderation is already functioning well, there were some weaknesses observed in the moderation process and there are some improvements that could be made to bolster the system. One improvement is in the timing of the submission of pupil profiles. Also, the assessments were based on a narrow range of pupils' work and generated an attainment level based on one or two pieces of work. Another area for improvement concerns the fact that the guidance from DfES does not stipulate the number of pieces of work required for submission from clusters of schools.

The samples submitted by school clusters for assessment were of variable quality and were not always annotated to evidence clearly why they had been awarded a particular level. It was also observed that no uniform agreement existed amongst schools and amongst clusters of schools as to what was required to comply with a best-fit notion for awarding a particular level.

There was no requirement for schools/clusters to resubmit their samples/portfolios when issues were raised, or crucially when moderators could not verify the level awarded by the cluster. Schools were free to suggest levels based on their "professional interpretation" which, in some cases, may be wide of the mark.

It was noted that there was minimal impact arising from the local authority cascading back the moderators' comments about improvements looked for in next year's submission. Another potential improvement is related to the fact that chief moderators and their deputies are seconded from their substantive posts for the period of moderation only.

Recommendation 3: Impact of the teacher assessment system

The teacher assessment programme in Wales has in place all of the main components that high-quality teacher assessment systems across the world have, but there is room for improvement across the system in the implementation and delivery of the programme. There is a lack of consistency in how it is being implemented and so it is not operating fully as planned. As such the operation of the system is not fully in line with best practice. The areas of operation that require attention are:

3.1 Schools' understanding of policy

There is a need to ensure that all schools' policy documents on assessment clearly define standardisation and moderation, and distinguish between them as well as describing arrangements for internal standardisation and moderation. Similarly, it should be ensured that school policy clearly describes the arrangements for cluster standardisation. Without the policies having this level of clarity, there is an insufficient base upon which to continue to operate the teacher assessment system and to train school staff in how it should be implemented.

3.2 Transition plans

Transition plans and actual practice in the system need to be matched, and that match needs to be monitored more effectively.

3.3 Teachers' understanding of the system

After four years of operation, teachers' understanding of the system and how it should work is still low. National training of teachers across both the secondary and primary sectors is needed to improve understanding, which will set the basis for more full implementation. Training in particular needs to cover the difference between standardisation and moderation, and on making best-fit judgments.

3.4 Pupil tracking systems

Although standardisation of the pupil tracking systems may not be possible, steps should be taken to ensure that data can be transferred between systems, particularly between KS2 and KS3.

3.5 Internal standardisation

Steps need to be taken to improve the implementation of internal standardisation among primary schools and to ensure that they have standardisation profiles in place that represent an appropriate range of evidence.

3.6 Internal moderation

Steps need to be taken to improve the implementation of internal moderation within primary schools and to ensure that they have learner profiles in place that represent an appropriate range of evidence.

3.7 Cluster moderation

More primary-school teachers should be participating in cluster moderation meetings. Changes should be made to ensure that no feeder primary ends up outside of the cluster moderation process as some do under the current 50% threshold rule. Also, attendance at cluster moderation meetings should be monitored closely enough to ensure that all who have to attend do so. Effort needs to be made to ensure that clusters have standardisation portfolios and learner profiles in place for all subjects, especially in mathematics and science that have been ignored until now.

Operability

1. *Do the assessments and the moderation programme represent an effective use of resources?*

The majority of respondents to the questionnaires agreed that the current teacher assessment procedures are an effective use of resources, but about one-quarter of respondents across the education system do not feel that. This may change for the better if the assessment processes identified as needing improvement in other parts of this report are addressed. Over one-third of schools, LAs and cluster coordinators do not view external moderation to be an effective use of resources.

2. *How useful are the national curriculum level descriptors?*

While the majority of schools considered the National Curriculum level descriptions useful for assessment, less than two-thirds of the LAs agreed. It is unclear why they have a different view to the schools. External moderators also viewed the NC level descriptions as useful for assessment and moderation, but they felt that some descriptions were imprecise and all would benefit from review.

3. *What are the barriers to reliable and consistent teacher assessment?*

   a. *What is working well and what is not working as well as desired?*

Many of the schools, LAs and cluster coordinators made written comments in their questionnaires to give views on many aspects of what they felt were barriers to reliable and consistent teacher assessment and contributed ideas about ways to improve the system. Themes that emerged from their comments were the reliability of teacher judgments; inflation of results; lack of external checking; lack of time; staff training and inclusion; simplification of the system; allowing time for the system to bed down; and pupil tracking systems as areas where changes need to be made to improve the effectiveness of the current system.

4. *Is the moderation programme fit for the intended purpose?*

   a. *Is the moderation programme assuring reliable and consistent teacher assessment outcomes nationally?*

The moderation programme has satisfied a clear and worthwhile purpose at system and school level and within the wider school community. Its design − consensus moderation − is in line with effective moderation models used elsewhere . The quality of the programme, as opposed to its fitness, can and should be improved in ways described throughout this report.

Experience from elsewhere demonstrates the added value of an external component. Barriers to reliable and consistent teacher assessment outcomes nationally include the lack of a national training exercise for teachers and the verbosity and/or vagueness of level descriptors, both of which detract from the aim of reliability and consistency.

Recommendation 4: Operability

Given that there is a need to increase the reliability of the teacher assessment system it would be wise to retain the external moderation scheme and extend it to other core subjects, as planned. It is a mechanism for ensuring that teacher judgments across the system are consistent and one of the few places where data are being gathered on reliability of those judgments at present.

While the phenomenon of teachers being unable to envision student work that matches a level description is not unusual in standards-based assessment, a shared understanding of standards in writing and in application is a cornerstone of moderated teacher assessment. There is a need to revise the level descriptions using a combination of curriculum experts, assessment experts, pedagogues, and editors with skills in instructional language and clear expression.

While national in-service programmes for teachers can be costly in both financial and human terms, teachers must not be subjected to conflicting messages from within the system and/or between schools and clusters. The costs involved in nation-wide professional development would be recovered in the future because less time and energy would be expended in dealing with anomalies of interpretation and/or the consequences of colleagues inadvertently exchanging inaccurate second-hand information.

Locating the consensus model within four main approaches

KS2 and KS3 appear to demand different approaches depending not only on fitness for intended purpose but also on the required degree of control and high/low-stakes nature of the programme. According to what Maxwell (2006) considers to be the advantages and disadvantages (or difficulties) of four different ways of going about social moderation, the approaches for KS2 and KS3 are located somewhere between the category “Assessor meetings” and the category “External moderation panels” (see Appendix 5, which incorporates many of the issues discussed throughout this report).

5. Conclusion

The picture that has emerged from this investigation is of a teacher assessment system that has the main components of successful systems elsewhere in the world. It has adequate levels of documentation about how the system should work and the responsibilities of the key participants in the process. However, the implementation of teacher assessment is an enormous task and there are many parts that must be functioning smoothly for the system to produce reliable teacher judgments and effective educational outcomes. The system has not yet achieved that level of functioning and it still requires some attention to the parts of the system that are not operating as designed. This will require careful scrutiny of the system and consultation with those involved in its operation. One thing that has become evident from this investigation is that the problems that need to be fixed are already known by many in the system, and what is more is that they have ideas for how to solve them.

References

Agresti, A. (1990). Categorical Data Analysis. New York: John Wiley & Sons.

Allen, J. R. (1987). Continuous quality control of written expression marks – a new technique. Paper presented at the annual conference of the International Association for Educational Assessment, Sydney.

Allen, J. R. (2003). Personal communication to the Director Assessment & New Basics, Queensland Department of Education, Brisbane.

Bernstein, B. (1990). The structuring of pedagogic discourse. London: Routledge & Kegan Paul.

Broadfoot, P., & Black, P. (2004). Redefining assessment? The first ten years of assessment in education. Assessment in Education: Principles, Policy & Practice, 11 (1), 7–26.

Brookhart, S. M. (1999).The art and science of classroom assessment, ASHE–ERIC Higher Education Report, 27 (1).

Brookhart, S. M. (2012).The use of teacher judgment for summative assessment in the USA. Assessment in Education: Principles, Policy and Practice, DOI: 10.1080/0969594X.2012.703170.

Brown, G. T. L. (2011). School-based assessment methods: Development and Implementation. Invited paper for First International Educational Conference on Assessment, New Delhi, January 2011.

Brown, G.T. L., Lake, R. I. E., & Matters, G. N. (2011). Queensland teachers' conceptions of assessment: The impact of policy priorities on teacher attitudes. Teaching and Teacher Education, 27, 210–220.

Canal, L., Bonini, N., Micciolo, R., &, Tentori, K. (2012). Consistency in teachers' judgments. European Journal of Psychology of Education, 27(3), 319–327.

Chatterji, M. (2003). Designing and using tools for educational assessment. Boston: Allyn & Bacon.

Crisp, V. (2010).Towards a model of the judgment processes involved in examination marking. Oxford Review of Education, 36(1), 1–21.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer (Ed.), Test validity. Hillsdale, N.J.: Lawrence Erlbaum.

David, H. A. (1988). The method of paired comparisons. New York: Oxford University Press.

Elwood, J. (2006). Formative assessment: possibilities, boundaries and limitations. Assessment in Education: Principles, policy and practice, 13(2), 215–232.

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.

Gipps, C. (2002). The Report of the Assessment and Reporting Taskforce. Brisbane: Department of Education.

Gipps, C. V. (1994). Beyond Testing: Towards a Theory of Educational Assessment. London: The Falmer Press.

Greatorex, J. Unable to be located by authors

Harlen, W. (2004). A systematic review of the evidence of reliability and validity of assessment by teachers used for summative purposes (EPPI-Centre Review), Research Evidence in Education Library, issue 3 (London, EPPI-Centre, Social Science Research Unit, Institute of Education).

Harlen, W. (2005). Teachers' summative practices and assessment for learning – tensions and synergies. The Curriculum Journal, 16(2), 207–223.

Harris, K. M. (2000). QCS Yearbook. Brisbane: Board of Senior Secondary School Studies.

Harris, K M., Kelly, D. J., & Matters, G. N. (2004). Putting Rich Tasks to the test. From a collection of research activities, New Basics Research Report. Brisbane: Department of Education and the Arts.

Hay, P., & MacDonald, D. (2008). A process for making reliable and consistent judgments. PDHPE, 11, 4–5.

Johnson, M., & Burdett, N. (2010). Intention, Interpretation and Implementation: some paradoxes of assessment for learning across educational contexts.

Klenowski, V. (2005). Evaluation of the effectiveness of the consensus-based standards validation process. Brisbane: Queensland Government.

Linn, R. L. (1993). Linking results of distinct assessments. Applied Measurement in Education, 6, 83-102.

Mansell, W., James, M., & the Assessment Reform Group (2009). Assessment in schools: Fit for purpose. Commentary by the teaching and learning research programme.

Masters, G. N., & McBryde, B. (1994). An investigation of the comparability of teachers' assessment of student folios. Brisbane: Queensland Tertiary Procedures Authority.

Matters, G. N. (2005). Good Data, Bad News, Good Policy Making. http://research.acer.edu.au/research_conference_2005/5

Matters, G. N. (2005). The grading master: a simpler way. EQ Australia The Assessment Agenda, Issue 2. Winter 2005, 12–15.

Maxwell, G. S. (2006). Quality management of school-based assessments: Moderation of teacher judgments. Paper presented at International Association for Educational Assessment (IAEA) 32nd Annual Conference, Singapore, 21–26 May 2006.

Maxwell, G.S. (2010). Moderation of student work by teachers. International Encyclopedia of Education, 3, 457–463.

Meisels, S. J., Bickel, D. D., Nicholson, J., Xue, Y., & Atkins-Burnett, S. (2001). Trusting teachers' judgments: A validity study of a curriculum-embedded performance assessment in kindergarten to Grade 3. American Educational Research Journal, 38(1), 73–95.

Meiers, M., Ozolins, C., & McKenzie, P. (2007). ACER Report: Improving Consistency in Teacher Judgments – An investigation for the Department of Education, Victoria.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. Review of Educational Research, 62(3), 229–258.

Moss, P. A. (1994). Can there be validity without reliability? Educational Researcher, 23(2), 5–12.

Moss, P. A., Girard, B. J., & Haniford, L. C. (2006). Validity in Educational Assessment

Myford. C. M. (1999). Assessment for accountability vs. assessment to improve teaching and learning: Are they two different animals? Paper presented at Australasian Curriculum, Assessment and Certification Authorities Conference. Perth.

Nunnally, J. C., & Bernstein, I. H. (2001). Psychometric Theory. New York: McGraw-Hill.

O'Brien, J. E. (1998). Incorporated in Matters, G. N., Pitman, J. A., & O'Brien, J. E. (1998). Validity and reliability in educational assessment and testing: A matter of judgment. Queensland Journal of Educational Research, 14(2) 57–88.

Parkes, C., & Maughan, S. (Eds). (2009). Methods for ensuring reliability of teacher assessments: Proceedings of the policy and research seminar on Tuesday 2 June 2009 at The Royal Institute of British Architects, Slough: NFER. Retrieved July 2012 from http://www.nfer.ac.uk/publications/99902/

Pitman, J. A., O'Brien, J. E., & McCollow, J. E. (1999). High-quality assessment: We are what we believe and do. Paper presented at the 25th annual conference of the International Association for Educational Assessment. Bled, Slovenia.

Pollitt, A. (2004). Let's stop marking exams. Paper presented at the 30th IAEA conference, Philadelphia, June 2004.

Queensland Studies Authority (2011). Moderation: enhancing teacher assessment culture. Retrieved 23 July 2012 from acaca.qsa. qld.edu.au/docs/ACACA_2011_Molloy_Miller.pdf

Sadler, D. R. (1986). Subjectivity, objectivity and teachers' qualitative judgments. (Discussion paper 5). Brisbane: Assessment Unit, Board of Secondary School Studies.

Sadler, D. R. (1987) Specifying and promulgating achievement standards. Oxford Review of Education, 13(2), 191−209.

Sadler, D. R. (2003). Re-visiting specifying and promulgating achievement standards. Paper presented to the Assessment and Reporting Framework Implementation Committee, Education Queensland, Brisbane.

Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading. Assessment & Evaluation in Higher Education. 34(2), 159–179.

Salvia, J., & Ysseldyke, J. E. (1998). Assessment in special and remedial education. Houghton Mifflin. Co.

Shavelson, R. J., Black, P. J, Wiliam, D., & J. Coffey. (2004). On linking formative and summative functions in the design of large-scale assessment systems. Educational Evaluation and Policy Analysis

Sireci, S. G. (2009). Packing and unpacking sources of validity evidence. In R. Lissitz (Ed.). The Concept of Validity: Revisions, New Directions and Applications, 19–37. Charlotte, NC: Information Age Press.

Smaill, E. (2012). Moderating New Zealand's national standards: teacher learning and assessment outcomes. Assessment in Education: Principles, Policy & Practice, DOI: 10.1080/0969594X.2012.696241

Stobart, G. (2004). The formative use of summative assessment: Possibilities and limits. Paper presented at the 30th IAEA Conference, Philadelphia.

Strachan, J. (2002). Assessment in change: Some reflections on the local and international background to the National Certificate of Educational Achievement (NCEA). New Zealand Annual Review of Education, 11, 245–262.

Wyatt-Smith, C. M., & Matters, G. N. (2009). Proposal for a new model of senior assessment: Realising potentials. Report commissioned by the Queensland Studies Authority. Brisbane: Griffith University and ACER.

Wyatt-Smith, C., Klenowski, V., & Gunn, S. (2010). The centrality of teachers' judgment practice in assessment: A study of standards in moderation. Assessment in Education: Principles, Policy & Practice, 17(1), 59–75.

Introduction to School-based Assessment. International Practice on School-based Assessment. Accessed July 2011 from http://www.hkeaa.edu.hk/DocLibrary/SBA/HKDSE/Eng_DVD/sba_practice.html

Appendices

Appendix 1: Literature review method

Procedure

The procedure had two components:

1. Scan the literature on teacher assessments and moderation, particularly at the Key Stages 2 and 3 level:
   - Clearly define the areas of interest and the search terms that will be used.
   - Identify target databases and journals and locate articles that meet the research terms and filter criteria.
2. Review and synthesise the literature to answer the following questions:
   a. What works (and does not work) in teacher assessment?

   b. What works (and does not work) in teacher assessment moderation?

   c. What works (and does not work) in becoming able to "level" work?

   d. What models of assessment exist internationally for similar teacher assessment and moderation schemes in this age range?

   e. Is there a relationship between type of assessment regime internationally and performance on international tests?

Key words

Moderation, teacher assessment, school-based assessment, teacher judgments, comparability in classroom assessments, validity of teacher assessments, reliability of teacher assessments, consistency (of teacher judgments)

Criteria for selection of texts

The prime criterion was relevance to the subject matter. Novelty also came into play once the references led to an exciting idea or procedure. At the risk of finding that novelty was unrelated to credibility, credibility was another of the criteria for selection. Authority and accessibility were two other criteria that were applied to the selection of texts. The criterion of accessibility was never a problem but ambiguity of terms was. The internet search gives an indication of the ease of access to key words for this review.

The results of a preliminary Google search of the World Wide Web yielded 40 million references for teacher assessment. It soon became apparent that much of these related to assessment of teachers (as in teaching standards, teacher accreditation etc.) as opposed to assessment by teachers of student achievement. School-based assessment (53 million) was too broad as a key term and classroom assessment (19 million) was often portrayed as a “soft touch”. Other reference frequencies of note are: teacher judgment (29 million); consistency of teacher judgments (3.1 million); moderation of teacher judgments (29 million); social moderation (14 million); moderation of assessment (2.0 million); and cluster moderation in assessment (1.5 million), for which the first ten referred to work in Australia. This goes some way to explaining why the reference list for this review is peppered with the names of Australians.

Sources of information

Several strategies were employed. They included:

- A search of data bases through the Cunningham Library at ACER – Australian Education Index, ERIC, Education Research Complete, and PsycInfo;
- An attempt to find reports, scholarly journal articles, and examples of teacher assessment not only in Australia and the UK, but beyond;
- Extensive use of the internet to track down references cited in bibliographies;
- Extensive use of the internet to acquire “e-texts”;
- Face-to-face and e-mail conversations with practitioners in the field;
- Trawling through papers from conferences especially the International Association for Educational Assessment (IAEA); and,
- Author’s personal library, which contains texts not available electronically.

All texts that could be acquired were at least skimmed and checked for conclusions and major points included in the abstracts and bibliographies.

The currency of 2007 and beyond as set down in the proposal was not delivered on because parts of the story on consistency of teacher judgments and validating them starts long before 2007 and remains relevant today. The general searches were undertaken in three stages: after 2007, 2002–2007, before 2007. The searches of databases related to work published between 2002 and 2012 and between 1992 and 2001.

Although the literature review is part of an investigation into assessment in the learning spans that align with key stages 2 and 3 in Wales, most examples of successful practice of school-based assessment (i.e., socially moderated teacher judgments) are in the senior span. Therefore, the review was not constrained to the schooling span of key stages 2 and 3, which is not to say that successful models for the senior years are suitable for the primary and middle-school years.

The currency of the research is attested to by the inclusion of two papers just published (July 2012) written by Brookhart in the USA and Smaill in New Zealand. This review would not be complete without the inclusion of work from outside the UK and Australia, especially from two countries (USA and New Zealand) that could be considered polar opposites in terms of assessment policy and practice.

Classification of selected texts

- Major policy documents from sources in the UK, Australia, and beyond
- International sources whose research has wide application
- Projects that have published significant findings or recommendations for the future
- The perceived popularity of the text in citations and referencing
- A sense that the text is seminal
- Unusual or unconventional takes on the subject
- Indications of significant implications for policy and practice
- Theory building (as opposed to theory testing)
- Principles-based.

Compositions of reference list

Approximately 170 sources including:

- Journal articles – 71
- Books and chapters in books – 19
- Conference papers – 21
- Review and reports to government – 13
- Research reports – 24
- Handbooks – 15
- Miscellaneous – 6

Background

An investigation is being undertaken into the effectiveness of teacher assessment of the National Curriculum for Wales at Key Stages 2 and 3 levels[14]. The investigation is examining evidence from a variety of sources to construct an understanding of how well the current system of teacher assessments is working and how it might be improved. In 2010, five years after the Welsh Assembly Government abolished statutory end-of-key-stage testing for 11–14-year olds and adopted a system of teacher assessment, an evaluation report (Estyn, 2010) concluded that teachers had not yet reached a common understanding of the standards that they were applying to student work in the assessment process.

Product

The literature review, which was undertaken in June and July 2012, outlines major work in the field of teacher assessment with particular regard to consistency of teacher judgments in the assessment process and validation of teacher judgments in the case where internal rather than external assessment is operating. Associated with these two major aspects of teacher assessment are notions of summative and formative assessment (and their relationship), standards-referenced assessment, moderation, validity, and reliability.

Because of the high degree of interrelatedness between the factors at work in teacher assessment and moderation, it was not easy to treat these particular facets in isolation. To a certain extent each one is a part of or shapes the other. Thus, the literature review is not structured around those two facets of assessment but around related concepts and themes and, as a consequence of this structure, there is repetition and overlap.

Despite the considerable volume of writings on the topics of teacher assessment and moderation, the literature is patchy. The patchwork nature of the literature is, in part, a function of the terminology that is used. Without a clear definition of terms and an understanding of concepts and subsidiary issues a proper consideration of the issues is difficult.

The intricacies of popular terminology are illustrated by the use of the verb “to moderate”, the use of the word “comparability”, and the often silent word after “comparability of”. The object of the verb to moderate is “teacher judgments” and comparability refers to “reported results/grades”. Also there is a distinction comparability of assessment and comparability of reported results; the former was not discussed in the review. As the literature review showed these are not unimportant issues. Also in the literature review were overtones of an uneasy coexistence of so-called “objective” assessment and a “subjective” model based on

[14] Key Stage 2 in Wales is equivalent to Years 3–6 (7 to 11 years of age). Key Stage 3 covers Years 7–9 (11 to 14 years of age). Key Stage 3 marks the end of compulsory schooling and is followed by the GCSE, A-levels etc.

the experience and skills of teachers as judges of standards.

A deliberate attempt was made to ensure that this literature review did not take on or appear to take on the genre of a comparative study – there is only passing mention of assessment in Wales, and the terminology in the literature review is not translated into the assessment terminology of Wales (or of any other country for that matter). Answers to the review questions were pursued at a general or global level (and even that was complicated by the terminology used across and within countries).

Conclusions of the literature review

What follows is not a summary of this literature review and therefore the reader cannot be fully apprised of the analytic base for these conclusions without reference to the main body of the review.

1. It would appear, from a preliminary reading of the literature, that there are many ways of executing teacher assessment and moderation, and that each has its own special way of operating. A more detailed analysis of the literature, however, reveals enormous similarities (at least, in statements of purpose, policy and practice), similarities that quite possibly arose from the export and import of moderated teacher assessment models from jurisdiction to jurisdiction over the past decade (at least) since external examinations and tests have been relegated a lesser role than in the past.

2. Descriptions of assessment systems that are based on moderated teacher judgments invariably include reference to evidence, standards, and validation, which correspond to Elements 3–5 in the representations of effective assessment systems, strategies for improving the consistency of teacher judgments, and processes of assessment that were described in the prelude to this review. Also, the literature is full of exhortations about consistency of teacher judgments, the use of exemplars, and the need for professional development of teachers. All of this indicates that there is a high level of agreement among educationists, policy makers and researchers about what constitutes good teacher assessment and moderation.

3. Unsurprisingly, there is variation in the moderation model applied: The moderation model may be different at different stages of schooling (compulsory versus non-compulsory years). The moderation model may vary according to the mode of assessment (written expression versus constructed response). The moderation model may be different for high-stakes and low-stakes assessment. The moderation model may differ depending on whether there is an internal or external locus of control. The international experience indicates that there is no single factor on its own that will be sufficient on its own to ensure that teacher assessment is good or to ensure that moderation is good. For assessment in the primary and middle years of schooling, however, cluster moderation meetings appear to be the most appropriate for the level of control required and the stakes of the assessment decisions.

4. Moderation meetings are seen to be powerful professional development and a networking opportunity for those involved although the primary purpose of moderation is to ensure comparability of results.

5. The current state of research reflects the small number of instances where moderated teacher assessment regimes (or systems) have been systematically implemented and supported over time. In most cases the system is new (the result of a reform process, for example) or evolving. The same language, more or less, is used across the assessment world. It is universally noted that a key component of successful teacher assessment is teacher expertise especially skills in making good judgments (i.e., applying relevant achievement standards or accurately levelling student work). Thus consistency of teacher judgments is a constant theme in research, theory and practice.

6. Models of assessment that exist internationally for socially moderated teacher assessment in Years 3–9 belong to one of two categories – emergent upwards or derivative downwards. The former category includes systems moving from a position where teacher assessment had been considered suitable for formative assessment only to a position where teacher judgments had come to be privileged for summative assessment, typically in the primary and middle school years. What these emergent upwards systems have in common is that they try to make sense of practices undertaken elsewhere in order to distil a distinctive model for their own circumstances. The “derivative downwards” category includes systems incorporating a model that has been established in the senior years of schooling to the junior years of schooling. What these derivative downwards systems have in common is that they need to develop an understanding of the principles upon which the senior model is based and then decide how to adapt that model to suit lower-stakes purposes.

7. Teacher assessment and moderation works if evidence is collected over time rather than at the end of a term or year or course of study, if learning accumulates and profiles are selectively updated, if all important aspects of the course are assessed including practical work. Teacher assessment does not work if the afore-mentioned conditions are not met; the same assessment could be accomplished through external measures.

8. Underlying concepts and practices are under-theorised and have not been the subject of substantial research to date (Wyatt-Smith, Klenowski, & Gunn, 2010; Crisp, 2010). In fact most examples of teacher assessment and moderation are based on principles rather than theory. Until very recently theory building rather than theory testing was the order of the day.

9. A recurring theme in the literature on teacher assessment is the level of expertise

required of teacher–assessors. The skill sets that can be identified as necessary are expertise in their disciplines, total immersion in their level statements, the ability to evaluate, and the ability to negotiate when seeking consensus at moderation meetings.

10. There are a few significant gaps in our understanding of teacher assessment and moderation:

    a. It is not clear how teachers make judgments and, in particular, about how competent teacher–assessors are in making on-balance or best-fit judgments.

    b. It has not been demonstrated that the most popular way of selecting exemplars (e.g., samples of student work at various levels) is the most effective way to build teacher capacity to "level" work. The question to be answered here is whether exposing teachers to samples of student work at various levels (work that by definition is not perfect) encourages misunderstandings about on-balance judgments.

    c. Missing or not fully developed in the discourse about moderation (theory and practice) is information about other possible reasons for varying the moderation model: An unanswered question is whether the same moderation model is suitable for all subjects being assessed. Most of the research emanates from assessment of written expression, literacy or language. Is there a general message here for mathematics and science?

    d. The relative merits of quantitative and qualitative research methods are acknowledged but it must be said that there is very little quantitative research to provide empirical support for some of the more recent forays into the field of teacher assessment and moderation.

11. Teacher assessment and moderation do not work if there is insufficient funding for proper standards-setting exercises and professional development of teachers. On the other hand, costs must be proportionate to benefits; for example, the scope of moderation meetings must be manageable.

12. Teacher assessment and moderation do not work if the public does not have faith in the accuracy of the outcomes. Even when the models are educationally sound, stakeholder perception might hold them to be otherwise. Success in convincing the public (and educationists for that matter) that the assessments are accurate requires sophisticated communication strategies.

13. This review did not produce a definitive statement on the relationship between type of assessment regime and performance on international tests because it is not a simple matter to categorise assessment regimes according to their locus of control for assessments because there are many variations within each (internal and external) as well as between them and also because the nature of the assessment

regime in a particular country can change rapidly. Furthermore, causal links are difficult to establish.

14. The intersection of assessment regimes with internal and external loci of control is not generally well understood by practitioners and not fully explored in the literature and research. Additionally, forms of assessment that developed at the end of the last century such as standards-referenced assessment and levelling are not well understood. This state of affairs is made worse by the ever widening range of vocabulary that is used in this field, possibly because of each system's desire to be distinctive.

Appendix 2: Questionnaires

A2.1 Questions in the Cluster Co-ordinator Questionnaire

Q1: To your knowledge, have any cluster staff attended external training regarding cluster standardisation / moderation during the last year?

Q1a: If Yes, briefly describe the attendance at training

Brief description of attendance at training

Q2: Does cluster moderation take place in all the core subjects (and Welsh 2nd language in English medium schools)? *Yes/No*

Q3: Is cluster moderation an annual event? *Yes/No*

Q4: Briefly describe the time allocation and organisation given to cluster moderation during 2011-2012

Brief description of time allocation and organisation

Q5: What funding stream is used to support cluster moderation in 2011-12?

Brief description of time allocation and organisation

Q6: Who is primarily responsible for co-ordinating cluster moderation meetings?

Position / role of co-ordinator(s):

Q7: Which core subject teachers attend the cluster meetings?

Position / role of teacher(s):

Q8: To your knowledge, do teachers who attend the meetings have opportunity to feedback cluster judgments to other KS3 subject teachers? *Yes/No*

Q9i: Does the cluster have standardisation portfolios in place for Oracy in English?

Q9ii: Does the cluster have standardisation portfolios in place for Reading in English?

Q9iv: Does the cluster have standardisation portfolios in place for Oracy in Welsh?

Q9v: Does the cluster have standardisation portfolios in place for Reading in Welsh?

Q9vi: Does the cluster have standardisation portfolios in place for Writing in Welsh?

Q9vii: Does the cluster have standardisation portfolios in place for Mathematics?

Q9viii: Does the cluster have standardisation portfolios in place for Science?

Q10: Do standardisation portfolios have appropriate commentaries linking the evidence (i.e. samples of pupil's work) to specific level descriptions? *Yes/Developing/No*

Q11: Does the evidence and judgments within cluster standardisation portfolios reflect the shared understanding of NC level descriptions of all KS2 and 3 teachers attending the cluster meetings? *Yes/Developing/No*

Q12: Do cluster representatives take and consider a range of learner profiles to determine the 'best fit' NC levels? *Yes/Developing/No*

Q13i: Does the cluster have specific learner profiles in place for NC Levels 4 and 5 for Oracy in English?

Q13ii: Does the cluster have specific learner profiles in place for NC Levels 4 and 5 for Reading in English?

Q13iii: Does the cluster have specific learner profiles in place for NC Levels 4 and 5 for Writing in English?

Q13iv: Does the cluster have specific learner profiles in place for NC Levels 4 and 5 for Oracy in Welsh?

Q13v: Does the cluster have specific learner profiles in place for NC Levels 4 and 5 for Reading in Welsh?

Q13vi: Does the cluster have specific learner profiles in place for NC Levels 4 and 5 for Writing in Welsh?

Q13vii: Does the cluster have specific learner profiles in place for NC Levels 4 and 5 for Mathematics?

Q13viii: Does the cluster have specific learner profiles in place for NC Levels 4 and 5 for Science?

Q14: For English, Welsh (and Welsh 2nd language in English medium schools), do the overall subject levels take account of aggregation of separate AT outcomes in oracy, reading and writing? *Yes/Developing/No*

Q15: Do cluster learner profiles have appropriate commentaries linking the evidence (i.e. named pupil's work) to specific NC levels? *Yes/Developing/No*

Q16: To your knowledge, do teachers use the cluster learner profiles to help moderate their overall pupil assessments at the end of KS2 and KS3? *Yes/Developing/No*

Q17: Does the evidence and judgments within cluster learner profiles reflect the shared understanding of NC levels and standards of all KS2 and KS3 teachers? *Yes/Developing/No*

Q18: Overall, how much confidence do you have in the accuracy and reliability of your existing cluster moderation procedures? *No confidence/some confidence/complete confidence*

Q19: Do you consider the national curriculum level descriptions useful for assessment? *Yes/No*

Q20: Do you consider current teacher assessment procedures to be an effective use of resources?

Q21: Do you consider the current KS2/3 external moderation programme to be an effective use of resources? *Yes/No*

Q22: Overall, do you consider that cluster assessments accurately reflect the actual ability of learners? *Yes/No*

Q23: What do you consider to be the main barriers to ensuring reliable and consistent teacher assessment?

Enter comments here

Q24: What do you think might improve the quality of pupil assessments?

Enter comments here

Q25: Final Comments

Enter comments here

A2.2 Questions in the Local Authority Questionnaire

Q1: Does the LA have copies (electronic or paper based) of assessment policies from all of its schools? *Yes/Developing/No*

Q2: Does the LA monitor school assessment policies to ensure they comply with statutory requirements? *Yes/Developing/No*

Q3: Does the LA provide guidance for schools in developing their assessment policies? *Yes/Developing/No*

Q4: Does the LA provide guidance for schools in developing their assessment practice? *Yes/Developing/No*

If No, skip to Q5

Q4a: If Yes, Does the guidance help schools to clearly distinguish between formative and summative assessment?

Q4b: If Yes Does the guidance help schools to clearly define and distinguish between standardisation and moderation?

Q5: Is the LA informed of arrangements for internal standardisation within all of its schools? *Yes/Developing/No*

Q6: Is the LA informed of arrangements for internal moderation within all of its schools? *Yes/Developing/No*

Q7: Is the LA informed of arrangements for cluster standardisation and moderation within all of its KS2/3 clusters? *Yes/Developing/No*

Q8: Do LA officers/advisers attend the following meetings?

Q8i: School Internal moderation

Q8ii: School Internal standardisation

Q8iii: Cluster standardisation/moderation

Q9: Does the LA have copies of the Transition Plans for all of its KS2/3 school clusters? *Yes/Developing/No*

Q10: Did the LA ensure that Transition Plans were renewed / updated at September 2010? *Yes/Developing/No*

Q11: Does the LA ensure that cluster moderation arrangements are the same as described in the current cluster Transition Plan i.e. as operable from September 2010? *es/Developing/No*

Q12: Are all LA officers / advisers familiar with the contents of Making the most of learning (DfES, 2008)? *Yes/Developing/No*

Q13: Are all LA officers / advisers familiar with the contents of Ensuring consistency in teacher assessment: Guidance for Key Stages 2 and 3 (DfES, 2008)? *Yes/Developing/No*

Q14: Are all LA officers / advisers familiar with the contents of Making the most of assessment (DfES, 2010)? *Yes/Developing/No*

Q15: Are all LA officers / advisers familiar with the contents of the annual DfES guidance Statutory assessment arrangements for the school year 2011-12? *Yes/Developing/No*

Q16: Do all LA officers / advisers clearly understand the difference between standardisation and moderation? *Yes/Developing/No*

Q17: Do all LA officers / advisers clearly understand the difference between standardisation portfolios and learner profiles? *Yes/Developing/No*

Q18: How does the LA promote / recommend that teachers record pupil attainment?

Select either NC Sub-levels or As 'best fit' in relation to the level descriptions

Q19: Does the LA use standardised tests to assess pupils in KS2 & 3? *Yes/Developing/No*

If Yes, Please indicate the range of tests currently used

If No please skip to Q.20

Q20: On a scale of 1 (least) to 5 (most), indicate the weighting you consider schools ought to give to standardised test outcomes in contributing to 'best-fit' NC levels.

Select either 1, 2, 3, 4 or 5

Q21: Overall, how much confidence do you have in the accuracy and reliability of your existing pupil assessment records? *No confidence/Some confidence/Complete confidence*

Please place any additional commentary on accuracy and reliability of pupil assessment records here

Q22: Has the LA provided any training regarding standardisation during the last year? *Yes /No. If No please skip to Q23*

If Yes, Please indicate what training was provided, for whom it was provided and who delivered it. Please add additional lines as required. Training type, recipients, delivered by.

Q23: Does the LA organise cluster / whole-authority meetings for schools re. standardisation? *Yes /No. If No please skip to Q25*

Q24 Indicate the funding steam used to support standardisation meetings in 2011-12

Brief description of funding stream

Q25i: Does the LA ensure that schools have standardisation portfolios in place for Oracy in English?

Q25ii: Does the LA ensure that schools have standardisation portfolios in place for Reading in English?

Q25iii: Does the LA ensure that schools have standardisation portfolios in place for Writing in English?

Q25iv: Does the LA ensure that schools have standardisation portfolios in place for Oracy in Welsh?

Q25v: Does the LA ensure that schools have standardisation portfolios in place for Reading in Welsh?

Q25vi: Does the LA ensure that schools have standardisation portfolios in place for Writing in Welsh?

Q25vii: Does the LA ensure that schools have standardisation portfolios in place for Mathematics?

Q25viii: Does the LA ensure that schools have standardisation portfolios in place for Science?

Q25ix: Does the LA ensure that schools have standardisation portfolios in place for Non-core subjects?

Q26: Does the LA ensure that schools standardisation portfolios have an appropriate range of evidence covering NC levels e.g. 3-5 for KS2 and 4-7 for KS3? *Yes /No*

Q27: Does the LA ensure that schools standardisation portfolios have appropriate commentaries linking the evidence (i.e. samples of pupil's work) to specific level descriptions? *Yes /No*

Q28: Overall, how much confidence do you have in the accuracy and reliability of your existing internal standardisation procedures? *No confidence/Some confidence/Complete confidence*

Please place any additional commentary on accuracy and reliability of existing internal standardisation procedures

Q29: Has the LA provided any training regarding moderation during the last year? *Yes /No. If No skip to Q31*

Q30: If Yes, please indicate what training was provided, for whom it was provided and who delivered it. Please add additional lines as required.

Training type, recipients, delivered by

Q31: Does the LA organise cluster / whole-authority meetings for schools re moderation? *Yes /No. If No skip to Q33*

Q32: Indicate the funding steam used to support moderation meetings in 2011-12.

Brief description of funding stream

Q33i: Does the LA ensure that schools have an appropriate range of learner profiles in place for Oracy in English?

Q33ii: Does the LA ensure that schools have an appropriate range of learner profiles in place for Reading in English?

Q33iii: Does the LA ensure that schools have an appropriate range of learner profiles in place for Writing in English?

Q33iv: Does the LA ensure that schools have an appropriate range of learner profiles in place for Oracy in Welsh?

Q33v: Does the LA ensure that schools have an appropriate range of learner profiles in place for Reading in Welsh?

Q33vi: Does the LA ensure that schools have an appropriate range of learner profiles in place for Writing in Welsh?

Q33vii: Does the LA ensure that schools have an appropriate range of learner profiles in place for Mathematics?

Q33viii: Does the LA ensure that schools have an appropriate range of learner profiles in place for Science?

Q33ix: Does the LA ensure that schools have an appropriate range of learner profiles in place for Non-core subjects?

Q34: Does the LA ensure that schools' learner profiles have appropriate commentaries linking the evidence (i.e. named pupil's work) to specific NC levels? *Yes /No*

Q35: Overall, how much confidence do you have in the accuracy and reliability of your schools' existing internal moderation procedures?

No confidence/Some confidence/Complete confidence. Please place any additional commentary on accuracy and reliability of existing internal moderation procedures

Q36: Has the LA provided any training regarding KS2/3 moderation during the last year?

Yes /No. If No skip to Q37

Q36a: If Yes, Please indicate what training was provided, for whom it was provided and who delivered it. Please add additional lines as required. Training type, recipients, delivered by.

Q37: Does the LA ensure that cluster moderation takes place for all core subjects (and Welsh 2nd language in English medium schools)? *Yes /No. If No skip to Q39*

Q38: Does the LA organise cluster / whole-authority meetings for schools re cluster moderation? *Yes /No. If No skip to Q39*

Q38a: If Yes, Who is primarily responsible for organising the meetings?

Person(s) responsible for arranging meetings

Q39: Briefly describe the time allocation and organisation given to cluster moderation during 2011-2012

Q40: Indicate the funding stream used to support cluster moderation in 2011-12.

Q41: To your knowledge, are all cluster schools appropriately represented at cluster meetings? *Yes /No*

Q42i: To your knowledge, does each cluster have cluster standardisation portfolios in place for Oracy in English?

Q42ii: To your knowledge, does each cluster have cluster standardisation portfolios in place for Reading in English?

Q42iii: To your knowledge, does each cluster have cluster standardisation portfolios in place for Writing in English?

Q42iv: To your knowledge, does each cluster have cluster standardisation portfolios in place for Oracy in Welsh?

Q42v: To your knowledge, does each cluster have cluster standardisation portfolios in place for Reading in Welsh?

Q42vi: To your knowledge, does each cluster have cluster standardisation portfolios in place for Writing in Welsh?

Q42vii: To your knowledge, does each cluster have cluster standardisation portfolios in place for Mathematics?

Q42viii: To your knowledge, does each cluster have cluster standardisation portfolios in place for Science?

Q43: Does the LA ensure that cluster standardisation portfolios have appropriate commentaries linking the evidence (i.e. samples of pupil's work) to specific level descriptions? *Yes /No*

Q44i: Does the LA ensure that each cluster has specific learner profiles in place for NC Levels 4 and 5 for Oracy in English?

Q44ii: Does the LA ensure that each cluster has specific learner profiles in place for NC Levels 4 and 5 for Reading in English?

Q44iii: Does the LA ensure that each cluster has specific learner profiles in place for NC Levels 4 and 5 for Writing in English?

Q44iv: Does the LA ensure that each cluster has specific learner profiles in place for NC Levels 4 and 5 for Oracy in Welsh?

Q44v: Does the LA ensure that each cluster has specific learner profiles in place for NC Levels 4 and 5 for Reading in Welsh?

Q44vi: Does the LA ensure that each cluster has specific learner profiles in place for NC Levels 4 and 5 for Writing in Welsh?

Q44vii: Does the LA ensure that each cluster has specific learner profiles in place for NC Levels 4 and 5 for Mathematics?

Q44viii: Does the LA ensure that each cluster has specific learner profiles in place for NC Levels 4 and 5 for Science?

Q45: Does the LA ensure that cluster learner profiles have appropriate commentaries linking the evidence (i.e. named pupil's work) to specific NC levels? *Yes /No*

Q46: Overall, how much confidence do you have in the accuracy and reliability of existing cluster moderation procedures? *No confidence/Some confidence/Complete confidence*

Q47: Do you consider the national curriculum level descriptions useful for assessment? *Yes?No*

Q48: Do you consider current teacher assessment procedures to be an effective use of resources? *Yes /No*

Q49: Do you consider the current KS2/3 external moderation programme to be an effective use of resources?*Yes /No.*

Q50: Overall, do you consider that your current Y6 & Y9 assessments accurately reflect the actual ability of learners? *Yes /No*

Q51: What do you consider to be the main barriers to ensuring reliable and consistent teacher assessment?

Enter comments here

Q52: What do you think might improve the quality of pupil assessments?

Enter comments here

Q53: Final comments

Enter comments here

A2.3 Questions in the Primary Schools (KS2) Questionnaire

Q1: Does the policy clearly distinguish between formative and summative assessment? *Yes/Developing/No*

Q2: Does the policy clearly define standardisation and moderation? *Yes/Developing/No*

Q3: Does the policy clearly distinguish between standardisation and moderation? *Yes/Developing/No*

Q4: Does the policy clearly describe arrangements for internal standardisation? *Yes/Developing/No*

Q5: Does the policy clearly describe arrangements for internal moderation? *Yes/Developing/No*

Q6: Does the policy clearly describe arrangements for cluster standardisation and moderation? *Yes/Developing/No*

Q7: Were cluster Transition Plans renewed / updated at September 2010? *es/Developing/No*

Q8: Are cluster moderation arrangements the same as described in the current cluster Transition Plan i.e. as operable from September 2010? *Yes/Developing/No*

Please place any additional comments you have on school assessment policy here

Q9: Are all teachers in your school familiar with the contents of Making the most of learning (DfES, 2008)? *Yes/Developing/No*

Q10: Are all teachers familiar with the contents of Ensuring consistency in teacher assessment: Guidance for Key Stages 2 and 3 (DfES, 2008)? *Yes/Developing/No*

Q11: Are all teachers familiar with the contents of Making the most of assessment (DfES, 2010)? *Yes/Developing/No*

Q12: Are all teachers familiar with the contents of the annual DfES guidance Statutory assessment arrangements for the school year 2011-12? *Yes/Developing/No*

Q13: Do all teachers clearly understand the difference between standardisation and moderation? *Yes/Developing/No*

Q14: Do all teachers clearly understand the difference between standardisation portfolios and learner profiles? *Yes/Developing/No*

Please place any additional comments you have on teacher understandings of assessment procedures here

Q15: Does the school have a pupil-level tracking system in place? *Yes/No If No please skip to Q16.*

Q15a: If Yes, what kind of tracking system is it?

Q15bi: Does the tracking system record pupil attainment in place for Oracy in English?

Q15bii: Does the tracking system record pupil attainment in place for Reading in English?

Q15biii: Does the tracking system record pupil attainment in place for Writing in English?

Q15biv: Does the tracking system record pupil attainment in place for Oracy in Welsh?

Q15bv: Does the tracking system record pupil attainments in place for Reading in Welsh?

Q15bvi: Does the tracking system record pupil attainment in place for Writing in Welsh?

Q15bvii: Does the tracking system record pupil attainment in place for Mathematics?

Q15bviii: Does the tracking system record pupil attainment in place for Science?

Q15c: How do teachers record pupil attainment? NC sub-levels.

Q15ci: How do teachers record pupil attainment? As 'best fit' in relation to the level descriptions

Q15cii: How do teachers record pupil attainment? Other (please describe)

Q16: Do you use standardised tests to assess pupils in KS2? *Yes/No If No please skip to Q17*

Q16a: If Yes, Please indicate the range of tests currently used.

Standardised tests currently used

Q17: On a scale of 1 (least) to 5 (most), indicate the weighting given to standardised test outcomes in contributing to pupils' 'best-fit' NC levels? *Select either 1, 2, 3, 4 or 5*

Q18: Overall, how much confidence do you have in the accuracy and reliability of your existing pupil assessment records?

No confidence/Some confidence/Complete confidence. Please place any additional commentary on accuracy and reliability of pupil assessment records here

Q19: Have any teaching staff attended external training regarding standardisation during the last year? *Yes/No If No please skip to Q20*

Q19a: If Yes, what are their positions/responsibilities?

Brief description of attendance at training

Q20i: Does internal standardisation take place in the school for Oracy in English?

Q20ii: Does internal standardisation take place in the school for Reading in English?

Q20iii: Does internal standardisation take place in the school for Writing in English?

Q20iv: Does internal standardisation take place in the school for Oracy in Welsh?

Q20v: Does internal standardisation take place in the school for Reading in Welsh?

Q20vi: Does internal standardisation take place in the school for Writing in Welsh?

Q20vii: Does internal standardisation take place in the school for Mathematics?

Q20viii: Does internal standardisation take place in the school for Science?

Q21: Is internal standardisation an ongoing or annual event?

Ongoing/Annual

Q22: Briefly describe the time allocation and organisation given to internal standardisation during 2011-2012

Brief description of time allocation and organisation

Q23: What funding stream (if any) is used to support internal standardisation in 2011-12?

Brief description of funding stream

Q24: Are all KS2 teachers involved in standardisation procedures? *Yes/No*

Q25i: Does the school have standardisation portfolios in place for Oracy in English?

Q25ii: Does the school have standardisation portfolios in place for Reading in English?

Q25iii: Does the school have standardisation portfolios in place for Writing in English?

Q25iv: Does the school have standardisation portfolios in place for Oracy in Welsh?

Q25v: Does the school have standardisation portfolios in place for Reading in Welsh?

Q25vi: Does the school have standardisation portfolios in place for Writing in Welsh?

Q25vii: Does the school have standardisation portfolios in place for Mathematics?

Q25viii: Does the school have standardisation portfolios in place for Science?

Q26: Do all standardisation portfolios have an appropriate range of evidence covering NC levels 1 to 5? *Yes/Developing/No*

Q27: In your opinion, do standardisation portfolios have appropriate commentaries linking the evidence (i.e. samples of pupil's work) to specific level descriptions? *Yes/Developing/No*

Q28: Does the evidence and judgments within standardisation portfolios reflect the shared understanding of NC level descriptions of all KS2 teachers? *Yes/Developing/No*

Q29: Overall, how much confidence do you have in the accuracy and reliability of your existing internal standardisation procedures? *No confidence/Some confidence/Complete confidence*

Please place any additional commentary on accuracy and reliability of existing internal standardisation procedures

Q30: Have any staff attended external training regarding moderation during the last year? *Yes/No*

Q30a: If Yes, what are their positions/responsibilities?

Brief description of attendance at training

Q31i: Does internal moderation take place in the school for Oracy in English?

Q31ii: Does internal moderation take place in the school for Reading in English?

Q31iii: Does internal moderation take place in the school for Writing in English?

Q31iv: Does internal moderation take place in the school for Oracy in Welsh?

Q31v: Does internal moderation take place in the school for Reading in Welsh?

Q31vi: Does internal moderation take place in the school for Writing in Welsh?

Q31vii: Does internal moderation take place in the school for Mathematics?

Q31viii: Does internal moderation take place in the school for Science?

Q32: Is internal moderation an ongoing or annual event?

Ongoing/Annual

Q33: Briefly describe the time allocation and organisation given to internal moderation during 2011012

Brief description of time allocation and organisation

Q34: What funding stream is used to support internal moderation in 2011-12?

Brief description of funding stream

Q35: Are all KS2 teachers involved in ongoing or annual moderation procedures? *Yes/No*

Q36: Do teachers consider a range of evidence to determine the 'best fit' NC level of individual pupils at the end of KS2? *Yes/Developing/No*

Q37i: Does the school have specific learner profiles in place for NC Levels 3, 4 and 5 for Oracy in English?

Q37ii: Does the school have specific learner profiles in place for NC Levels 3, 4 and 5 for Reading in English?

Q37iii: Does the school have specific learner profiles in place for NC Levels 3, 4 and 5 for Writing in English?

Q37iv: Does the school have specific learner profiles in place for NC Levels 3, 4 and 5 for Oracy in Welsh?

Q37v: Does the school have specific learner profiles in place for NC Levels 3, 4 and 5 for Reading in Welsh?

Q37vi: Does the school have specific learner profiles in place for NC Levels 3, 4 and 5 for Writing in Welsh?

Q37vii: Does the school have specific learner profiles in place for NC Levels 3, 4 and 5 for Mathematics?

Q37viii: Does the school have specific learner profiles in place for NC Levels 3, 4 and 5 for Science?

Q38: For English, Welsh (and Welsh 2nd language in English medium schools), do the overall subject levels take account of aggregation of separate AT outcomes in oracy, reading and writing? *Yes/Developing/No*

Q39: Do learner profiles have appropriate commentaries linking the evidence (i.e. named pupil's work) to specific NC levels? *Yes/Developing/No*

Q40: Do teachers use the learner profiles to help moderate their overall pupil assessments at the end of KS2? *Yes/Developing/No*

Q41: In your opinion does the evidence and judgments within learner profiles reflect the shared understanding of NC levels and standards of all KS2 teachers? *Yes/Developing/No*

Q42: Overall, how much confidence do you have in the accuracy and reliability of your existing internal moderation procedures? *No confidence/Some confidence/Complete confidence*

Please place any additional commentary on accuracy and reliability of existing internal moderation procedures.

Q43: Have any staff attended external training regarding cluster standardisation / moderation during the last year? *Yes/No*

Q43a: If Yes, what are their positions/responsibilities?

Brief description of attendance at training

Q44i: Does cluster moderation take place in the school for Oracy in English?

Q44ii: Does cluster moderation take place in the school for Reading in English?

Q44iii: Does cluster moderation take place in the school for Writing in English?

Q44iv: Does cluster moderation take place in the school for Oracy in Welsh?

Q44v: Does cluster moderation take place in the school for Reading in Welsh?

Q44vi: Does cluster moderation take place in the school for Writing in Welsh?

Q44vii: Does cluster moderation take place in the school for Mathematics?

Q44viii: Does cluster moderation take place in the school for Science?

Q45: Is cluster moderation an annual event? *Yes/No*

Q46: Briefly describe the time allocation and organisation given to cluster moderation during 2011-2012

Brief description of time allocation and organisation

Q47: What funding stream is used to support cluster moderation in 2011-12?

Brief description of funding stream

Q48: Who is primarily responsible for co-ordinating cluster moderation meetings?
Position / role of co-ordinator(s):

Q49: Which of your teachers attend the cluster meetings? *Position / role of teacher(s):*

Q50: Do teachers who attend the meetings have opportunity to feedback cluster judgments to other KS2 teachers? *Yes/No*

Q51i: Does the cluster have standardisation portfolios in place for Oracy in English?

Q51ii: Does the cluster have standardisation portfolios in place for Reading in English?

Q51iii: Does the cluster have standardisation portfolios in place for Writing in English?

Q51iv: Does the cluster have standardisation portfolios in place for Oracy in Welsh?

Q51v: Does the cluster have standardisation portfolios in place for Reading in Welsh?

Q51vi: Does the cluster have standardisation portfolios in place for Writing in Welsh?

Q51vii: Does the cluster have standardisation portfolios in place for Mathematics?

Q51viii: Does the cluster have standardisation portfolios in place for Science?

Q52: Do standardisation portfolios have appropriate commentaries linking the evidence (i.e. samples of pupil's work) to specific level descriptions? *Yes/Developing/No*

Q53: Does the evidence and judgments within cluster standardisation portfolios reflect the shared understanding of NC level descriptions of all KS2 and 3 teachers attending the cluster meetings? *Yes/Developing/No*

Q54: Do cluster representatives take and consider a range of learner profiles to determine the 'best fit' NC levels? *Yes/Developing/No*

Q55i: Does the cluster have specific learner profiles in place for NC Levels 4 and 5 for Oracy in English?

Q55ii: Does the cluster have specific learner profiles in place for NC Levels 4 and 5 for Reading in English?

Q55iii: Does the cluster have specific learner profiles in place for NC Levels 4 and 5 for Writing in English?

Q55iv: Does the cluster have specific learner profiles in place for NC Levels 4 and 5 for Oracy in Welsh?

Q55v: Does the cluster have specific learner profiles in place for NC Levels 4 and 5 for Reading in Welsh?

Q55vi: Does the cluster have specific learner profiles in place for NC Levels 4 and 5 for Writing in Welsh?

Q55vii: Does the cluster have specific learner profiles in place for NC Levels 4 and 5 for Mathematics?

Q55viii: Does the cluster have specific learner profiles in place for NC Levels 4 and 5 for Science?

Q56: For English, Welsh (and Welsh 2nd language in English medium schools), do the overall subject levels take account of aggregation of separate AT outcomes in oracy, reading and writing? *Yes/Developing/No*

Q57: Do cluster learner profiles have appropriate commentaries linking the evidence (i.e. named pupil's work) to specific NC levels? *Yes/Developing/No*

Q58: Do teachers use the cluster learner profiles to help moderate their overall pupil assessments at the end of KS2 and KS3 *Yes/Developing/No*

Q59: Does the evidence and judgments within cluster learner profiles reflect the shared understanding of NC levels and standards of all KS2 and KS3 teachers? *Yes/Developing/No*

Q60: Overall, how much confidence do you have in the accuracy and reliability of your existing cluster moderation procedures? *No confidence/Some confidence/Complete confidence*

Q61: Do cluster learner profiles have appropriate commentaries linking the evidence (i.e. named pupil's work) to specific NC levels? *Yes/No*

Q62: Do teachers use the cluster learner profiles to help moderate their overall pupil assessments at the end of KS2 and KS3? *Yes/Developing/No*

Q63: Do you consider the current KS2/3 external moderation programme to be an effective use of resources? *Yes/Developing/No*

Q64: Overall, do you consider that your current Y6 assessments accurately reflect the actual ability of learners? *Yes/Developing/No*

Q65: What do you consider to be the main barriers to ensuring reliable and consistent teacher assessment?

Enter comments here

Q66: What do you think might improve the quality of pupil assessments?

Enter comments here

Q67: Final comments

Enter comments here

A2.4 Questions in the Secondary Schools (KS3) Questionnaire

Q1: Does the policy clearly distinguish between formative and summative assessment? *Yes/Developing/No*

Q2: Does the policy clearly define standardisation and moderation? *Yes/Developing/No*

Q3: Does the policy clearly distinguish between standardisation and moderation? *Yes/Developing/No*

Q4: Does the policy clearly describe arrangements for internal standardisation? *Yes/Developing/No*

Q5: Does the policy clearly describe arrangements for internal moderation? *Yes/Developing/No*

Q6: Does the policy clearly describe arrangements for cluster standardisation and moderation? *Yes/Developing/No*

Q7: Were cluster Transition Plans renewed / updated at September 2010? *Yes/Developing/No*

Q8: Are cluster moderation arrangements the same as described in the current cluster Transition Plan i.e. as operable from September 2010? *Yes/Developing/No*

Please place any additional comments you have on school assessment policy here.

Q9: Are all teachers in your school familiar with the contents of Making the most of learning (DfES, 2008)? *Yes/Developing/No*

Q10: Are all teachers familiar with the contents of Ensuring consistency in teacher assessment: Guidance for Key Stages 2 and 3 (DfES, 2008)? *Yes/Developing/No*

Q11: Are all teachers familiar with the contents of Making the most of assessment (DfES, 2010)? *Yes/Developing/No*

Q12: Are all teachers familiar with the contents of the annual DfES guidance Statutory assessment arrangements for the school year 2011-12? *Yes/Developing/No*

Q13: Do all teachers clearly understand the difference between standardisation and moderation? *Yes/Developing/No*

Q14: Do all teachers clearly understand the difference between standardisation portfolios and learner profiles? *Yes/Developing/No*

Please place any additional comments you have on teacher understandings of assessment procedures here

Q15: Does the school have a pupil-level tracking system in place?

Yes/No If No please skip to Q16

Q15a: If Yes, what kind of tracking system is it?

School based, LA-Based, Commercial Package

Q15bi: Does the tracking system record pupil attainment in place for Oracy in English?

Q15bii: Does the tracking system record pupil attainment in place for Reading in English?

Q15biii: Does the tracking system record pupil attainment in place for Writing in English?

Q15biv: Does the tracking system record pupil attainment in place for Oracy in Welsh?

Q15bv: Does the tracking system record pupil attainments in place for Reading in Welsh?

Q15bvi: Does the tracking system record pupil attainment in place for Writing in Welsh?

Q15bvii: Does the tracking system record pupil attainment in place for Mathematics?

Q15bviii: Does the tracking system record pupil attainment in place for Science?

Q15bix: Does the tracking system record pupil attainment in place for non-core subjects?

Q15ci: How do teachers record pupil attainment? NC sub-levels.

Q15cii: How do teachers record pupil attainment? As 'best fit' in relation to the level descriptions

Q15ciii: How do teachers record pupil attainment? Other (please describe)

Q16: Do you use standardised tests to assess pupils in KS3?

Yes/No If No please skip to Q18

Q16a: If Yes, Please indicate the range of tests currently used.

Standardised tests currently used

Q17: On a scale of 1 (least) to 5 (most), indicate the weighting given to standardised test outcomes in contributing to pupils' 'best-fit' NC levels?

Select either 1, 2, 3, 4 or 5

Q18: Overall, how much confidence do you have in the accuracy and reliability of your existing pupil assessment records?

No confidence/Some confidence/Complete confidence.

Q19: Have any teaching staff attended external training regarding standardisation during the last year?

Yes/No If No please skip to Q19

Q19a: If Yes, what are their positions/responsibilities?

Brief description of attendance at training.

Q20i: Does internal standardisation take place in the school for Oracy in English?

Q20ii: Does internal standardisation take place in the school for Reading in English?

Q20iii: Does internal standardisation take place in the school for Writing in English?

Q20iv: Does internal standardisation take place in the school for Oracy in Welsh?

Q20v: Does internal standardisation take place in the school for Reading in Welsh?

Q20vi: Does internal standardisation take place in the school for Writing in Welsh?

Q20vii: Does internal standardisation take place in the school for Mathematics?

Q20viii: Does internal standardisation take place in the school for Science?

Q20a: Does internal standardisation take place in the school for non-core subjects? *Yes/No*

Q21: Is internal standardisation an ongoing or annual event? *Ongoing/Annual*

Q22: Briefly describe the time allocation and organisation given to internal standardisation during 2011-2012

Brief description of time allocation and organisation

Q23: What funding stream (if any) is used to support internal standardisation in 2011-12?

Brief description of funding stream

Q24: Are all KS3 teachers involved in standardisation procedures? *Yes/No*

Q25i: Does the school have standardisation portfolios in place for Oracy in English?

Q25ii: Does the school have standardisation portfolios in place for Reading in English?

Q25iii: Does the school have standardisation portfolios in place for Writing in English?

Q25iv: Does the school have standardisation portfolios in place for Oracy in Welsh?

Q25v: Does the school have standardisation portfolios in place for Reading in Welsh?

Q25vi: Does the school have standardisation portfolios in place for Writing in Welsh?

Q25vii: Does the school have standardisation portfolios in place for Mathematics?

Q25viii: Does the school have standardisation portfolios in place for Science?

Q25ix: Does the school have standardisation portfolios in place for non-core subjects?

Q26: Do all standardisation portfolios have an appropriate range of evidence covering NC levels 3 to 7? *Yes/Developing/No*

Q27: In your opinion, do standardisation portfolios have appropriate commentaries linking the evidence (i.e. samples of pupil's work) to specific level descriptions? *Yes/Developing/No*

Q28: Does the evidence and judgments within standardisation portfolios reflect the shared understanding of NC level descriptions of all KS3 subject teachers? *Yes/Developing/No*

Q29: Overall, how much confidence do you have in the accuracy and reliability of your existing internal standardisation procedures? *No confidence/Some confidence/Complete confidence*

Please place any additional commentary on accuracy and reliability of existing internal standardisation procedures

Q30: Have any staff attended external training regarding moderation during the last year? *Yes/No*

Q30a: If Yes, what are their positions/responsibilities?

Brief description of attendance at training

Q31i: Does internal moderation take place in the school for Oracy in English?

Q31ii: Does internal moderation take place in the school for Reading in English?

Q31iii: Does internal moderation take place in the school for Writing in English?

Q31iv: Does internal moderation take place in the school for Oracy in Welsh?

Q31v: Does internal moderation take place in the school for Reading in Welsh?

Q31vi: Does internal moderation take place in the school for Writing in Welsh?

Q31vii: Does internal moderation take place in the school for Mathematics?

Q31viii: Does internal moderation take place in the school for Science?

Q31b: Does internal moderation take place in the school for non-core subjects? *Yes/No*

Q32: Is internal moderation an ongoing or annual event?

Ongoing/Annual

Q33: Briefly describe the time allocation and organisation given to internal moderation during 2011-2012

Brief description of time allocation and organisation

Q34: What funding stream is used to support internal moderation in 2011-12?

Brief description of funding stream

Q35: Are all KS3 teachers involved in ongoing or annual moderation procedures *Yes/No*

Q36: Do teachers consider a range of evidence to determine the 'best fit' NC level of individual pupils at the end of KS3? *Yes/Developing/No*

Q37i: Does the school have specific learner profiles in place for NC Levels 4, 5, 6 and 7 for Oracy in English?

Q37ii: Does the school have specific learner profiles in place for NC Levels 4, 5, 6 and 7 for Reading in English?

Q37iii: Does the school have specific learner profiles in place for NC Levels 4, 5, 6 and 7 for Writing in English?

Q37iv: Does the school have specific learner profiles in place for NC Levels 4, 5, 6 and 7 for Oracy in Welsh?

Q37v: Does the school have specific learner profiles in place for NC Levels 4, 5, 6 and 7 for Reading in Welsh?

Q37vi: Does the school have specific learner profiles in place for NC Levels 4, 5, 6 and 7 for Writing in Welsh?

Q37vii: Does the school have specific learner profiles in place for NC Levels 4, 5, 6 and 7 for Mathematics?

Q37viii: Does the school have specific learner profiles in place for NC Levels 4, 5, 6 and 7 for Science?

Q37ix: Does the school have specific learner profiles in place for NC Levels 4, 5, 6 and 7 for non-core subjects?

Q38: For English, Welsh (and Welsh 2nd language in English medium schools), do the overall subject levels take account of aggregation of separate AT outcomes in oracy, reading and writing? *Yes/Developing/No*

Q39: Do learner profiles have appropriate commentaries linking the evidence (i.e. named pupil's work) to specific NC levels? *Yes/Developing/No*

Q40: Do teachers use the learner profiles to help moderate their overall pupil assessments at the end of KS3? *Yes/Developing/No*

Q41: In your opinion does the evidence and judgments within learner profiles reflect the shared understanding of NC levels and standards of all KS3 teachers? *Yes/Developing/No*

Q42: Overall, how much confidence do you have in the accuracy and reliability of your existing internal moderation procedures? *No confidence/Some confidence/Complete confidence*

Please place any additional commentary on accuracy and reliability of existing internal moderation procedures.

Q43: Have any staff attended external training regarding cluster standardisation / moderation during the last year? *Yes/No*

Q43a: If Yes, briefly describe the attendance at training.

Brief description of attendance at training

Q44i: Does cluster moderation take place in the school for Oracy in English?

Q44ii: Does cluster moderation take place in the school for Reading in English?

Q44iii: Does cluster moderation take place in the school for Writing in English?

Q44iv: Does cluster moderation take place in the school for Oracy in Welsh?

Q44v: Does cluster moderation take place in the school for Reading in Welsh?

Q44vi: Does cluster moderation take place in the school for Writing in Welsh?

Q44vii: Does cluster moderation take place in the school for Mathematics?

Q44viii: Does cluster moderation take place in the school for Science?

Q45: Is cluster moderation an annual event? *Yes/No*

Q46: Briefly describe the time allocation and organisation given to cluster moderation 2011-2012?

Brief description of time allocation and organisation

Q47: What funding stream is used to support cluster moderation in 2011-12? *Brief description of funding stream*

Q48: Who is primarily responsible for co-ordinating cluster moderation meetings? *Position / role of co-ordinator(s):*

Q49: Which of your core subject teachers attend the cluster meetings? *Position / role of teacher(s):*

Q50: Do teachers who attend the meetings have opportunity to feedback cluster judgments to other KS3 teachers? *Yes/No*

Q51i: Does the cluster have standardisation portfolios in place for Oracy in English?

Q51ii: Does the cluster have standardisation portfolios in place for Reading in English?

Q51iii: Does the cluster have standardisation portfolios in place for Writing in English?

Q51iv: Does the cluster have standardisation portfolios in place for Oracy in Welsh?

Q51v: Does the cluster have standardisation portfolios in place for Reading in Welsh?

Q51vi: Does the cluster have standardisation portfolios in place for Writing in Welsh?

Q51vii: Does the cluster have standardisation portfolios in place for Mathematics?

Q51viii: Does the cluster have standardisation portfolios in place for Science?

Q52: Do standardisation portfolios have appropriate commentaries linking the evidence (i.e. samples of pupil's work) to specific level descriptions? *Yes/Developing/No*

Q53: Does the evidence and judgments within cluster standardisation portfolios reflect the shared understanding of NC level descriptions of all KS2 and 3 teachers attending the cluster meetings? *Yes/Developing/No*

Q54: Do cluster representatives take and consider a range of learner profiles to determine the 'best fit' NC levels? *Yes/Developing/No*

Q55i: Does the cluster have specific learner profiles in place for NC Levels 4 and 5 for Oracy in English?

Q55ii: Does the cluster have specific learner profiles in place for NC Levels 4 and 5 for Reading in English?

Q55iii: Does the cluster have specific learner profiles in place for NC Levels 4 and 5 for Writing in English?

Q55iv: Does the cluster have specific learner profiles in place for NC Levels 4 and 5 for Oracy in Welsh?

Q55v: Does the cluster have specific learner profiles in place for NC Levels 4 and 5 for Reading in Welsh?

Q55vi: Does the cluster have specific learner profiles in place for NC Levels 4 and 5 for Writing in Welsh?

Q55vii: Does the cluster have specific learner profiles in place for NC Levels 4 and 5 for Mathematics?

Q55viii: Does the cluster have specific learner profiles in place for NC Levels 4 and 5 for Science?

Q56: For English, Welsh (and Welsh 2nd language in English medium schools), do the overall subject levels take account of aggregation of separate AT outcomes in oracy, reading and writing? *Yes/Developing/No*

Q57: Do cluster learner profiles have appropriate commentaries linking the evidence (i.e. named pupil's work) to specific NC levels? *Yes/Developing/No*

Q58: Do teachers use the cluster learner profiles to help moderate their overall pupil assessments at the end of KS2 and KS3? *Yes/Developing/No*

Q59: Does the evidence and judgments within cluster learner profiles reflect the shared understanding of NC levels and standards of all KS2 and KS3 teachers? *Yes/Developing/No*

Q60: Overall, how much confidence do you have in the accuracy and reliability of your existing cluster moderation procedures? *No confidence/Some confidence/Complete confidence.*

Q61: Do you consider the national curriculum level descriptions useful for assessment? *Yes/No*

Q62: Do you consider current teacher assessment procedures to be an effective use of resources? *Yes/No*

Q63: Do you consider the current KS2/3 external moderation programme to be an effective use of resources? *Yes/No*

Q64: Overall, do you consider that your current Y9 assessments accurately reflect the actual ability of learners *Yes/No*

Q65: What do you consider to be the main barriers to ensuring reliable and consistent teacher assessment?

Enter comments here

Q66: What do you think might improve the quality of pupil assessments?

Enter comments here

Q67: Final comments

Enter comments here

Appendix 3: Respondents to questionnaires

List of respondents to the questionnaires sent to samples of local authorities, secondary schools, primary schools and clusters.

Local Authority Questionnaires (18)

Gwynedd & Anglesey LEAs
Denbighshire County Council
Flintshire
Wrexham County Borough Council
Powys
Pembrokeshire
Carmarthenshire
Swansea
Neath Port Talbot
RCT
CBS CONWY
Ceredigion
Sir Abertawe
Consortiwm De Ddwyrain Cymru (includes Blaenau Gwent, Caerphilly, Monmouthshire, Newport and Torfaen)

Cluster Questionnaires (17)

Bassaleg
Maesteg
St Cyres
Tredegar
Clywedog
Ysgol Bryn Elian
Caerau
Llangatwg School
St Alban's (No 2)
St Alban's
Penarth
St Teilo’s
Cwm Rhymni
Botwnnog
Tredegar
DalgylchYsgol David Hughes
Teulu’r Emlyn

Primary Questionnaires (48)

Ysgol LLanddulas
Ysgol Y Plas
Christchurch CP School
Bryn Hedydd
Bwlchgwyn
Victoria CP
Barker’s Lane
Gwenfro Community Primary
Acton Park Primary
Llanidloes Primary
Brynmill Primary School
Grange Primary
Whitestone Primary School
Blaengwrach Primary School
Cwmnedd Primary
Ynysfach Primary School
Cwmfelin Primary
Garth Primary School
Caerau Primary School
Cogan Primary
Cwmfelin Primary School
Brynnau Primary
Glyn Gaer Primary
Ystrad Mynach Primary
Undy Primary
Rogiet Primary
Green Lane Junior

Secondary Questionnaires (18)

Alun School
YsgolRhosnesni High School
Ysgol Clywedog
Ysgol Gyfun Emlyn
Bishop Gore School
Bishop Vaughan Catholic School
Llangatwg Community School
Pencoed Comprehensive School
Afon Taf High School
Lewis School Pengam
Caldicot School
Glyn Derw High School
St. Teilo’s Church in Wales High School
Ysgol David Hughes
Botwnnog
Preseli
YsgolTre-Gib
Cwm Rhymni

Primary questionnaires (continued)

Archbishop Rowan Williams VA CIW
Primary
Llanmartin Primary School
Pentrepoeth Primary School
Brynsiencyn
Dwyran
Llangoed
Ysgol Gynradd Llandegfan
Tudweiliog
Eglwyswrw
Maenclochog
Gynradd Gymraeg Glan Cleddau
Clydau
Brynsaron
Ysgol Eglwys Llanllwnik
YGGD Cwmgors
Y G GRhosafan
Y G GCwmnedd
Ysgol Y Wern
Ysgol Gymraeg Pontardawe
YsgolTyle'rYnn
YGG Caerphilly

Appendix 4: Overview of the status of school-based assessment in 10 jurisdictions

NB: The cycle for change in assessment systems is quite short; the currency of entries was that at the time of drafting the report.

Country/ province	School-based assessment (SBA)	Practice
Africa	Increasingly seen as a valuable tool for South Africa and other African nations such as Nigeria, Ghana and Zambia	Illustrative only Grade Reception to 3 is internal, set and marked by the teacher and moderated externally within. Continuous and formative assessment for the Foundation Phase, from grade R to 3, will be internal, set and marked by the teacher and moderated externally within guidelines of the Provincial Education Departments. Summative assessment at Grade 3 will be external, administered as part of the national assessment programme Continuous and formative assessment, as well as summative assessment for the Intermediate Phase, from grade 4 to 6, will be internal and external, set and marked by the teachers and moderated externally within guidelines of the Provincial Education Departments. Grade 6 summative assessment will form part of the national systemic evaluation. Continuous and formative assessment, as well as summative assessment for the Senior Phase, from grade 7 to 9, will be internal and external, set and marked by the teachers and moderated externally within guidelines of the Provincial Education Departments. Grade 9 summative assessment will form part of the national systemic evaluation.
Australia	Has been established practice in Australia for over twenty years. In Queensland, where SBA was	The practice in Queensland and the research that has been conducted within the system is discussed in this report.

Country/ province	School-based assessment (SBA)	Practice
	introduced in the 1970s, teacher-based assessment is used for all assessment in secondary school, even for high-stakes purposes. The Australian Capital Territory also uses SBA for assessment for senior secondary level. Other states such as New South Wales have incorporated large-scale, school-based assessment into their public examinations.	
Canada	Has been the standard mode of assessment in Canadian schools for many years with teachers taking responsibility for all assessment processes and judgments at the school-level. For an example see the work on school based assessment in Saskatchewan.	In Canada the curriculum implicitly requires student performance to be assessed, and in principle, the class teacher is responsible for this work. Reform of the curriculum has considerably increased work demands on students and teachers.
Northern Ireland	Strong policy commitment to school-based assessment and assessment for learning.	
New Zealand	http://www.nzqa.govt.nz/ncea/acrp/secondary/5/5.html Has a long history of school-based assessment in senior secondary school, and has developed a wide variety of teacher support materials and associated research studies.	
Scotland	Much interesting work on teacher-based assessment is being conducted by the Scottish AfL group, supported by the Ministry of Education and involving many schools.	
Scandinavia	Finland and Sweden both have long-established school-based assessment systems utilising a wide range of open-ended authentic tasks and challenging classroom-based assignments. Such SBAs, embedded	Finland's assessment system relies heavily on self-assessment and other mechanism that put relatively little emphasis on external measures. Marking is based on teacher assessment.

Country/ province	**School-based assessment (SBA)**	**Practice**
	in the curriculum, are often cited as an important reason for the high levels of educational achievement in those countries.	There is no final examination at the end of compulsory schooling Sweden grades from Year 8 with responsibility lying entirely with the teaching staff. No external control.
Singapore	Ministry of Education in Singapore adopted an official policy of assessment for learning and is encouraging teachers to experiment with different forms of SBA though the school system is still dominated by externally-set and assessed examinations.	
The Netherlands		Just before completing, primary school students sit a test assessing their skills in Dutch language numeracy, working with data and environmental studies. Final national test in lower secondary education has been abolished.
USA	Many states have developed and implemented SBAs as a supplement to or complement to national and state standardised testing (e.g. Iowa), although the standing of such assessments in the wider educational community is still low.	

Sources: http:/www.hkeaa.edu.hk/en/sba; Parkes & Maughan (2009)

Appendix 5: Advantages and disadvantages of some approaches to social moderation

Approach	*Advantages*	*Disadvantages*
External moderators	Offer authoritative interpretations of standards Can carry the standards from site to site and assessor to assessor Can offer advice on assessment approaches and procedures Can observe actual assessments not just view folios Be a trouble-shooting resource for assessors to draw on Can induct novice assessors quickly into high quality assessment	Substantial costs involved (salary, base office, travel accommodation, communications, training, moderator conferences etc.) Logistic problems to be overcome in covering all assessors adequately Authoritative interpretations not always right or appropriate External authority can be rather stultifying
External moderation panels	Are likely to make more consistent decisions than individuals Can cover a larger group of assessors than a single moderator Can offer more comprehensive advice on assessment approaches and procedures (by being able to draw on a wider cross-section of examples) Represent collective authority rather than single person authority Provide powerful professional development for those involved	Substantial costs involved (travel, accommodation, communications, training, moderator conferences etc.) Need some full-time officers to organise the process

Assessor meetings	Opportunity for direct comparison and sharing among assessors Less judgmental atmosphere than for external moderators or panels Personal ownership of any new insights and understandings and ideas Opportunity to develop networks of support for resolving new problems Powerful professional development for those involved Mandatory participation and public scrutiny of one's own practices and judgments could encourage serious attention to the issue of quality Useful supplement to other quality control procedures	Meetings need to be organised and facilitated (with attendant costs) Would assessors come to meetings if they were voluntary? (or alternatively, what sanctions could be used to encourage participation?) Need to be some full-time officers to organise the process (though the costs for this would probably be less than for moderators or panels) Substantial logistic obstacles to covering all assessors No guarantee of quality outcomes without a formal approval process Would not satisfy the need for quality control in a high stakes situation
Assessor partnerships	Can be locally organised and do not need bureaucratic support Few external costs apart from any promotional material Can be promoted as providing mutual benefit to partners Participation is personally empowering, reducing uncertainties and enhancing assessment capabilities	Possible that nothing may happen if it is voluntary Participation depends on individual initiative and intrinsic motivation Level of involvement depends on personal commitment Successful partnerships require personal compatibility May need to be some professional support and resources Partners may simply reinforce each other's errors and misconceptions

Source: Maxwell (2006)