assessment europe

A focus on teacher assessment reliability in GCSE and GCE

Sandra Johnson

Ofqual/11/4807
January 2011

Ofqual This report has been commissioned by the Office of Qualifications and Examinations Regulation

Preface

The principal aim of this project, carried out under Contract 3224 within Ofqual's Reliability Programme, was production of a comprehensive and critical review of the literature on the reliability of teacher assessment, which emerged as extremely sparse, with a particular focus on assessments conducted in England, Wales and Northern Ireland.

Reliability was pre-defined as follows:

> *Reliability refers to the consistency of outcomes that would be observed from an assessment process were it to be repeated. High reliability means that broadly the same outcomes would arise. A range of factors that exist in the assessment process can introduce unreliability into assessment results. Given the general parameters and controls that have been established for an assessment process – including test specification, administration conditions, approach to marking, linking design and so on – (un)reliability concerns the impact of the particular details that do happen to vary from one assessment to the next for whatever reason.*

Four potentially important sources of measurement error were identified: occasion-related, test-related, marker-related and grading-related.

The following research questions were to be specifically addressed:

1. What is the nature of the tasks assigned to teachers as the basis for forming judgements about pupils' knowledge, skills or abilities?
2. What rules and procedures are in operation that guide or standardize the conditions under which pupils produce the evidence that their teachers use to assess them?
3. What is the nature of pupils' work – reports or artefacts – that teachers are required to assess, and what rules or requirements govern these?
4. What is the nature of any formal marking schemes that teachers use to arrive at their assessments, and what procedures are in place for checking reliability?
5. What methods are employed to check on the reliability of sets of submissions, and what are the criteria that would trigger action to address discrepancies?
6. What scaling or other adjustment methods are employed to the assessments before aggregation with test results to arrive at final awards, and what are the potential effects of these on the overall reliability of those final awards?

The focus of the review is teacher summative assessment in GCE and GCSE examinations. In England, Wales and Northern Ireland there are just five awarding bodies – also traditionally known as examining boards – that offer these qualifications. In England these are the Assessment and Qualifications Alliance (AQA), Edexcel, and the Oxford, Cambridge and RSA Examinations (OCR). For Wales there is the Welsh Joint Education Committee (WJEC) and for Northern Ireland the Council for Curriculum, Examinations and Assessment (CCEA). The websites of the five examining boards were scrutinised for relevant information, and key board professionals were contacted for further detail as necessary.

This search for information about current examining board practice was supplemented by a general review of the literature on teacher assessment, with a particular search for empirical reports on the *reliability* of summative teacher assessment. The search covered the most relevant educational research and assessment journals, including *Assessment in Education, British Journal of Educational Research, Education 3-13, Educational Assessment, Educational Research, Educational Research and Evaluation,* and *Research Papers in Education*. It also embraced the 'grey literature' to be found on the websites of the examining boards, university research groups, research funding bodies, and so on. Previous review reports on this and closely related topics were also consulted: these included reviews commissioned by the Qualifications and Curriculum Authority (QCA), now the Qualifications and Curriculum Development Authority (QCDA), and the Office for Qualifications and Examinations Regulation (Ofqual).

Acknowledgements

The literature on teacher assessment is large. But within this literature reports on the levels of reliability that have been achieved with various forms of teacher assessment, as opposed to expressions of belief that these should be high, are relatively few and far between. This present report draws on previous reviews of the literature on teacher assessment, including teacher assessment reliability, particularly those of Wilmut, Wood and Murphy (1996) for the Qualifications and Curriculum Authority, Harlen (2004) for the EPPI Centre, and Stanley, MacCann, Gardner, Reynolds and Wild (2009) for the Qualifications and Curriculum Authority.

Information about the current form and role of teacher assessment within GCSE and GCE examinations was gathered mainly from the website of the Office for Qualifications and Examinations Regulation (Ofqual) and from those of the six awarding bodies that offer these academic qualifications in the UK: the Assessment and Qualifications Alliance (AQA), the Council for the Curriculum, Examinations and Assessments (CCEA), Edexcel, Oxford, Cambridge and RSA Examinations (OCR), the Scottish Qualifications Authority (SQA), and the Welsh Joint Education Committee (WJEC).

Key individuals in the examining boards were also contacted for information about practices and procedures. I take this opportunity to express my gratitude to Martin Taylor of AQA , Ann Comac of CCEA, Alistair Drewery and Jeremy Pritchard of Edexcel, Elizabeth Gray of OCR, and Jo Richards and Margaret Franks of WJEC for their invaluable input. I thank Paul Black of Ofqual's Technical Advisory Group for helpful comments on an earlier draft, and Jo Taylor of Ofqual for her efficient and flexible management of this project from conception to report publication.

Contents

1 Introduction

1.1 The outcomes of learning and the assessment challenge

If the outcomes of learning could be assessed as easily and as accurately as students' heights or weights then there would be little call for reviews such as this on assessment reliability. As it is, the literature on assessment in general, and on teacher assessment in particular, is large and continually growing, even if rather little of this extensive literature focuses directly on the issue of assessment reliability. So why does the assessment of learning engender such research interest and continue to pose interesting challenges for those involved in this activity?

The principal reason has to do with the fact that some of the outcomes of learning that we want to assess, or measure, cannot be directly accessed. We can look at a student and even without an appropriate measuring device have a reasonable idea of how tall that student is. We cannot do the same with knowledge, skills, abilities or attitudes. These only become visible to us when demonstrated by the student in some way, typically by responding to a question or carrying out some action in response to a request or instruction. And even then we cannot always be sure that the evidence provided to us in the form of responses or actions is sufficient to infer with confidence anything about the individual's general state of knowledge, skill or attitude.

If we ask a young person for the date of the Battle of Waterloo or for the chemical symbol for mercury then provided we are given an answer we will have some assessment-relevant information about that individual. We will know that at that particular point in time, under the circumstances in which the question was posed, the student knew or did not know – perhaps knew once but could not now recall – the battle date or the chemical symbol for mercury.

But even for such factual knowledge the assessment challenge increases as we attempt to move away from isolated atomistic assessment exercises in attempts to measure knowledge, abilities and skills more globally. Knowing the date of the Battle of Waterloo is one thing, but this tells us little about the student's broader state of knowledge of that battle, or of that period, or of history in general. Knowing the chemical symbol for mercury gives us no information about how many other chemical symbols the student knows and can correctly attach to the relevant chemicals. It also tells us nothing about the student's knowledge and understanding of chemistry more generally, theoretical or practical.

Now suppose that we ask the student to multiply 206 by 75, or to give us the past participle of the French verb *finir.* These two new requests move us into even more complex territory. Should the student provide the right answer to the multiplication question then we would be able to infer that the student possessed both the procedural knowledge necessary to perform integer multiplication tasks *and* could apply that knowledge skilfully – unless, of course, this particular calculation had been memorised, which would be unlikely. A wrong answer is less easy to interpret. For had the answer been wrong then unless we could directly question the student we would not know whether this resulted from a careless mistake in application, or whether the skill of integer multiplication could not be demonstrated because the underlying procedural knowledge was lacking.

If the student offered the correct response for the past participle of *finir* there are again different possible inferences that can be drawn from the assessment evidence. If the student had come across this verb before then we would know that this particular piece of grammatical knowledge had been absorbed and retained. On the other hand, if the student had never met *finir* before, but had been taught the general grammatical rules for regular "ir" verbs, then the correct response to the question would tell us that the student was not merely recalling a previously taught fact, but was using more abstract assimilated knowledge about rules of language to make a correct prediction about the form of the past participle in this case.

If the question had been posed to the student in French rather than in English, and had the student responded correctly on that occasion, then we would have more general evidence about the student's linguistic competence, which would extend beyond knowledge of one atomistic element of grammar to more general comprehension skills. However, should the response have been wrong on this occasion then we would not know whether this was because the student had forgotten the past participle of *finir*, or could not apply taught language rules to infer the past participle correctly, or had insufficient capability in the language generally to understand the question.

These few examples already demonstrate some of the interesting problems that assessors face in this particularly challenging occupation. Knowing some subject-specific facts tells us little about the breadth or depth of a student's general knowledge in that subject field. Being able to solve correctly a single multiplication question tells us nothing about the security of the student's ability with multiplication, and nothing at all about other aspects of numeracy or mathematics. Knowing the past participle of *finir* tells us nothing about the extent of the student's knowledge of French grammar and vocabulary, and certainly nothing at all about the student's facility to interact orally in the language.

Yet even if we could formally assess every possible fact and every possible skill that is embodied in a subject curriculum we would need to find a way to pull all of the resulting evidence together in some way, to produce summative statements about individuals that might be meaningfully interpreted. It was the requirement that teachers assess large numbers of relatively small aspects of subject performance that proved the greatest difficulty when national curriculum assessment was first launched in England, Wales and Northern Ireland 20 years ago (Sainsbury 1994, Clarke & Gipps 1998).

Beyond meaningful interpretation, which is an aspect of assessment validity, we need to know how much relevant evidence we would need to gather about a student for the resulting assessment to be dependable. Would 20 factual questions about a particular period in history produce the same assessment outcome for a student as 50 questions? What would be the outcome if 10 of the original questions were substituted with 10 different ones? If any of the questions required an extended written response, would the assessment outcome vary depending which person marked the student's response? If an oral assessment stimulus were replaced with another would the student's performance differ? And so on. Consistency in outcome across assessment conditions, i.e. repeatability, is essential for assessment reliability, and assessment reliability is a necessary condition for assessment validity.

It is this need adequately to summarise the outcomes of an individual's learning at specific points in time that challenges teachers and professional assessment agencies alike. How we address the challenge, and how much time, effort and resource we expend in the process, depends in great part on the purpose of the assessment, and the purpose can vary from informal use in the classroom to student certification to system monitoring (see Newton 2010 for a comprehensive account of the different possible purposes of assessment).

1.2 Teacher assessment and tests

The principal advantage of teacher assessment over impersonal written testing is generally acknowledged to be the fact that teachers are with their students for long and continuous periods of time, constantly interacting with them inside and outside the classroom, posing questions to develop their thinking and observing them as they carry out assigned tasks and activities. In consequence, teachers are assumed to have a more comprehensive, and by default more valid, picture of their students as learners and achievers than any set of test results can alone provide. Formal tests are by their nature limited in scope. They inevitably sample the knowledge/skill curriculum more narrowly than any teacher might, and they therefore lack that ability to provide a truly holistic view of development and achievement.

For their part, tests have some advantages over teacher assessment. Their degree of assessment validity, for example, is relatively transparent, and reliability can (at least in principle) be checked. Both the range of knowledge and skill being assessed and the relative importance being given to various different aspects within a curriculum, or to different intellectual or personal qualities, are usually clear from a review of the content of a test and its associated mark scheme. Teacher assessment does not always lend itself to this degree of transparency, especially when the scope of the assessment is wide, for example 'science knowledge and understanding', or relatively loosely defined, 'hospitality skills' perhaps, or the assessment itself is highly subjective, 'artistic creativity' for instance.

Despite guidance, different teachers might operationalise their assessment of a subject domain differently, perhaps giving greater emphasis to some aspects than others – investigation skills versus factual knowledge in science, for example, or computational skills versus graphical skills in numeracy. They might also judge the evidence in different ways, using different criteria and different expectations of standards. Little research has apparently been carried out to investigate these possibilities, a point noted by Harlen (2004) in a comprehensive review of what emerged to be a rather sparse academic literature on the reliability of teacher assessment.

Equally importantly, if well constructed, tests can be demonstrated to be unbiased, in the sense of not favouring one type of candidate, or one student personality, over another. Teachers, working as they do within the human dynamics of the classroom, are potentially influenced both in their teaching interaction and in their assessment activity by student characteristics other than those that are in principle being assessed (Morgan & Watson 2002; Harlen 2004, 2005; Martinez, Stecher & Borko 2009). These 'construct irrelevant' characteristics include the student's gender (Lafontaine & Monseur 2009), ethnicity (Burgess & Greaves 2009), socioeconomic status (Hauser-

Cram, Sirin & Stipek 2003; Wyatt-Smith & Castleton 2005), EAL and SEN status (Thomas, Madaus, Raczek & Smees 1998; Reeves, Boyle & Christie 2001), and personal qualities, such as behaviour and effort (Bennett, Gottesman, Rock & Cerullo 1993; Morgan & Watson 2002; Wyatt-Smith & Castleton 2005). The same phenomenon has been noted with respect to workplace assessors in vocational education and training (Wolf 1995).

Teachers are widely assumed, and have been shown, to be very capable of rank ordering their own students in the same way that an external test would (Martinez, Stecher & Borko 2009). In other words, correlations between teacher assessments and test results are typically high for individual teachers working within the contexts of their own classrooms. But across classroom, schools, subjects and pupil year groups the association can be inconsistent (see Hoge & Coladarci 1989 for a review of research). It is this type of inconsistency that causes concern where teacher assessment features prominently in regional or national assessment systems (Wijkstrom 2006; Stanley et al 2009), and leads to the need for moderation in some form.

Despite its flaws, teacher assessment does in principle have a role to play in high-stakes examinations, in providing assessments that more fully span the intended subject domain than any test alone can do. In particular, teacher/instructor assessment is essential for contributing to validity in situations where traditional written testing is not a possibility, and performance assessment is practised: the assessment of oral skills, creative writing, bricklaying skills, and so on.

Stobart (2009), for example, speaking of national curriculum assessment in England, notes that national tests:

> … sample only a limited part of the curriculum, for example ignoring speaking and listening in English, and so construct underrepresentation becomes a key validity problem. A simple solution here would be a name change, for example, Reading and Writing rather than English. The more radical option is to give teacher assessment a role that counts – so that all the main components of the curriculum can be assessed. (Stobart 2009, pp.174-175)

Harlen (2007) shares this view:

> Teachers' judgements can, when moderated, provide more accurate information than external tests because they can cover a much wider range of outcomes and so provide a more complete picture of students' achievements. (Harlen 2007, p.138)

Wiliam (2001) extends expectations of the potential benefits of teacher assessment beyond greater curriculum coverage, i.e. higher validity, to include improved assessment reliability. Again in the context of national curriculum assessment, he claimed that:

> By using a teacher assessment, we would in effect be using assessments conducted over tens, if not hundreds, of hours for each student, providing a degree of reliability that has never been achieved in any system of timed written examination. (Wiliam 2001, p.19)

and

> The key to improved reliability lies with increased use of teacher assessment, standardized and moderated to minimize the potential for bias. (Wiliam 2001, p.20)

These authors, and others, are right in claiming that formal tests typically, sometimes inevitably, focus on a part only of the intended curriculum, and that involving teachers in the assessment process would be a way of overcoming this particular type of assessment bias. Few would argue against claims that the involvement of teachers in summative assessment offers the possibility to increase validity, in the sense of covering a multifaceted curriculum more comprehensively.

But how justified are the assertions about a positive impact on assessment reliability? Wiliam's first assertion is particularly strong, and can readily be questioned. For what is this 'degree of reliability' that has *neve*r been achieved in *any* system of timed written examinations? What type of examinations, in what subjects – mathematics, history, creative writing? In any subject, badly designed and unduly short tests can certainly have poor measurement properties, leading to low score reliability. But not all tests in all subjects are poorly designed, and, equally importantly, not all aspects of all subjects will lend themselves to reliable teacher assessment.

At this point in time there is no credible evidence to support or to refute the claim of superiority for teacher assessment over tests, on the basis that teachers can base their assessments on the cumulated evidence of very large numbers of assessment opportunities. Indeed, doubt has been expressed that teachers could in practice succeed in assessing individual students over tens or hundreds of hours, as Wiliam assumes, given the number of subjects to be taught and assessed in busy classrooms (MacCann & Stanley 2010). And even if teachers could feasibly achieve this amount of individual student assessment, it must be questionable whether higher assessment reliability would result – this would depend on the degree to which some form of standardisation and moderation could indeed "minimise the potential for bias"? Unfortunately, once again there is little if any empirical evidence available at this point to judge the validity of this belief one way or the other.

Increasing potential assessment validity by covering the curriculum more broadly through teacher assessment will not necessarily lead to an increase in overall assessment reliability. Indeed, while not an inevitable consequence, reliability might rather be reduced, so that any apparent gain in validity will be spurious.

There is little doubt that teachers represent a wealth of knowledge about students' achievements and capabilities that is indispensable in the assessment of pupils' learning progress, and which *in principle* could be harnessed for the good in summative assessment in high-stakes examinations. But what do we know *in practice* about how effective such harnessing might be? There are issues surrounding teacher assessment that have to do with potential bias, application of different, sometimes personal, assessment criteria, and differences in the available evidence base when implemented curricula and standards of judgement differ from class to class and school to school.

Teacher assessment has been widely used in UK high-stakes examinations for some time, as Chapter 2 confirms. The question is what impact might the use of teacher assessment be having on assessment reliability in this high-stakes context?

1.3 The focus of this review

Whether high-stakes examinations are based on external tests alone or on a combination of tests and internal teacher assessment, the validity and reliability of those examinations must be demonstrated, as far as this is possible to do. Awarding bodies have practices and procedures in place that in principle assure assessment validity and reliability, both of their own externally marked tests and of any teacher assessed components that contribute to whole-examination grades. But what are these practices and procedures for teacher assessment, and how well do they perform in terms of achieving assessment reliability? This is the principal question addressed in this review.

Reliability was pre-defined for the review as follows:

> *Reliability refers to the consistency of outcomes that would be observed from an assessment process were it to be repeated. High reliability means that broadly the same outcomes would arise. A range of factors that exist in the assessment process can introduce unreliability into assessment results. Given the general parameters and controls that have been established for an assessment process – including test specification, administration conditions, approach to marking, linking design and so on – (un)reliability concerns the impact of the particular details that do happen to vary from one assessment to the next for whatever reason.*

and four potentially important sources of measurement error were identified, *viz*. occasion-related, test-related, marker-related and grading-related.

It is important to stress that this review is concerned with the reliability of teachers' *summative* assessments where these contribute to high-stakes examinations. There is no attempt to extend consideration to formative assessment practice, or 'assessment for learning' (Black & William 1998; Gardner 2006), which has been defined as "the use of day-to-day, often informal, assessments to explore pupils' understanding so that the teacher can best decide how to help them to develop that understanding" (Mansell, James and the Assessment Reform Group 2009, p.9).

After overviewing the nature and extent of teacher assessment in the different testing and examination systems currently operating in the UK, the review focuses specifically on internal assessment within GCE and GCSE examinations, addressing in particular the following given questions:

1. What is the nature of the tasks assigned to teachers as the basis for forming judgements about pupils' knowledge, skills or abilities?
2. What rules and procedures are in operation that guide or standardize the conditions under which pupils produce the evidence that their teachers use to assess them?
3. What is the nature of pupils' work – reports or artefacts – that teachers are required to assess, and what rules or requirements govern these?
4. What is the nature of any formal marking schemes that teachers use to arrive at their assessments, and what procedures are in place for checking reliability?

5. What methods are employed to check on the reliability of sets of submissions, and what are the criteria that would trigger action to address discrepancies?
6. What scaling or other adjustment methods are employed to the assessments before aggregation with test results to arrive at final awards, and what are the potential effects of these on the overall reliability of those final awards?

Questions 1, 2 and 3 and the first part of question 4 are addressed in Chapter 3, while the second part of question 4, along with questions 5 and 6 are the focus of Chapter 4. Chapter 5 offers a summary of the salient points of the review, and reflects on the implications of teacher assessment for the reliability of certifying examinations.

2 Teacher assessment within UK testing and examining systems

2.1 The UK's testing and examining systems

High-stakes assessment in the UK extends in one form or another from the early primary school, where it is driven by a strong accountability agenda that affects teachers and schools rather than pupils, to qualifications-awarding examinations taken by secondary school students and their peers in vocational education and training (for a complete overview see Isaacs 2010). At all levels of the system teachers are involved in assessing their own students, with the results of this teacher assessment frequently feeding directly into student certification.

The UK's high-stakes qualifications system is strongly market driven. Over 120 different awarding bodies offer a total of well over 6000 nationally accredited academic, vocational and occupational qualifications to candidates in schools, colleges and businesses. In England the system is regulated by the Office for the Regulation of Qualifications and Examinations (Ofqual), in Wales by the Department for Children, Education, Lifelong Learning and Skills (DCELLS), in Northern Ireland by the Council for the Curriculum, Examinations and Assessment (CCEA) and in Scotland by the Scottish Qualifications Authority (SQA).

Just six awarding bodies, known historically as 'examining boards', offer academic school leaving qualifications, with the same subject qualification at the same level typically being offered by more than one board. For England the examining boards are the Assessment and Qualifications Alliance (AQA), Edexcel, and the Oxford, Cambridge and RSA Examinations Board (OCR). Wales, Northern Ireland and Scotland have, respectively, the Welsh Joint Education Committee (WJEC), CCEA and SQA. Currently the two principal school leaving examinations are the *General Certificate of Secondary Education* (GCSE), normally taken at the end of Year 11 (16 year olds) and the *General Certificate of Education* (AS/A level), taken as a school leaving examination at the end of Year 13 (18 year olds) and used for university entry. But there are more recent options now available, such as the recently launched Diploma, that attempts to combine academic and vocational elements, and the Welsh Baccalaureate. Scotland has its own national qualifications.

The remaining 100+ awarding bodies, many of which are small businesses but some of which are large companies with household names, provide thousands of competence-based vocational and occupational qualifications in hundreds of different fields of employment, from law and legal services through hotel management and child care to glass manufacture and equine transport. Like the GCSE, the system of *National Vocational Qualifications* (NVQ), and their Scottish equivalents (SVQ), was launched in the late 1980s, and has continued to grow in scope and scale ever since.

Also in the late 1980s, the national curriculum assessment system was set up by the London-based government, to assess and monitor the new national curriculum that was simultaneously introduced into England and Wales. Since then, blanket assessment of pupils has taken place annually at the end of 'key stages' in schooling – essentially at ages 7, 11 (end of primary) and 14, managed by the Qualifications and Curriculum Development Authority (QCDA - previously the Qualifications and Curriculum Authority (QCA). A recent important development is the abandonment in

2003 of the mandatory testing of pupils at the end of key stages by the now devolved Welsh Assembly, with formal blanket testing replaced by moderated teacher assessment (Daugherty 2007). Another is the more recent (2009) abandonment of blanket key stage 3 testing in England itself, to be replaced by a sample-based survey yet to be introduced (see Isaacs 2010 for further detail).

Until recently, Northern Ireland assessed its own national curriculum at the end of the same key stages as England and Wales. But here assessment at ages 7 and 11 was always based on teacher judgement rather than tests. And while the assessment of English, mathematics and science at the end of key stage 3 was based on external tests, paralleling the situation in England and Wales, there is currently no cohort assessment at age 14 in Northern Ireland (Elwood 2009).

In Scotland a voluntary system of submission of teacher assessments was introduced in the early 1990s, with teacher judgements being made for every pupil from 5 to 14 against a six-level criterion-referenced national curriculum framework. This system, which rapidly evolved into census submissions, ran alongside Scotland's sample-based monitoring programme – the Assessment of Achievement Programme (AAP) – until both were replaced in 2005 by the Scottish Survey of Achievement (SSA). The SSA was a sample-based monitoring programme that used formal testing to estimate the proportions of pupils at the various 5-14 levels at the monitored stages (Hayward 2007; Johnson 2007), as the AAP had done, but which also gathered teachers' level judgements for the sampled pupils tested in the survey (though the rationale for this aspect of information gathering was never made clear). The SSA was in 2010 replaced by the Scottish Survey of Literacy and Numeracy (SSLN), which will report against the new "Curriculum for Excellence" level-based framework (Scottish Government 2010). Teachers' judgements, whatever form these might eventually take, are not to be gathered in the new system monitoring programme.

2.2 Teacher judgement in national curriculum assessment

England's national curriculum covers the full period of compulsory education from early primary to Year 11 (age 16). Pupils' progress through the curriculum is assessed informally by teachers whenever they wish, but is *required* to be assessed in core subjects at the end of each of three 'key stages', with reference to an eight-level progression framework. Key stage 1 covers learning in the early years foundation stage and primary years 1 and 2, key stage 2 covers learning in years 3 to 6, key stage 3 covers years 7 to 9. Years 10 and 11 comprise key stage 4, at the end of which pupils take their GCSEs. The results of the required assessments have been used, and continue to be used, for a variety of different purposes, including reporting on pupils' learning progress to parents and other teachers, and using aggregated results in school, authority and system evaluation (Newton 2010; QCDA 2010a, b, c).

Since its first incarnation the national curriculum assessment system has undergone a variety of evolutions – see Whetton (2009) for a useful and insightful historical account. Initial requirements of all teachers at key stages 2 and 3 to assess each of their pupils against a long list of aspects within each core and non-core subject were soon abandoned because of protests about workload and manageability. External tests were introduced in the core subjects of English (reading and writing), mathematics and science, soon to be themselves replaced with shorter more cost-effective and less

disruptive versions. Teachers' level judgements continued to be gathered, in principle to carry equal weight alongside the test results in attainment reporting, but in practice given a secondary status.

The most recent, and arguably one of the most radical, changes in England's national curriculum assessment system is the abandonment of cohort testing in English, mathematics and science at the end of key stage 3 in favour of a sample survey of English and mathematics attainment, due to begin in 2011 (QCDA 2010a). In the meantime, teachers' level judgements will continue to be gathered in English, mathematics, science and modern foreign languages. In each case, these will be level judgements for each of a number of separate attainment targets and for the subject as a whole (determined as a weighted average of target level judgements). Teachers are advised that they:

> …should base their judgements on the level descriptions in the national curriculum. They should use their knowledge of a pupil's work over time to judge which level description is closest to the pupil's performance. In reaching a judgement, they should take into account written, practical and oral work, as well as classroom work, homework and the evidence from any tasks and tests. (QCDA 2010a, p.10)

Changes are occurring at the end of key stage 2 also, in that cohort testing in science is giving way to sample-based testing in this subject from 2010 (QCDA 2010b), though blanket testing in English and mathematics remains. Teachers' level judgements for English, mathematics and science will continue to be gathered.

There is no external testing at the end of key stage 1. Teachers will continue to provide level judgements for each of their pupils (QCDA 2010c), for reading, writing, speaking and listening separately, for mathematics, and for each of four science attainment targets. For their more able pupils teachers are required to use any number of available tasks and tests provided by the QCDA to support them in forming their judgements. Task and test results, however, are not gathered centrally.

A system of moderation is in place for key stage 1 teacher assessment, organised by local education authorities. This, however, would seem closer to verification than moderation, understandably in the circumstances. Moderators periodically make half-day visits to schools, and talk with the head teacher and class teachers, to establish that they are fully conversant with the assessment requirements and diligent in implementing them. They might also speak for a few minutes with a handful of individual pupils, though in doing this they could not be expected to ratify with any great confidence the particular level judgements given by class teachers to every Year 2 pupil in the school.

In addition to assessment at the end of key stage 1, schools are required, with certain specific exceptions, to complete an early years foundation stage (EYFS) profile in the final term of the academic year in which a child reaches the age of five. This is a summative record of development in 13 different areas, based on "cumulated observational evidence recorded over the course of the year" (QCDA 2010c, p.12), and is also moderated/ratified (see, for example, Cale & Burr 2007).

The entire national curriculum assessment system has been the subject of both interest and frustration for assessment professionals and academics in the UK ever since its inception. The reliability of national curriculum tests has been questioned (Hutchison & Schagen 1994; Wiliam 2001, 2003) and explored (see, for example, Newton 2009). And attempts have been made to evaluate the reliability of teachers' judgements, which have never been subject to moderation at key stages 2 and 3, by reviewing rates of agreement with test results (see, for example, Reeves, Boyle & Christie 2001).

One difficulty faced when exploring the reliability of teacher assessments through comparison with national curriculum test results is a question mark over the reliability, and indeed the validity, of the tests themselves. One short test, such as the annual reading test, provides little scope for taking into account in reliability measurement the impact of the particular questions comprising that test, or, at least as importantly in this context, the impact of the source material on which the questions are based. If the source material, or the questions based on it, had been different then what effect would this have had on the attainment outcome? If the topic for a single writing assignment had been substituted by another what difference would this have made to the test results for individual pupils and entire cohorts? Another weakness for reliability estimation is that the external tests are marked by a single marker, leaving no scope for marker effects and their impact to be investigated. There has, though, been a recent review on the issue of marker reliability and its implications for national curriculum assessment (QCA 2009a), so that perhaps there will be change in this respect.

A second and extremely important difficulty with comparing teachers' level judgements with national curriculum test results is that the two measures, whilst ostensibly of the same construct, are not necessarily independent. Schools have a generous window within which to submit teachers' level judgements, and, crucially, this window extends beyond the testing date. Judgements submitted after testing has occurred could well have been 'contaminated' by knowledge of the test results. But unfortunately no record is kept of the date on which individual schools submit their judgements, so that an analysis based on independent assessments is not possible to undertake. Reeves et al. (2001) noted this, as did Rose (1999), in the report of the independent scrutiny panel on the 1999 key stage 2 tests.

This lack of independence between teachers' level judgements and test results could in part explain the relatively high rates of agreement that were recorded by Reeves and colleagues (2001) in their research, which focused on the results of national curriculum assessment in English, mathematics and science at the end of key stage 2 in 1996, 1997 and 1998 for samples of 11 year olds ranging from almost 1,300 pupils to over 2,600 per year. Interestingly, exact agreement rates on level classification between teachers and tests were consistently around 75% in all three years in all three subjects. Where teachers' judgements and test-based classifications differed, teachers tended more often to award higher levels than the tests in mathematics and lower levels than the tests in English and science.

In Scotland, where a similar level-based progression framework has been in operation for around 15 years, rates of agreement between independently produced teachers' level judgements and test results in reading, mathematics and science – the latter originating in the sample-based Scottish Survey of Achievement (SSA) – were also

investigated (Johnson & Munro 2008). In this context agreement rates were found to be lower than those reported by Reeves et al. (2001) for national curriculum testing in England.

The test-based level classifications of pupils in the SSA were arguably more reliable than those produced through national curriculum testing in England, for the following reason. In the SSA multi-item reading tasks, and atomistic mathematics and science items, were individually assigned to one or other of the six levels in the 5-14 progression framework prior to survey use, through the majority decisions of three or more teacher validators working independently. Several dozen reading tasks and several hundred atomistic mathematics or science items were used in any one survey, multiple-matrix sampling used to distribute tasks and test booklets among the sampled pupils. In the reading assessment, pupils were each randomly assigned three reading tasks (or tests), one at each of three consecutive levels appropriate to the age-group. In mathematics each pupil was randomly assigned two 'interchangeable' (i.e. domain-sampled) tests, from among several used that year, the set of items comprising each test spanning the appropriate three levels so that the pair of mixed-level tests that any particular pupil attempted actually embodied three single-level tests (for full details of the mathematics assessment strategy see Johnson 2008).

Pupils' performances on the three single-level tests determined their level classification for that subject, a fixed cut-score of 65% being applied to each level test to classify the pupil as having demonstrated attainment at that level or higher (this cut-score had been agreed several years earlier by subject specialists as being appropriate for the purpose). To give an indication of the reliability of the tests for classifying pupils by level, the overall percentages of pupils who reached the cut-off score at one level having failed to reach it at a lower level were approximately 6.5% for reading, 1.5% for numeracy and 4% for science.

Rates of exact level agreement by both teachers and tests were higher for numeracy/mathematics, at between 45% and 60% depending on pupil stage, and reading, at around 40% at each stage, than for science, where rates varied from 10% to just under 35% (Johnson & Munro 2008). Several possible explanations for the particularly poor result in science were offered by the researchers, including a disparity in the nature of the 'science' being assessed by teachers (process) and by tests (knowledge).

In Wales, where formal tests were abandoned for national curriculum assessment in 2005, teachers' level judgements have been routinely gathered year by year at the key stages, and published on DCELL's website. From September 2007 schools have been expected to engage in school-based standardisation and moderation at key stages 2 and 3 "involving suitably robust systems and procedures to ensure that they have appropriate opportunities to discuss their pupils' work and agree a shared understanding of standards", while from September 2008 teachers have been engaged in cluster-group moderation of examples of their pupils' work "to ensure alignment with national standards" (DCELLS 2008).

During 2010 the WJEC managed an external moderation pilot study on behalf of DCELLS. From their pupils' work, school cluster groups selected two core subject profiles at each key stage at levels 4 and 5 (for details see WJEC 2010), which they

moderated and levelled within the cluster. These profiles were produced principally as sources of evidence and as benchmarks for schools. They were also used within the external moderation pilot. External moderators worked in pairs – one primary and one secondary specialist – to review the profiles, the cluster group level judgements and annotations, and on this basis produced reports on whether they thought the cluster judgements appropriate. They external moderators were not required to make independent judgements themselves (Margaret Franks, personal communication, October 2010). To date, therefore, there are no empirical data available for Wales that would offer any direct comment on the current reliability of teacher summative assessment in that country.

2.3 Workplace assessment in vocational qualifications

The unique feature of assessment in the vocational field is that it is heavily, if not always uniquely, workplace based, a characteristic that is assumed to guarantee high validity with respect to the future employment intentions of the qualification candidates. Workplace assessors might be site managers, warehouse supervisors, senior care home assistants, experienced hairdressers, and so on. They train and supervise qualification candidates, and during their training assess their competence against the appropriate set of national occupational standards (NOS) using detailed criterion-referenced assessment schemes. The standards and criteria are statements drawn up by sector skills councils that describe the knowledge and skills required for an individual to carry out a particular job competently (for an example see Harth & van Rijn 2011, p.11). The end-decision of assessment is that the candidate is declared competent or not competent – there is no finer grading.

Typically, workplace assessors observe candidates as they perform the kinds of task that might be required in the course of carrying out the occupation concerned. The assessment might be of a process or of an end-product. The evidence against which competence is evaluated, and the techniques used to produce the evidence, can take a variety of forms, including observations of performance in the workplace, diaries, discussions, projects, assignments, witness statements, and formal tests. Assessors are required to ensure that candidates produce sufficient evidence "to enable reliable and consistent judgements to be made about the achievement of all the learning outcomes against the stated assessment criteria" (Ofqual, 2008, p.26).

Internal verifiers are responsible for ensuring that assessors carry out proper procedures, and apply assessment criteria appropriately. They are expected to do this by observing the assessor making assessments, whenever this might be possible, and by reviewing candidate evidence. External verifiers, who are appointed by the awarding bodies, are responsible for "ensuring that assessment decisions are fair, consistent and meet the requirements set out by the national occupational standards" (Harth & van Rijn 2011, p.14).

But even in a criterion-referenced system, where assessment criteria can be quite specifically stated, it cannot be assumed that assessments are entirely reliable. Yet this is exactly the assumption that was apparently made by the architects of the NVQ system, an assumption that set the tone for the status of the concept of 'reliability' from the start. Paralleling the situation in the new GCSE, assessment validity, or

authenticity, was seen to be paramount, and assessment reliability almost an irrelevance. In the words of Jessup (1991):

> The new model of assessment requires a re-evaluation of the concepts of validity and reliability. It certainly raises questions about the traditional emphasis placed upon reliability of assessments….reliability is important only in so far as it contributes to valid assessment decisions. There are circumstances in which attempts to increase reliability reduce validity. (Jessup 1991, p.50)

Jessup appropriately understood the concept of reliability as consistency in outcomes. He noted that inconsistency might arise should an assessor assess the same competence evidence on two different occasions, or should different assessors independently evaluate the same candidate evidence (Jessup 1991, p.191). And yet he did not see, or did not accept, the implication for NVQs. He had a strong belief in the ability of clear assessment criteria to guarantee 'correct' competence-based assessment decisions on the part of assessors, the inference being that assessors should be able to make consistent, and therefore by default reliable, assessment decisions unless they were themselves incompetent. Or perhaps he was simply reacting against the norm-referenced approach that was, and remains, dominant in the academic qualifications system and the attention which had been given to reliability in that context:

> Reliability has been given great prominence in norm-referenced systems of assessment because, by definition, the assessment is about comparing individuals with each other. Less emphasis is given to what is actually being measured….Typically, in such norm-referenced systems, which have tended to prevail in educational assessment, there is seldom any attempt to relate the assessment to any external criterion, In fact there is often a lack of clarity as to what the objectives of the assessment are, except to discriminate between individuals in some way. (Jessup 1991, p.192)

On the basis of his reasoning, Jessup proposed (p.191) that education professionals should "just forget reliability altogether and concentrate on validity, which is ultimately all that matters". No wonder, then, that reliability has been so little explored in this field, with genuine empirical studies of assessor agreement to this author's knowledge remaining non-existent. Two rare assessor agreement studies have been conducted recently, but while a number of experienced assessors participated in each study the candidate evidence assessed was just one portfolio in one case (Johnson, M. 2008) and two portfolios in the other (Greatorex 2005). If truly valid and generalisable findings are to emerge from similar reliability studies then assessors will need to be required independently to evaluate greater numbers of candidate work samples than this.

Another recent study (Harth & van Rijn 2011), conducted within Ofqual's reliability programme, was larger scale in terms of candidate (portfolio) numbers, and explored agreement rates between workplace assessors and internal verifiers, as opposed to two or more assessors rating independently. For a number of logistic reasons the study was able to investigate just three different qualifications from two different occupational areas – electrotechnical services and hairdressing. Agreement rates were very high – perhaps not too surprising a result, given a likely tendency towards acquiescence among colleagues working side by side in the same organisation with

shared understandings about competence requirements and established social relationships. A study in which internal verifiers reviewed the work of candidates from other centres would be more revealing.

It is undoubtedly challenging to organise meaningful assessor agreement studies where workplace assessment is concerned. But recent work in the health sciences field has shown that this can be done (see Murphy, Bruce, Mercer & Eva 2009 for examples). It is regrettable that generalizability theory (Cronbach, Nanda, Gleser and Rajaratnam 1972; Shavelson & Webb 1991; Brennan 1992, 2001) was so little known in the UK before the new examination and qualification systems were launched, given that this offers an appropriate approach to reliability estimation in competence-based assessment, as in any other assessment field (see Johnson & Johnson 2009 for an overview).

2.4 Internal assessment in GCE and GCSE examinations

The UK's academic school leaving qualifications systems have undergone numerous transformations over time (QCA 2006), partly to address changes in society that rendered existing systems inappropriate and partly in response to an insatiable consumer demand for qualifications. Evolution, if no longer revolution, continues apace.

In the first half of the last century there was the *School certificate,* replaced in the early 1950s by single-subject *General Certificate of Education* (GCE) examinations at ordinary and advanced levels (O and A levels), to be taken at ages 16 (Year 11) and 18 (Year 13) respectively (with the exception of Scotland, which has its own national qualifications). The GCE was designed for the academically most able pupils, who at the time were being educated in grammar and independent schools. The O level, for example, was expected to be taken by the top 20% of the age population by ability, and the A level by an even smaller proportion – those pupils destined for what was then a highly selective university system. The mid-1960s saw the launch of the *Certificate of Secondary Education* (CSE), which was designed to serve the qualification needs of the next 40% of the population at age 16, who were at that time studying in secondary modern schools. In the event, CSEs were in practice taken by a much higher proportion of pupils than originally intended, extending the possibility of a school leaving qualification to almost all 16 year olds. 'GCE boards' and 'CSE boards', which tended to have regional catchments, were respectively responsible for the two different kinds of qualification.

In an attempt to link the new qualification with the old, a decision was made that a grade 1 in the CSE, which was graded 1 to 5, would be equivalent to a grade C in the GCE, whose passing grades were A, B and C. With what degree of validity this linking was satisfied remains an open question. One particularly interesting feature of the CSE was 'Mode 3'. While Mode 1 and Mode 2 examinations were conducted by the examining boards, Mode 3 CSEs were based on syllabuses and examinations that were not only developed by teachers in individual schools or school groups, but were also assessed by those teachers (for further details see Mobley, Emerson, Goddard, Goodwin & Letch 1986).

The conversion from a selective to a comprehensive system of public education during the 1960s led to some timetabling and ethical difficulties for teachers in terms of the qualifications system. Schools were faced with the need to accommodate both GCE and CSE possibilities within their range of course offerings, and one practical way to do this was in effect to continue segregating pupils by ability, at least in some subjects, so that particular pupils would be prepared for one level of examination or the other. This led to a degree of social divisiveness that was almost as blatant as that which had operated when pupils attended entirely different kinds of school. It also meant that many pupils, sometimes for pragmatic reasons such as limited classroom space, were being denied the possibility of aiming as high as their abilities might have justified.

For these and other reasons, in the late 1980s, after several years of deliberation, research and preparation, GCE O level and the CSE were both replaced in England, Wales and Northern Ireland by the General Certificate of Secondary Education (GCSE).

One particularly important feature of the GCSE is that from the start teachers were to assume an increasingly important role as assessors. It was observed that:

> In some ways the GCSE might be described as the 'coming of age' examination for teachers. It acknowledges publicly that teacher assessment is important and respectable in public examinations. (Mobley et al. 1986, p.114)

Coursework was to be included as an assessable component in almost all the new GCSE subject examinations, to increase pupil motivation for courses, which would now be much less constrained by the requirements of external assessment, and at the same time to increase assessment validity:

> The set of subject-specific criteria for GCSE examinations in almost all subjects included a significant element of coursework as well as external examinations. The setting and assessment of the coursework was intended by the government to help teaching and learning processes by measuring and encouraging the development of important skills not easily tested in timed, written examinations, including practical and oral skills and the ability to tackle extended pieces of written work. (QCA 2006, p.2)

Those schools that had previously been organising Mode 3 CSEs would have been quite at home with the mix of coursework and written examinations planned for the new GCSE. For other schools, however, this would have been an interesting innovation. Immediately prior to the launch of the GCSE, Mobley et al (1986) noted that while teachers were in full accord with its aims there were nevertheless anxieties about its implementation:

- Many teachers fear that the courses will be under-resourced and introduced into schools without the necessary supporting structures and facilities.
- In addition, the GCSE will impose, initially, a heavier workload on them as they undertake preparations for the introduction of the new examinations.
- Then there is the role that teachers will have to take on board in the setting and assessment of coursework for external examination purposes (a role which may be new to many teachers). (Mobley et al. 1986, p.20)

In the event the examining boards and local education authorities organised a comprehensive training programme for teachers. The programme included orientation to the new syllabuses and assessment criteria, and advice on how to conduct coursework and keep assessment records. Strategies for offering effective but non-intrusive pupil differentiation were also covered: to accommodate the very wide range of ability: in many subjects written papers were to be 'tiered' (currently 'foundation tier' and 'higher tier') with overlapping grades, and teachers would need to decide which tier would be the most appropriate for which pupils. Crucially, training at this stage did not address the issue of teacher assessment reliability. Indeed, a policy intention was that the issue of reliability would be resolved through the development and application of grade-related criteria, but such criteria ultimately proved elusive.

From the start validity was recognised as "of supreme importance" in the new GCSE, and examining boards were advised to "rethink their attitudes to reliability" (Mobley et al 1986, p.135). As in the other major assessment programme developments that were launched in this period, and that have been discussed earlier in this chapter, assessment reliability was not an issue of great concern in government circles, and indeed it might reasonably be claimed that 'reliability' had for a number of reasons become an uncomfortable and inconvenient concept (see Willmott & Nuttall 1975 for an overview of the findings that emerged from numerous School Council funded reliability research projects, and of the issues raised).

This negative attitude to the issue of reliability persisted over time, and to some extent explains the almost complete absence of relevant empirical research over the past 20 years, with the notable exception of an impressive volume of research into the reliability of written test marking (see Meadows & Billington 2005 for a review of the literature). In her extensive search for quantitative studies on teacher assessment reliability, Harlen (2004), for example, located just one small-scale study that focused on teacher assessment in the GCSE (Good 1988). Yet threats to the reliability of coursework assessment were clearly identified in the early years of the GCSE, with differences noted between teachers and schools in the nature of the coursework itself, the resources available to support it, the amount of parental input to assignments, the assessment techniques and tasks employed and the assessment criteria applied (Scott 1991).

The GCSE system continued virtually unchanged for almost a decade, with increasing anxiety expressed about the impact that the high weighting given to internally assessed coursework in some subjects was perceived to be having on standards in the qualification. One of the highest-entry subjects, GCSE English, was a particularly interesting example. This had a 100% coursework weighting, later reduced to 40% (QCA 2006, p.7). The political response to growing public concern about standards was to introduce in the mid-1990s an upper limit of 20% on the weighting attached to coursework in the GCSE. Subject specification and assessment criteria were revised at this time, and revised again in the early 2000s.

In the QCA's 2006 review of coursework in the GCSE, coursework is described as:

> Any type of assessment activity undertaken by candidates in accordance with the specification during their course of study and that contributes to the final grade awarded for a GCSE qualification. Typically, though, it is an assessment activity that

> is set and marked by a teacher and not carried out under close supervision. (QCA 2006, p.4)

Commenting on the activities that could be included in coursework, such as written work, project work, investigations, production of artefacts, group performance work, oral work, and statistical tasks, the QCA report goes on:

> These different activities present a variety of challenges to the quality assurance of coursework assessment including teachers setting tasks that allow candidates to demonstrate their abilities, ensuring teachers interpret mark schemes correctly, and being confident that the work submitted is the candidate's own. (QCA 2006, p.4)

Concern had been growing about the nature and challenge of coursework and the assessment tasks within it. There were issues to do with teacher workload, and also to do with pupil workload, given that many pupils were engaging in examination-focused coursework in a variety of different subjects. And, not least, there were questions about the validity and reliability of the resulting internal assessments, given the kinds of variation in practice noted from the start in the Scott (1991) study. Validity and reliability were feared to be threatened by inappropriate assessment tasks, misapplication of assessment criteria by teachers, differing standards of marking, pupil cheating (for example by having another pupil or a parent contributing to assignments) and plagiarism (particularly through downloading from the internet).

In response, in the mid-2000s a political decision was made to replace the relatively unconstrained coursework in the GCSE by 'controlled assessment' (QCA 2007, 2009a; QCDA 2009). This would simply mean that while coursework itself might not change in nature or duration the assessments within it would do so: assessment tasks would be provided, or at least approved, by the examining boards, and the process of assessment would be formally supervised by the teacher whenever possible. Where coursework was not considered essential within a subject, i.e. where any relevant knowledge, skills and abilities could be assessed through written tests, then the previous coursework element would be eliminated. Where coursework remained as an examination component this would be weighted at 25% or 60% of the subject examination, as appropriate.

New GCSE subject specifications were developed for first teaching in September 2009, with first examinations in 2010. Since most GCSE subjects still retain a coursework component, in the form of anything between one and three units in a four-unit GCSE, internal assessment is still very much a part of the teacher's role in the secondary school. Moderators, too, are generally practising or retired teachers. So we have two types of summative teacher assessment occurring here. Subject teachers assess their own students' skills and abilities, submitting the results for external moderation. And the external moderators assess the work of samples of other teachers' students. For both internal and external assessors the performance evidence might include oral assessment videos, pieces of writing, sculptures, portfolios, and so on.

The GCE was also undergoing modification during this same period. In response to Curriculum 2000 GCE subject specifications were revised, and qualifications became modularised, or unitised. In addition a new intermediate qualification, the AS (Advanced Subsidiary), was introduced. This was essentially half an A level, and was

originally intended for those candidates who wanted something more than GCSE qualifications but could not for one reason or another reach a full A level. Candidates currently achieve AS levels by taking a subset of the units (AS units) comprising the full A level (AS and A2 units).

In 2010 A level examinations became uniformly 4-unit whereas earlier they could be up to six. Content was updated and where necessary re-packaged to fit four units. Coursework was either eliminated altogether, or made compulsory, so that centres and their candidates would no longer have the possibility of a written paper alternative, or it was moved from AS units to A2 units. And a new A* grade was introduced to offer the best students 'stretch and challenge' and simultaneously to further differentiate the most able students for university entrance and other purposes.

Chapters 3 and 4 focus on the new 'controlled assessment' within the GCSE, as well as the traditional coursework that continues within some GCE subjects, in particular considering the likely impact on examination reliability.

3 Internal assessment in GCSE and GCE examinations

3.1 Coursework and controlled assessment

It has been noted in Chapter 2 that in order to address growing concerns about the impact of relatively flexible coursework on the validity and reliability of teacher assessment (QCA 2006) 'controlled assessment' was introduced into the GCSE from 2010 (though not into the GCE, where traditional coursework elements continue in some subjects).

Controlled assessment is defined as:

> …a new form of internal assessment that replaces coursework in GCSEs. It encourages a more integrated approach to teaching, learning and assessment and enables teachers to confirm that students carried out the work involved. As the name suggests, it applies increased control over assessment of students' work at three critical points:
> - task setting – teachers can choose from a wide range of tasks set by awarding bodies, which can be contextualised to suit local circumstances. Arrangements will differ by subject, with some subjects allowing centres to set tasks
> - task taking – there are several levels (and types) of supervision under which assessment can take place, depending on the skills involved; generally this will be done by subject teachers in regular lesson time
> - task marking – awarding bodies provide mark schemes or criteria.
>
> (QCDA 2009, p.3)

What this means in practice is that coursework, where this is considered relevant in a subject syllabus, will continue much as before, but that any assessments made during the coursework will now be based on tasks set or approved by the examining boards, and wherever possible formally supervised by the teachers managing the course. This is mainly to ensure a) that the work that is assessed by the teacher and/or the awarding body moderator is appropriate, in the sense of enabling the student to demonstrate relevant knowledge, skills and abilities, and b) that it is indeed the student's own work, and not the work of a fellow student, parent or other individual. Both teacher and student are required to sign statements confirming authenticity. These statements are submitted to the awarding body along with centres' mark lists.

Not all GCSE examinations have controlled assessment components, and where they do the weighting is now 25% or 60%: "which group a specific GCSE falls into will depend on the range of skills being assessed and the best way of assessing them"(QCDA 2009, p.4). Within the small group of eight subjects identified for 100% external examination we find psychology, religious studies, mathematics and law. The next group of 12 subjects identified for 25% controlled assessment includes English literature, the sciences, humanities and business studies. The third and largest group of 21 subjects that now comprise 40% external examination and 60% controlled assessment includes art and design, dance, English language, media studies, modern foreign languages and music (QCDA 2009, p.5). In a four-unit GCSE examination up to three units can be focused on controlled assessment, depending on the overall weight given to internal assessment in the subject specification.

Controlled assessment is designed to be embedded in the curriculum, taking place within the normal teaching timetable, for example in the classroom, laboratory or

workshop. It is intended to be used "for assessment of subject-related skills and their application when external assessment is not the best way of assessing them" (QCA 2009b, p.7). In other words, teacher assessment input to GCSE certification is now explicitly focused on promoting maximum assessment validity by ensuring maximum breadth in curriculum coverage within the subject qualification as a whole.

Examples from different subjects of the skills that are expected to be assessed through controlled assessment include (QCA 2009b):

- undertaking research and gathering, selecting and organising materials and information
- planning investigations and/or tasks
- carrying out investigations and/or tasks
- performance and production skills
- working with others and devising creative approaches
- extracting and interpreting information from a range of different sources
- selecting and applying tactics, strategies and compositional ideas
- taking informed and responsible action
- analysis and evaluation of processes and products
- presenting ideas and arguments supported by evidence.

Where controlled assessment features in a subject specification levels of control are identified for each aspect in the process – task setting, task taking and task marking. Levels of control can be high, medium or low/limited, and have been decided subject by subject by the examination and qualification regulators, with the aim of addressing "issues of authenticity, plagiarism, and comparability of process and demand across specifications in the same subject offered by different awarding bodies." (CCEA 2010, p.7).

For task setting, the control level is 'high' when tasks are set by the awarding body and must be used unchanged. The control level is 'medium' when tasks are set by the awarding body but can be adapted by the teacher. Adaptation might be to better tailor the context to the students' interests and experiences, or to better target the task for students of differing abilities, for example by adding more demanding questions in speaking tests for the more able language students, or less demanding ones for the less confident or less gifted students. The control level is 'low/limited' when the centre sets the task following awarding body guidelines and criteria.

In the general case the control level for task setting is high. But among subjects with 60% controlled assessment there are exceptions. Modern foreign languages is a particular case in point. Here teachers are offered several options. They can choose to use one or more tasks from a list of exemplar tasks set by the awarding body, or they can choose to adapt one or more of these tasks as they deem appropriate for their students, or they can create their own tasks bearing in mind the assessment criteria that task performances will be marked against. Where awarding bodies set tasks these will be replaced every year for subjects with 25% controlled assessment and at least every two years for subjects with 60% controlled assessment.

Control levels for task taking essentially dictate the degree of teacher supervision required as students undertake their controlled assessments and related activity. A high level of control means that the tasks must be undertaken under the direct and continuous supervision of a teacher or other centre-nominated individual, in other words formal supervision as in a traditional testing situation. With a medium level of control students can work individually on their tasks, within or outside the classroom, with some guidance from the teacher but without tight supervision. A low level of control implies limited or no direct teacher supervision.

Depending on the level of control defined within the examination specification, the assessment might take place within a supervised classroom, or under less formal conditions within the centre generally, or even outside the centre with limited supervision of research and field work. Indeed, in sympathy with the nature of the subject and in the interests of task authenticity, controlled assessment tasks are not necessarily tasks undertaken wholly by individual students, but can involve a more or less substantial element of group work, provided only that assessable evidence for individual candidate evaluation is produced by that candidate alone. For example, students "may carry out a group activity and write it up individually, drawing out their own contribution to the group activity and commenting on how they developed and demonstrated their own skills" (QCA 2009b, p.8). This is a particularly interesting option, given the difficulties that are known to apply when attempting to make fair assessments of individual contributions to group efforts.

Within any subject, different types of activity within a broad project that includes or ends with the taking of a controlled assessment task can have different levels of control. Thus, research, planning, data collection, practice and preparation are typically subject to limited control, while the development of an artefact, the production of a written evaluation of work undertaken, or completion of an oral, practical or written test will be subject to high control.

Task marking is based on given awarding body criteria, and has two possible levels of control: medium and high. The default is a medium level of control, which means that the work of candidates is marked internally by the centre and externally moderated by the awarding body. In some cases, however, task marking is under high control, which means that the work of candidates is marked externally by the awarding body. Modern foreign languages are a notable exception to the general rule, since writing tasks here are the only writing tasks that are marked externally rather than being internally marked and externally moderated.

Teachers are expected to view controlled assessment as part of their normal course work rather than as a separate time-consuming activity. In other words, like coursework elements in the previous GCSE system, and that continue in GCE, controlled assessment as a process is intended to be heavily curriculum-embedded. In sympathy with this aim, and with the exception only of those subjects with 25% controlled assessment, controlled assessments can in principle take place at any time during a course. Centres do, though, need to bear in mind the terminal assessment regulation, which requires that 40% of the entire examination must be taken in the final session.

3.2 Examples of controlled assessment tasks

The tasks devised by the examining boards or centres for the first new GCSE examinations in 2010 are not yet accessible by the general public. But a number of case studies have been made available by the QCDA that serve to provide a flavour of the likely nature and variety of the tasks that will be used. Here are brief details of a selection of the case studies.

Example 1: GCSE History
(http://www.qcda.gov.uk/resources/508.aspx)

The stimulus task here involved students in an historical enquiry into the "interpretation of an individual", at the end of which they were to reach a substantiated judgement "using sources to justify their line of argument and using these sources within their historical context". In brief, students worked in groups to produce presentations on Al Capone, before working individually on final written reports under controlled conditions.

The task was set by the awarding body but adapted by the school (medium task setting control). The teacher provided several contextualising sessions, after which the students worked in small groups to produce a Moviemaker presentation in answer to an enquiry question focusing on different interpretations of Al Capone. Students recorded their findings in research diaries and shared them with other students and with the teacher via an online blog on the school's virtual learning environment. Together the members of each student group then produced their presentation, having been encouraged to use and reference a variety of different information sources. Through the presentation they were to address the enquiry question, "following a clear line of argument and coming to a substantiated conclusion". Each group presented its work to the rest of the class for peer and teacher assessment. All of this activity would be subject to low/limited control.

A high control assessment task concluded the whole activity. The students worked individually under timed and supervised conditions to produce a written response of up to 2000 words to the enquiry question, including references to their information sources. They had controlled access to their own research notes during this session.

The students' write-ups and research diaries were evaluated by the teacher against awarding body guidelines, and a sample of the work was later moderated by the awarding body (medium task marking control).

Example 2: GCSE Modern Foreign Languages: Speaking
(http://www.qcda.gov.uk/resources/499.aspx)

This task, which culminated in a controlled assessment of speaking, was set in the context of magazine publishing. Students "took part in a simulated job interview for the post of an assistant at a magazine for young people in France". The task, which was set by the awarding body and adapted by the school to suit its students' interests and available resources (i.e. medium task setting control), required students to research the magazine industry, using ICT.

Students began this assignment, which lasted roughly four weeks, by continuing work they had carried out on a previous unit, for which they had investigated French

magazines for young people. They read paper-based and online publications, and completed a questionnaire about their magazine preferences (medium to low control). Students individually analysed the results of the questionnaire enquiry and wrote a report on their findings. They then read a variety of job advertisements in online newspapers and listened to young people talking about various jobs – requirements, advantages and disadvantages – before being given information about a specific job that they were to gather information about by telephone before writing a letter of application. The simulated job enquiry saw the school's French assistant role playing on the end of a telephone, giving information about the job and requiring the student to provide personal details.

The controlled assessment task itself took the form of a 2-3 minute simulated job interview, with the teacher playing the part of the job interviewer (high control). The students were given details about the task three days before undertaking it, including a checklist of "likely points" that they might be asked about, such as their education, their computer skills, their knowledge about young people's magazines, their hobbies, and so on. The teacher discussed the task with the students, introducing useful vocabulary and giving advice on how they might prepare themselves for it. When the assessment day arrived each student was individually interviewed by the role-playing teacher, while the rest of the class worked on regular classroom activities. The teacher adapted the questions to individual students as appropriate, and assessed each student's performance using the awarding body's criteria (medium task marking control).

In this particular case study the teacher did not record the students' interviews. But in a live GCSE examination the interviews would be recorded and securely stored, with a sample being forwarded to the awarding body along with the full mark list for external moderation.

Example 3: GCSE Media Studies
(http://www.qcda.gov.uk/resources/484.aspx)
This task was again set by the awarding body and adapted to suit students' interests and school resources (medium task setting control). Students were to research the medium of radio, plan the delivery of a school radio station, and produce a 10-15 minute extract for broadcast. As they worked they put together a portfolio of written evidence that would eventually be evaluated alongside their controlled assessment task. Task taking controls varied from one part to another of the activity, "but included reliable and valid individual evidence".

The students were given an orientation session by the teacher before starting their individual and group research into public service broadcasting, some of which took place under teacher supervision in class. This culminated in ICT-aided preparation of individual presentations on an aspect of their planned school radio station. The presentation became the first element in the students' personal portfolios. In a second stage the students, with some teacher support, worked individually to produce 20-second soundscapes. These were shared with the whole group and informally peer evaluated before being saved onto a USB memory stick for inclusion in the portfolio alongside a short personal written evaluation produced in class under controlled conditions. A visit to a local radio station followed, in which students were to research the audience demographic in preparation for writing up a case study in class later,

again under controlled conditions but with access to their visit notes. This case study was added to the portfolio. A final pre-production activity had students working individually on proposals for a two-minute item on local radio about a forthcoming school open evening. The teacher provided information about the evening for inclusion in the proposals, supervised initial proposal drafting in class, commented on the drafts, within awarding authority guidelines, and then supervised redrafting in class. The proposals were then added to the students' portfolios.

With the benefit of a teacher-provided brief, and working in groups of four, the students now started work producing a 10-15 minute extract for the proposed school radio, using a range of available technologies. As they worked over the following five weeks they were monitored by their teacher, who took notes but gave minimal feedback. Each group evaluated its own work as it progressed, modifying and adapting as appropriate. Finally, the students peer assessed their group's work and their own contributions to this, in terms of aspects such as reliability, organisation, creativity and effort. This evaluation, along with the teacher's in-class observations, contributed to the teacher's assessment of each project and of each student's individual contribution to it. In a final controlled assessment, with scaffolding support from the teacher in the form of guiding questions, the students were given two hours in which to produce a 1000-word personal overview of their whole project experience, from initial research through planning and production to evaluation (high task taking control).

The task outcomes – portfolio, production and evaluation paper – were marked by the teacher, and the work of a sample of students was externally moderated by the awarding body (medium task marking control).

Example 4: GCSE Design and Technology
(http://www.qcda.gov.uk/resources/521.aspx)

This curriculum-embedded activity focused on designing a digital camera for teenagers. It is not clear whether the task was provided by the relevant awarding body or set by the school and approved by the awarding body. Either way, the research and analysis phase of the task began with individual homework, following awarding body guidelines. The objective for students was to explore camera manufacturers' websites to gather information about the commercial design process. Findings were cut and pasted into electronic research diaries. Under teacher supervision in the classroom the students next explored potential camera design scenarios, developing a mind map for teenagers, which formed the basis for their product brief. They then once more independently, though under teacher supervision, consulted camera manufacturers' websites, to explore existing markets before analysing an actual mobile phone/camera and developing their observational design drawing skills and techniques. The students then considered the technological development of cameras and image capture, using a teacher-supplied collection of cameras of different types. They shared their analytic findings with the rest of the class, and recorded their conclusions in their electronic research diaries.

In the project development phase the students began by recording in appropriate ways the essential attributes of their intended camera in a worksheet prepared by the teacher following awarding body guidelines. Still under teacher supervision, the students were encouraged to review examples of previous students' work, as supplied by the

teacher. Following this they took their ideas further, individually and in group discussion, before producing drawings and a sequence of increasingly detailed 3D models. The models were produced under teacher supervision, and kept in school under controlled conditions (i.e. securely). The students then used CAD to generate a working drawing of their designed camera, and CAD/CAM to generate a foam model. Throughout this development phase students were adding notes to their research diaries. Finally, the students used their CAD rendered views to produce a promotional poster for their camera, and staged a presentation to the rest of the class, during which the teacher monitored and recorded activity.

For each student the project resulted in four assessable outcomes: research diary, worksheet (containing information about their new camera attributes), a CAD working drawing and a foam model. The teacher used these outputs, presumably along with observations made throughout the process, to arrive at an assessment of each student's performance. The work of a sample of the students was later moderated by the awarding body.

Other examples

Further example controlled assessment tasks are to be found on QCDA's website and also on the websites of some of the examining boards. AQA, for example, offers specimen tasks in Drama (AQA 2008), English Literature (AQA 2009a) and Modern Foreign Languages (AQA 2009b), among others.

Edexcel offers a complete set of specimen units for its GCSE in French (Edexcel 2008), including traditional mark schemes for the reading and listening units and 'best fit' rubrics for the speaking and writing components. The relatively short 40-mark reading and listening tests – foundation and higher tiers – offer candidates no question choice. In contrast, the speaking and writing units offer choice of theme, from among the four themes in the examination specification: media and culture, sport and leisure, travel and tourism, business, work and employment. There is also a choice of task within and across themes. In addition, there is scope for teacher modification to, or substitution of, stimulus materials (photos and drawings) and bulleted questions (to guide speaking assessments) or information prompts (things to include in writing tasks). Task setting is therefore subject to limited control. This enables teachers, for example, to change the given context to better suit their students' interests and experiences, or to modify the number and wording of bulleted questions and prompts to further challenge the more able students and to better motivate others. Edexcel does, however, add this caveat:

> Clearly, the facility to modify task stimuli enables teachers to target activities to the level and needs of individual students. However, as all changes to tasks can impact significantly on their overall level of demand, it is imperative that teachers do not constrain or compromise the performance of their students through an inappropriate task stimulus. (Edexcel 2008, p.111)

Students were individually to undertake two different speaking tasks with their teacher (task taking being therefore subject to high control), from a choice of three types: presentation with discussion following, picture-based free flowing discussion, role-play open interaction. Each task was to last 4-6 minutes, with student performance assessed by the teacher using the awarding body's assessment criteria

(see the next section for details). The teacher's mark allocations for a sample of students would be moderated later by the examining board on the basis of the two recordings. Task marking was therefore subject to medium control, with each task meriting 30 marks.

Interestingly, assessment workload issues have already arisen in the context of speaking assessment, and centres were advised before the first 2010 examinations that they would in fact be required to submit a recording of one only of the two tasks for moderation (JCQ 2009). Although this will have been an essential compromise in the circumstances, reducing two task recordings to one only will clearly make the moderators' own marking responsibility that much more difficult to carry out effectively.

In the case of writing, students were each required to undertake two tasks from a choice of three tasks set by the awarding body within each of the four themes (high task setting control), under teacher supervision (high task taking control). Each task was worth 30 marks. But in contrast with all other subjects, controlled assessment writing tasks in modern foreign language GCSEs are marked externally by all the awarding bodies. There would therefore be no internal teacher assessment in this case (so high task marking control).

3.3 Assessment evidence and assessment criteria

Every subject specification includes a set of learning objectives that underpin the design of courses, of summative written tests, and of controlled assessment tasks. The evidence upon which judgements about the degree to which the learning objectives have been achieved by candidates will include completed written tests, some of which will have originated as controlled assessment tasks, electronically recorded role-play interactions with teachers, research diaries, artefacts such as models, sculptures and paintings, and portfolios. There will also be ephemeral evidence in some cases, such as drama productions and musical performances.

The way that the different types of evidence are assessed will clearly be different for the different kinds of evidence, and the effectiveness of moderation will vary too. Where the outcome of a controlled assessment unit is a completed written test then the moderator's marking responsibility will be little different from that of a regular marker working on external written test units. But where portfolios and artefacts are involved, the marking task will be much more challenging.

The assessment criteria that teachers and moderators are required to use when marking candidates' work are, like the learning objectives, included in the relevant subject specification. For example, for the GCSE French examination, for which specimen units were briefly described in the previous section, the reading and listening tests (externally marked) are composed of a series of objective items, and have traditional readily applied mark schemes, with most test questions binary-scored. The assessment schemes for speaking and writing, however, are inevitably different in kind.

The speaking tasks are rated for 'content and response' (18 marks), 'range of language' (6 marks) and 'accuracy' (6 marks). The writing tasks are rated for

'communication and content' (15 marks), 'knowledge and application of language' (10 marks) and 'accuracy' (5 marks). Each aspect is assessed using a 'best fit' scheme. For 'knowledge and application of language', for example, the mark allocation runs from 0 marks for "no language worthy of credit" to 9-10 marks for the combination of a "wide range of vocabulary and structures, fully appropriate to the task and used effectively", "little or no repetition", "confident use of more complex structures, such as object pronouns, negatives, superlatives and range of tenses, with very few lapses" and "clear ability to manipulate language and to produce longer, fluent sentences with ease" (Edexcel 2008, p.118).

A feature of best fit scheme application, or levels based marking, is that while it is usually relatively easy to identify candidate performances that merit the highest marks, and those where no marks at all are merited, distinguishing performances between these extremes is not so straightforward. Take for example, the 'middle range' performance descriptors in the writing assessment scheme for 'knowledge and application of language', shown in Table 3.1.

Table 3.1: Extract from a best fit assessment scheme for French writing

Knowledge and application of language	*mark*
• Quite a wide range of vocabulary and structures appropriate to describe and to express and justify opinions. • Some attempt to use ambitious structures (subordinate clauses, object pronouns, tenses, etc) with a fair measure of success. • Tenses are generally used correctly. • Some ability to manipulate language although not always successful.	7-8
• Vocabulary and structures are generally appropriate to the task. • Correct syntax when using simple, short sentences. • Some longer sentences where syntax is not always correct. • Attempts enhancement of fact with adjectives and adverbial phrases with some success. • Some evidence of correct use of a range of tenses, with some lapses. • Attempts to use subordinate clauses/simple linking with some success.	5-6
• Limited vocabulary and structures, often repetitive and stereotyped. • Language is basic and sometimes inappropriate to the task. • Pre-learnt, set phrases predominate but there are some short simple sentences, which are more or less correct. • Some attempts at tenses, but many mistakes. • Some attempt to use adjectives. • There may be some simple subordination.	3-4

Source: Edxecel 2008, p.118

For each mark range a number of different characteristics of written language are identified. What would a teacher or moderator do should a candidate demonstrate achievement of a subset of the features but not all? Would some features be given higher conscious or unconscious weighting than others by the rater in such cases? In what circumstances would seven marks be more appropriate than eight marks? And how easy can it be in practice to clearly distinguish between candidate performances worth 7-8 marks from those worth 5-6 marks?

Evaluating writing using such a rating scheme will be quite challenging. The challenge will surely be greater for the evaluation of speaking skills, on the basis of a relatively contrived teacher-student interaction, especially for the moderator who will be working from single video recordings. Applying such schemes in art and design, or music, will be more challenging still.

Even among the most experienced teachers there will inevitably be variation in judgements in the middle of such a rating scale. The question is how large or how small are these inevitable differences in practice? We can only begin to have an idea about assessment reliability if we can answer such questions empirically, and at this point in time such empirical studies appear to be lacking.

4 Quality assurance for GCSE and GCE qualifications

4.1 Internal standardisation

Where centres have more than one teacher involved in the assessment of coursework or of controlled assessment tasks for a particular subject specification then they are expected by the awarding bodies to undertake some form of internal standardisation. The same applies to consortia, where it is expected that standardisation activities will involve teacher assessors from all the centres constituting the consortium. Interestingly, there is no expectation that centres with a single teacher assessor for a particular subject will be involved in inter-centre standardisation.

Several approaches to internal standardisation have been suggested. One of these is consideration of exemplars that would then act as benchmarks. Another is consideration of a small number of pieces of work from the top, middle and bottom of the range from each teaching group, these pieces of work to be marked independently by the teacher participants before being jointly discussed. Independent marking would highlight any tendency for some teachers to rate work more or less highly than others, and in this way standardisation might be reached.

Notwithstanding the expectations that centres carry out internal standardisation there seems to be little if any evidence that internal standardisation is indeed practiced by centres, and, if so, what proportion of centres engage in it, with what effect, in the various subjects examined at GCSE and GCE levels. In the 2006 QCA evaluation of coursework it was noted that few centres were indeed practising any form of internal standardisation:

> Standardisation within a centre is required and there is much good and often very thorough practice taking place. However, internal standardisation is not apparent or consistent across all centres. Awarding bodies need to carry out further checks and provide better guidance. (QCA 2006, p.11)

The awarding bodies offer training and standardisation meetings for centre participants. These meetings, however, are usually brief – one half day to one day at best – and much of the time is spent in administrative matters, task orientation and an overview of assessment criteria. It is unlikely that within the short time allocations any independent evaluation of candidates' work is undertaken, results compared and the process repeated until a useable and demonstrable degree of agreement is achieved. Whether and how centres carry out such agreement checks is unknown – this author could not locate any relevant documentation.

4.2 External moderation

Moderators mark, or in other appropriate ways evaluate, samples of candidates' work in each subject from each centre. On the basis of this review a centre's marks for the unit concerned might be accepted without change, or the marks might be accepted after some statistical adjustment, or a total remark might be requested. So the first question has to be: how are the work samples selected?

Work sampling

The way that work samples are selected for external moderation varies from one awarding body to another. Some boards draw random samples of candidates from each centre's subject entry list, or have centres themselves do this using a given sampling scheme, before centres mark their candidates' work. Others draw random or judgemental samples after centres have marked their candidates' work, using the centres' submitted mark lists for this purpose. In every case the samples are designed to represent a spread of candidate achievement within the centre. Should the sample drawn ahead of marking turn out not to contain the highest and lowest performing candidates then the centre is asked to include those candidates as additions.

Sample sizes are a minimum of 10 candidates for centres with more than 10 candidates in total, but can be up to 20 for large-entry centres. Centres with 10 or fewer candidates submit work for all their candidates.

Some of the boards draw a subsample from within each centre's sample, and use this as the basis of a first decision point about the quality of a centre's marking. Subsamples can be as small as five candidates for centres in which up to 10 candidates were entered for the unit, and six for centres with larger entry sizes than this. Or they can vary markedly in size, every other candidate in the centre's full sample being included in the subsample. Table 2.1 provides details of the sample and subsample sizes that all boards have agreed to use in the future (Taylor 2009).

Table 2.1: Moderation sample and subsample sizes for centres

Total number of candidates	*Sample size*	*Subsample size*
1-5	All	All
6-10	All	5
11 - 100	10	6
101 - 200	15	8
Over 200	20	11

The boards' different sampling strategies have different advantages and disadvantages in terms of predictable and experienced impact on the time required for the moderation process. Where samples are identified ahead of centre marking then the moderation process will be shorter than it might otherwise be. This is because centres would know before they sent in their mark lists which candidates they would also need to send assessment evidence for (products, portfolios, writing scripts, and so on), and so could package that up immediately for despatch to the moderator. Where samples are identified on the basis of submitted mark lists then candidates' work is requested in a second stage, with a consequent delay between receipt of mark lists and receipt of evidence for moderation (it is anticipated that future automation of mark submissions and development of appropriate sample selection software will alleviate this problem somewhat).

All the examining boards reserve the right to request further samples of candidates' work should initial moderation raise issues, and to call in the work of all candidates where significant problems with the initial marking are perceived by moderators.

There is one difference in moderation practice that might usefully be mentioned here, and this is between those subjects for which physical evidence can be submitted for moderation, such as completed written tests, recordings, portfolios, and so on, and those for which such evidence is not available because the end-product is ephemeral. Where the outcome of a course is, or includes, a drama production, for example, or a musical performance, the moderator visits the centre as the teacher's assessments are being made. The moderation sample in such cases will be opportunist, since some candidates might have been assessed before the moderator's visit.

Arriving at decisions about the quality of centre marking

Whichever method is used for drawing work samples, the next question of interest must be: what do moderators do with the submitted assessment evidence in order for the awarding body to make a decision about the adequacy or otherwise of centre/consortium marking? Once again, the answer, at least in part, varies from awarding body to awarding body.

In all boards and across all subjects each centre or consortium is assigned to a single moderator, who reviews and marks the submitted work samples. The moderator has the teacher's original marks available, along with any annotations the teacher might have made on the work samples, such as justifications for mark allocations. Depending on the results of the moderator marking three outcomes are possible:

1. the centre's marks are accepted unchanged
2. the centre's marks are adjusted
   (to address differences in marking standards between the centre and the moderator overall or for parts of the mark range)
3. the work of all the centre's candidates is remarked
   (the centre's marking standards appear too inconsistent for a rational adjustment to be identified).

All awarding bodies apply the same tolerance to the results of comparisons of teacher and moderator marking in all subjects, in order to make a first decision about acceptance or not of the teachers' marks without adjustment. The tolerance is 6%, i.e. where the difference between a teacher mark and a moderator mark is within 6% of the unit total mark (rounded up) then the teacher's mark is considered to be within tolerance. For an internally assessed unit with a maximum mark under 17 the tolerance would be one mark, for a unit with a maximum mark anywhere between 17 and 33 the tolerance would be 2 marks, and so on. The tolerance criterion was apparently originally based on awarding body research several years ago for application in coursework moderation for GCSE, GCE and vocational counterparts – see, for instance, the notes on common principles and practice for moderation of centre-assessed GCE and GCSE unit and components issued by the Joint Council on Qualifications in the early 2000s (JCQ 2003a, 2003b).

Once a centre's sample or subsample has been marked by the moderator the results are reviewed by board professionals. In those cases where boards begin by reviewing a subsample then should all the differences between centre mark and moderator mark be within tolerance the centre's marks for *all* candidates from that centre are accepted unchanged. Otherwise the moderator moves on to mark the work of the remaining candidates in the centre's full sample. When both teacher and moderator marks are

available for all sample candidates boards review the pattern to decide between the three options, *viz*. accept marks, modify marks, request a complete remark.

Two boards subject the two sets of marks to a linear regression analysis at this point, to produce a line of best fit, from which regression-based marks for all the centre's candidates can be produced, whether the candidates formed part of the centre work sample or not. They also visually review the regression plots and use a combination of statistical and judgemental criteria to decide whether to accept the centre's marks unchanged, whether to replace them with the regressed marks, or whether to request a complete remark. Other boards review the evidence of centre and moderator mark differences, without a regression analysis, to make the same decisions. Without the possibility of regressed marks the centre's marks would be linearly adjusted up or down by an appropriate amount to compensate for any perceived difference in marking standards between centre and moderator, across the mark scale or for different mark ranges.

Where a picture of clear inconsistency in centre marking is perceived in the data, or is signalled by the moderator to the board, then a complete remark of candidates' work will be requested. Inconsistency essentially means that the rank order of candidates based on internal marks is noticeably different from that based on moderator marks, possibly, though not necessarily, because the centre had not applied the marking criteria as intended. In other words there will be evidence that the general marking standards of teacher(s) and moderator are not simply different overall, but that one or other's standard was applied unequally to candidates or that different criteria were being applied by the two individuals. The assumption is that the moderator is consistent in the application of appropriate assessment criteria and standards and the teacher(s) not. In such cases a uniform increase or decrease in candidates' marks would be less fair for some of the candidates than for others.

When a remark is considered necessary, some boards require the centre to undertake the remarking, which is then further moderated, while others have the remarking carried out directly by the moderator.

GCSE moderation experience in recent years is that a large majority of centres/consortia have their marks accepted unchanged, and typically between 10% and 20% have their marks linearly adjusted, the proportion being higher or lower in some subjects and years than others. In a small proportion of cases, around 1%, the internal marks are considered to be too inconsistent for a linear scaling adjustment to be justified, and a complete remark is undertaken by the centre or the moderator.

Again, a comment on moderation practice in subjects producing ephemeral evidence of achievement is in order. Apart from the sampling difference noted earlier, another very important difference here is that the moderator and teacher rate candidates simultaneously. Their two judgements on each candidate are therefore independently made, unlike the situation described above where moderators have the benefit of seeing the marks already allocated by the candidates' teachers. The moderator's standards of judgement are as usual considered the correct ones, so that where differences in views occur the moderator decides what action should be taken. If action is needed then this is retrospectively applied to candidates who might have been marked before the moderator's visit.

But can a single moderator 'carry' assessment standards to the extent that that individual is necessarily more correct when making judgements than the candidates' original teacher? "Where there is disagreement between teacher and moderator is the moderator necessarily right?", asked Taylor (1992, p.12), who designed and implemented a rare empirical study to explore the question. In the study, for each of four centre-assessed GCSE/GCE coursework components, three experienced moderators independently evaluated the work of the same candidates. There were 80 candidates for each of three GCSE subjects – English, mathematics and history – and 60 for GCE A level psychology. The research confirmed that indeed a single moderator cannot be assumed to carry standards:

> Even where there was no significant difference between the means for two moderators, it was possible that those moderators nevertheless disagreed substantially on the marks awarded to individual candidates. (Taylor 1992, p.5)

Because of the differences in moderators' judgements, many candidates would have received a different grade depending which moderator had marked their work: for example between roughly 15% and 25% of the GCSE English candidates whose work was reviewed would have had a different grade had their centre mark been replaced by the mark of one or other of the moderators (Taylor 1992, p.6). In GCSE history the proportion varied between approximately 20% and 40% (p.8), in GCSE mathematics between roughly 15% and 30% (p.9), and in A level psychology between 10% and 20% (p.10). This was an important piece of work which has not, to the author's knowledge, been repeated.

An investigation into examination-based grade comparability that was undertaken in pre-GCSE days produced a similar finding about experienced examiners being able or not to carry standards. In each of three different subject studies, three senior examiners from each of three different examining boards independently judged each script in random samples of scripts from those same three boards. There was as much variability in the grade judgements of examiners from within any one board as there was between examiners in different boards (Johnson & Cohen 1984; Johnson 1988). In other words, a prevailing assumption that senior examiners could 'carry' their boards' standards was shown not to be valid, just as Taylor's study provided evidence that single moderators, however senior, cannot do so either. And yet this assumption continues to hold, both for examiners and moderators, despite the empirical evidence.

In light of the Taylor (1992) research, the fact that just one single moderator marks the work of a sample of candidates to service decisions about the quality of the centre's marking in a qualification unit is concerning. The fact that comparisons between centre and moderator marks are being based on such relatively small work samples is also something that merits attention, as previously noted by Wilmut (2005), who advised that sample sizes be increased.

4.3 Grade awarding and uniform mark conversion

Grade awarding is a complex process in the heavily, though not exclusively, norm-referenced GCSE/GCE system. It involves principal examiners or, in the case of centre-assessed units, principal moderators agreeing two critical grade boundaries on the basis of a number of different kinds of qualitative and quantitative information.

For the GCE, the responsible examiners and moderators must decide the appropriate mark that would divide candidates receiving grade A awards from those receiving grade B awards, and the appropriate mark that would separate those candidates receiving grade E awards from those left ungraded. For GCSE the boundary marks that are judgementally determined are for the A/B and C/D boundaries for higher tier unit papers and the C/D and G/fail boundaries for foundation tier papers (see Wheadon & Béguin 2010 for further detail on the grading of tiered papers).

Principal examiners, and principal moderators, carry a heavy burden of responsibility when they make judgements about appropriate mark boundaries for key grades. And their judgements are critical.

The evidence that examiners use to support the decision-making process includes: the current unit paper/task itself plus its mark scheme, along with those for previous series; examiners' feedback reports on candidates' performances, also compared with previous years; any published grade or performance descriptions; archived scripts at or close to the relevant boundary marks for previous papers/tasks plus current script samples at or near possible boundary marks for the paper in question; mark distributions for previous papers and the current paper; grade distributions from previous papers, and also for the current paper on the basis of potential boundary marks; a comparison of entry patterns from previous years to the current one; and centres' estimated candidate grades.

Note that much of the comparative evidence that is supplied to examiners for their grade boundary determinations derives from past papers of similar kind. Where subject specifications or paper structures change then this supporting evidence is no longer available and the task of boundary mark determination is more difficult:

> In considering the Principal Examiner's proposed range, it must be recognised that Principal Examiners vary in their experience, their own internalised standards and ability to focus on sound boundary marks without the provision of mark distributions. When the standard is new or the assessment structure has changed the difficulties in recommending sound boundary marks should not be under-estimated. (JCQ 2006, p.2)

Once the key judgemental boundaries are agreed the remaining boundaries are automatically determined, simply by dividing the mark range between the two critical boundary marks so that there is an even range of marks for each of the intermediate grades. Depending on the shape of the underlying mark distribution this practice can, and frequently has, resulted in intermediate grades being separated by just two or three marks (see, for instance, Johnson & Johnson 2011, reporting Ofqual-funded research), a feature noted over 30 years ago on the evidence of Schools Council funded research into CSE, O-level and A-level examinations (see Willmott & Nuttall 1975).

In the past this procedure was carried out for entire examinations, using the overall marks gained on the examination, which would have been a weighted sum of the marks achieved on the examination's various component written papers and coursework (see JCQ 2006 and Robinson 2007 for full accounts). Now that GCSE and GCE examinations are unitised every unit is separately graded, using the same

procedure. But this is not as simple as it sounds. In fact, the grading task has become very much more difficult.

The greater degree of flexibility that examination candidates currently benefit from, in terms of different specifications in the same subject, unitisation, timing of assessment, and resit possibilities, has consequences for the validity of traditional norm-referenced approaches to grading. In particular, fewer candidates will be taking any particular unit paper at any given time than might have been the case in the past. And the ability characteristics of these candidate groups might vary over time more markedly than for a previous whole-examination candidature. Norm referencing becomes of questionable validity as a result.

This must be particularly true for units that are based entirely on coursework assessment, especially where this involves the evaluation of artefacts, recordings, portfolios, and so on. Such end-products are already known to be difficult to evaluate objectively. This difficulty is compounded by the fact that for logistical reasons a mere handful of candidate work samples are made available for review during the grade awarding process. How does this affect the result? How might grading outcomes vary depending on the particular small work samples that are reviewed? And what is the situation for those units whose internally assessed coursework culminates in ephemeral evidence that by its nature cannot be taken into account by those with responsibility for determining grade boundaries?

Putting aside the new challenges for dependable grade boundary determination for single examination units, a way had to be found to try to ensure that candidates' results from one unit to another would be as comparable as possible, whenever those units might have been taken and whatever their relative difficulty might have been. This would be fair to candidates taking the same unit at different points in time, since the controlled assessment tasks or written papers comprising any unit would be different on each occasion, and their general difficulty would probably be different also.

The same raw mark for the same unit (but different question paper) taken on different occasions would not necessarily result in the same grade award for the candidates achieving that mark. An A/B boundary mark might be 75 on one occasion and perhaps 78 on another, for the same unit, depending on the decisions of the boundary setters. Clearly this issue had to be addressed in the interests of fairness to candidates. It was also essential to provide a fair basis for combining achievement results across different units within a qualification to arrive at overall grades for that qualification for individual candidates. And so the uniform mark scale (UMS) was conceived (see AQA 2010 for a comprehensive overview).

'Uniform marks', on a 0-100 scale, map to grades as follows:

A: 80-100
B: 70-79
C: 60-69
D: 50-59
E: 40-49

Once uniform marks are available for the different units comprising an examination these are added after appropriate weighting to produce a uniform mark for the whole-examination mark, and this automatically determines the appropriate grade award. The maximum number of uniform marks available for a unitised qualification is a multiple of 100, the multiplier being the number of units concerned. Thus, a four-unit A level or GCSE has a maximum of 400 uniform marks, while a two-unit AS and a 2-unit GCSE each has a maximum of 200 uniform marks.

Multiplying up the mark-grade correspondence above, a candidate would need a combined UMS of at least 160 marks to achieve an A grade across a 2-unit AS level qualification or a GCSE, and at least 100 marks for a D grade. And in a 4-unit GCSE or A level examination a candidate would need at least 320 uniform marks for a Grade A, 240 for a C and 160 for an E. To achieve the newly introduced A* grade at A level, candidates need to achieve an A grade overall on their A level, all four units combined, *and* to achieve an average of 90% or more of the maximum uniform marks on their A2 units, thus satisfying the 'stretch and challenge' element. It should be noted that in 2010, the year of A* launch, examining boards might be required by their regulators to adjust the agreed general criterion for A* awards should the proportion of candidates in large entry subjects potentially eligible for A* awards exceed a 2% tolerance on proportions calculated on the basis of 2009 examination results.

But how are raw marks transformed into uniform scale marks? The answer is simple for a single unit. The 'raw' boundary marks that emerge for the unit in the grade awarding process are mapped directly onto the boundary marks of the UMS. Thus whatever raw mark the examiners and moderators decide is the appropriate one to distinguish candidates deserving an A grade from the rest, this mark becomes 80 on the UMS. And so on. For raw marks which fall between boundary marks the raw mark is mapped proportionately to the UMS. Thus, if a raw mark is a quarter of the way between the C/D boundary and the B/C boundary then the UMS for that candidate will be a quarter of the way between the minimum UMS marks for grades B and C. These UMS marks are 70 and 60, respectively, so that the appropriate UMS mark for the candidate would presumably be 62 (61.5 rounded up).

Combining UMS marks across units to determine appropriate candidate grades for a qualification is more complex, because unit weightings must now be taken into account. While a four-unit GCSE will have a maximum uniform mark of 400 this does not mean that every unit within that qualification will carry 100 uniform marks. The distribution of the 400 marks over the four units will depend on the weighting given to each unit in the qualification. Thus a unit with a 30% weighting in the total qualification will have a maximum uniform mark of 120 (30% of 400), a unit with a 20% weighting will have a uniform mark allocation of 80, and so on. The uniform marks achieved by candidates on the various units must clearly be suitably weighted before summing to produce the overall uniform mark for the qualification as a whole.

5 Implications for reliability investigation

5.1 Coursework, tasks and conditions of assessment

There are a variety of factors that can affect the validity and the reliability of any assessment of educational performance. These include the nature of the test and tasks that are devised to enable students to show evidence of their knowledge, abilities and skills. A text-heavy test of numeracy might be a valid test of numeracy skills for able readers but less so for poor readers. An investigation task that is designed to allow a student to demonstrate research and analysis skills will be more or less valid as an assessment of those skills should much of the investigation take the form of group work. A portfolio-based assessment might in principle be a valid format for assessing art work, but would be less valid should different students' portfolios contain such a different variety of content that they show differential evidence of the skills and abilities that together form the basis for the art qualification.

The introduction of a degree of structure and control into coursework assessment, in the form both of tasks, task taking conditions and assessment criteria, must have a positive impact on reliability, without unduly jeopardising validity. However, it should be clear from the few example controlled assessment tasks described in Chapter 3 that there remains very wide variety in the nature of the tasks that teachers are required or free to use for GCSE internal assessment, and in the nature and extent of supporting coursework. Teachers have advanced knowledge of the assessment tasks, even when these are set by the awarding bodies. They also often have some choice of theme and topic, to allow them to plan their coursework in advance and to focus this on their students' local circumstances and abilities and their own and their students' interests. How tempting must it be for teachers to "teach to the task", a criticism frequently identified as a drawback of tests? And how strong would be the temptation to offer a little extra help in one form or another to a weaker student – especially given that the student's chances of gaining the qualification would in part depend on the outcome of this particular assessment (for an interesting discussion on this and related issues see Stanley et al 2009).

Students, for their part, also have prior knowledge of the tasks that they are to undertake, and can even legitimately benefit from teacher support in the form of prompts beforehand. Even when this prior knowledge is short, a few days perhaps, this does mean that undertaking a controlled assessment task is different from taking an unseen external test. What are the implications for assessment validity? And if the focus of the coursework had been different, or had the teacher or student made a different choice of controlled assessment task, what would be the implications for assessment reliability?

We know from a large body of research (see, for example, Meadows & Billington 2005 and QCA 2009a) that different markers can produce different marking outcomes for the same piece of work, even when relatively tight mark schemes are applied. We know less about the impact on assessment reliability of changes in tests and tasks, which would surely be greater in some subjects than others. We do now have some information about such effects for written GCE and GCSE papers (Johnson & Johnson 2011), as an important outcome of Ofqual's reliability programme. And studies in the US have demonstrated that in performance assessment, specifically the

assessment of science investigation skills, the tasks used have a greater influence on candidate outcomes than the raters who rate their performances (e.g. Shavelson, Baxter & Pine 1992). The same phenomenon has emerged during the exploration of the reliability of performance assessment in the medical and health sciences (Murphy et al 2009). Yet for assessed coursework in the GCSE and GCE, with its high degree of individual teacher and student choice of topic and task, there appears to be little if any empirical evidence about the likely impact of task choice on reliability.

Factors affecting assessment reliability also include the conditions under which the assessment takes place, and the way that the outcomes of the assessment are recorded for evaluation. When a teacher interacts with a student in a one-to-one controlled assessment of oral skills, does it matter whether this interaction occurs in a doorway, so that the teacher can simultaneously keep an eye on the rest of the class, or in a quiet anteroom? What difference would it have made? What difference would it have made had a different teacher, perhaps an external moderator, replaced the students' own class teacher for this particular assessment? Would the student have felt more or less relaxed? Would the task have seemed more or less contrived? Would the outcome have been different? We have no evidence one way or the other at the present time. Moreover, knowledge about such issues is lacking elsewhere, too. Even in countries and states that have long-established systems of high-stakes teacher assessment, with or without parallel testing operations, little is known about exactly how teachers arrive at their judgements about student attainment, nor about the reliability of the different kinds of task that teachers use with students to provide relevant evidence (MacCann & Stanley 2010 note this in the case of Australia).

5.2 Internal assessment: performance evidence and rating criteria

The next issue concerning assessment reliability resides in the internal centre marking process. The judgements that are required to be made in some GCSE and GCE subjects when evaluating coursework, even when this involves controlled assessment tasks, are often complex. And yet little if any research has been carried out to investigate the 'state of health' of summative teacher assessment at this level. When there is no tangible outcome to evaluate then assessment reliability becomes an entirely elusive concept with no possibility of empirical investigation. With an end-product available – a sculpture, piece of writing, video clip, portfolio – reliability investigation does become possible. But exploring assessment reliability, even if in principle an essential activity, will be a useful activity only where the end-product is agreed by all to be a fair (valid) representation of the candidate's achievement in the subject domain.

As we have seen in Chapter 3 the tangible results of a controlled assessment in the GCSE, paralleled in project work in the GCE, can take many forms, and can vary in quantity and variety, from a single short piece of writing or a video-recorded oral interaction, to any combination of research diary, portfolio, creative production, evaluation paper and working model. For some of these individual forms it is possible to produce rating schemes of one sort or another, traditional marking schemes, best-fit schemes or schemes that combine best-fit with mark allocation. This provides the strongest basis for exploring rater reliability. Rather little reliability research has been carried out, unfortunately, although there have been a few studies that have explored

the reliability of writing evaluation and of the even more challenging portfolio assessment.

The assessment of writing skills has been explored in the context of national curriculum assessment in England at key stage 3 (Baker et al 2008), and also in the context of the Scottish Survey of Achievement (SSA) in Scotland for pupils aged 10 to 14 (Johnson 2009). In the first study, five different markers independently rated each of 40 scripts in each of several script batches, producing average inter-rater correlations of around 0.85. In the second study, 20 lower primary teachers rated 48 scripts produced by 8 and 10 year olds, while 25 upper primary/lower secondary teachers independently rated the same number of scripts produced by 12 and 14 year olds: rating involved the production of level judgements using national writing criteria. Generalizability analysis revealed that rater-script interaction was a larger contributor to measurement error than inter-rater differences in standards. Triple rating of scripts within the survey itself (8000 scripts in total across the age groups) produced reliability coefficients for 'absolute measurement' of between 0.85 and 0.90. Double rating would have produced coefficients of under 0.8 and single rating coefficients of around 0.7, replicating a similar finding reported by Falk, Ort & Moirs (2007) in a different context.

The reliability of portfolio assessment, too, has received attention, with mixed results (see Elton & Johnston 2002, Harlen 2004 and Stanley et al 2009 for reviews). The fuller and the more varied the content of a portfolio the more difficult it will inevitably be to evaluate the content as a whole to produce a mark, level, grade or even simply a competency decision. Because portfolio evaluation is a very time-consuming process, in empirical studies in vocational education that have been carried out in England to explore inter-rater agreement rates extremely few portfolios have typically been reviewed. As a result no generalisable quantification of rating reliability has been possible (see, for example, Greatorex 2005; Johnson, M. 2008; Murphy et al 1995).

The fact that in 2006 QCA observed that internal standardisation in schools for GCSE assessment was not widespread must be of concern in this sense. It is clearly vital that if teachers are to undertake internal assessments that contribute to high-stakes examinations then this internal pupil assessment needs to be as dependable as possible. Relying on external moderation to address problems, even where problems can be assumed with confidence to be properly exposed through this strategy, is 'end of line' quality control and not ongoing quality assurance. There is a generally acknowledged need to address the issue more directly, by (re)developing the assessment skills of classroom teachers, and building their confidence for making assessment judgements of their own pupils.

This need was confirmed in a recent study that focused on an in-depth investigation of the nature and quality of summative assessment practice amongst lower secondary English and mathematics teachers in three schools in England (Black, Harrison, Hodgen, Marshall and Serret 2010). The external testing regime – national curriculum assessment at key stage 3 – was found to have had a profound impact on the teachers who participated in the study, to the extent that they no longer had confidence in their own abilities to judge their pupils' subject achievement. The researchers observed that:

> Our initial exploration of the existing practices within the three schools showed that these were weak in the application of the principles that guide quality in assessment.

and after their initial research concluded that:

> Overall, the application of concepts of validity and reliability, to the extent of taking these seriously in auditing their own work, was seriously in need of development amongst the teachers.

Support for teachers engaged in internal assessment can take a variety of forms, including involvement in the development of the assessment criteria themselves, provision of exemplar assessment materials, review and discussion of pre-rated pieces of student work in 'agreement trials', and so on (see Stanley et al. 2009 for a comprehensive description of the kind of support currently being offered to primary teachers in England under the Assessing Pupil Progress (APP) initiative). This kind of professional development for assessment is indisputably necessary for working towards standardisation of judgements in internal assessment. Unfortunately, it is not guaranteed to attain the objective of standardisation. Without empirical checks on the levels of inter-rater agreement actually reached as a result of teacher participation in such activities we cannot know how effective that participation has actually been in contributing to rating standardisation. Unfortunately, 'agreement trials' typically do not actually result in empirical standardisation evidence – the APP pilot evaluated in Stanley et al 2009 is an exception in this sense.

In some countries, such as Sweden, New Zealand and now Wales, entire systems of high-stakes assessment are based on internal teacher assessment, as is the UK's vocational education system. There are concerns about the likely reliability of this assessment (Wikstrom 2006, Crooks 2002), in terms of comparability in standards across districts and schools and between teachers within the same school. But there has as yet been no attempt in any of these contexts to address this issue in any direct way, other than through the encouragement of the kinds of internal standardisation activity noted above, including standards exemplification through benchmark work samples – something that is now under development in Wales (DCELLS 2008). Responsibility for ensuring, or attempting to ensure, comparability in ratings across classrooms and schools is left to external moderation of one kind or another.

5.3 External moderation of internal assessments

There are two principal types of external moderation: moderation by inspection, as practised by the examining boards in the UK, and statistical moderation, as practiced extensively in, for example, Australia (Stanley et al 2009; MacCann & Stanley 2010).

Moderation by inspection describes the situation in which actual samples of candidates' work are scrutinised by an education professional, usually a practising teacher other than the students' own teacher, and decisions made about the appropriateness or otherwise of the original assessments on the basis of the rating comparisons. Statistical moderation describes the situation in which sets of teacher assessments, by school, district or across the state or nation, are compared against sets of external test results, and adjusted to share some of the properties of the test results. Some of these standardisation, or data manipulation, procedures are very complex (see, for instance, Taylor 2005 and Wilmut & Tuson 2005 for examples of different

statistical moderation approaches, and Stanley et al 2009 for specific details of some of the principal methods in use in Australia).

Moderation by inspection would be the preferred strategy, where this is feasible to implement, principally because the procedures involved are transparent and can be easily understood by all stakeholders. In two of the Australian states whose high-stakes assessment systems are described in some detail by Stanley et al (2009), *viz*. New South Wales and Victoria, statistical moderation has become a highly sophisticated activity, with several layers of standardisation being applied at different stages to produce the assessment results that are finally delivered to different stakeholder groups. There must be questions here about the legitimacy of the assumption that tests are inherently more reliable than teachers' internal assessments. At least as importantly, the interpretability of the final assessment outcomes must be an issue when these are sometimes quite distant from students' original achievement scores. Moreover, while statistical moderation has the potential to expose gross differences in the assessment standards of different centres it cannot do the same for inconsistency in the ratings of individual teachers. The only way to do this would be to inspect samples of candidates' work, preferably on a scale large enough to deliver robust evidence to support decisions about the appropriateness or otherwise of the internal assessment.

In the words of Stanley et al (2009):

> If the school assessment scores in a course were accurately equated through consensus moderation across school groups, and public confidence was placed in this process, then it would be logical to merge all these moderated assessments to form a statewide distribution of assessments for the course. … This, however, does not occur. (Stanley et al 2009, p.33)

They note further that "when the assessment shifts to high stakes, all systems rely on some form of external testing" (ibid). It is perhaps the assumption that statistical moderation adequately addresses any issues of internal assessment unreliability that explains why empirical research in this area is lacking in Australia, as it is for different reasons in the UK.

One recent Australian study set out to remedy this situation (MacCann & Stanley 2010). The study exploited test and teacher assessment data furnished by the large-scale examination system for Year 12 school leavers in New South Wales for the five years 2004 to 2008 inclusive. In English and mathematics students take a base examination and an extension examination. The English examinations are essay-based while the mathematics examinations are essentially composed of structured questions. In both subjects the extended examination is intended to be more difficult than the base examination, tapping higher order skills or simply posing more challenging questions. Scripts are externally double marked. Test scores are complemented by teacher assessments for both examinations (i.e. for the coursework underpinning each examination). These are moderated. Without entering into detail – readers are advised to consult the article for this – score distributions for raw test scores and for moderated teacher assessments were converted in the reliability study to percentile grade distributions, and the degree of mismatch in student grade classifications

computed, first for the paired test results and the paired teacher assessments separately and then for the paired composite classifications.

The results in principle showed teacher assessment to be slightly more reliable than test results in the sense of classification consistency, with composite score classifications more reliable than test or teacher assessments alone. However, this increased reliability could be spurious, given the necessary but questionable assumption that the two teacher assessments, like the two test results, are independent measurements. Two assessments of the same student by the same teacher can surely not in reality be considered independent. The fact that the superiority of teacher assessments over test results occurred mainly at the top and middle percentiles would suggest that the extension tests were more successful in discriminating amongst the better students than were the teachers. In other words, it could be claimed that the test results were in fact the more reliable assessments.

External moderation in the GCSE and GCE relies on moderation by inspection, with teachers, or retired teachers, evaluating the work of samples of other teachers' students. This has to be preferable to statistical moderation, given that the teachers are in principle assessing elements of the curriculum that the external tests cannot, and there is no reason therefore to assume that the test results and the teachers' judgements should necessarily coincide, even for individual students.

But there are issues to do with the actual practice of this moderation by inspection. Firstly, the work samples, and even more so the work subsamples, that moderators are required to mark are small, even for large centres, as noted in Chapter 4. Secondly, the moderator has the benefit of knowing the original teacher's marks, so that the moderator's marking is not independent. Given this, to what extent are moderators in some sense ratifying marks rather than genuinely remarking work? If teachers' marks were not available to moderators how many decisions about the quality of centre marking might be changed? Thirdly, if one moderator were to be replaced by another what would be the outcome? The research by Taylor (1992), who demonstrated that moderators, like markers generally, can vary in their marking standards and consistency, with measurable impacts on candidates' grade awards, is highly pertinent here.

There is no systematic form of moderator training in the GCE/GCSE world, as there is for written test markers, and little is known about inter-moderator consistency at the present time. It would be costly and logistically complex to organise inter-moderator reliability studies on a regular basis for every subject that has a coursework element, but some further research does surely need to be carried out. Otherwise any unfairness to candidates in the current system will persist. In the words of Massey and Raikes (undated), offered in the context of a recent study of inter-marker agreement at item level:

> One important contribution to the reliability of examination marks is the extent to which different examiners' marks agree when the examiners mark the same material. Without high levels of inter-examiner agreement, validity is compromised, since the same marks from different examiners cannot be assumed to mean the same thing. Although high reliability is not a sufficient condition for validity, the reliability of a set of marks limits their validity. (Massey & Raikes undated, p.2).

Where it is known, or simply suspected, that markers or raters vary in their judgements of candidates' work, and where it might be the case that individual markers or raters do not always apply standards of marking consistently across candidates, then the best strategy to maximise reliability is to have more than one individual mark samples of candidates' work. This would be a disruptive strategy to implement within a centre during ongoing coursework, but feasible for external moderation, even if costly, time consuming and logistically challenging, given the apparent difficulty in recruiting sufficient teachers to undertake the volume of marking and moderating that this would require.

If there are found to be inter-moderator differences and intra-moderator inconsistency in work sample assessment, the final result of that assessment should be the average moderator mark, since no single moderator can be assumed to be 'the expert'. Indeed, the original teacher's assessment could be included in that averaging process. Further, the practice should properly extend beyond work samples to embrace the work of all internally assessed candidates, just as external tests should be multiple-marked and marks averaged when high marker variation is evident.

5.4 The need for relevant research

It is interesting, if somewhat depressing, to note that in several important reviews of the literature on teacher assessment reliability, as in this one, one of the principal findings has been the sparseness of the literature on the reliability of teacher assessment, and in particular of reports offering quantified evidence one way or the other (Wilmut et al 1996; Harlen 2004; Johnson 2006; Stanley et al 2009; Harth & van Rijn 2011). There is a large literature on formative assessment, and an equally impressive literature on the reliability of the marking of written tests. There are also numerous publications promoting greater use of teacher summative assessment in high-stakes assessment with an expectation of improving validity *and* reliability, such as those noted in Chapter 1. There is in contrast little in the academic literature in the way of reported empirical studies that provide robust quantifications of the degree of reliability achieved.

For example, in her review of teacher assessment reliability Harlen (2004, 2005) initially located over 400 articles that on the basis of abstracts and keyterms appeared to be relevant, but eventually filtered these down to just 30 that were relevant in practice, with just one of these focusing on teacher assessment in GCSE. The same experience is reported by Stanley et al (2009), and by Johnson 2006 and Harth and van Rijn 2011 for vocational education. Johnson (2006, p.60), for example, in his review of UK vocational research in the early to mid-2000s, noted that "there are significant gaps in recent reliability research in the UK, and it appears that we need to go back some time to find empirical work addressing such issues". This comment was made in the context of vocational education, but it applies equally to academic qualifications, and going "back some time" in practice meant going back to the 1980s, before the NVQ, GCSE and national curriculum assessment systems were introduced.

With such a small evidence base available to them it was difficult for these reviewers to offer confident statements about the current state of play. Whether reliability results were positive or less positive depended not only on the nature of the assessments being made (writing, oral skills, art, science investigation skills, etc), but also on the

nature of the achievement evidence being assessed (scripts, portfolios, oral interactions, practical performances) and the tightness of the criteria being used to make the assessments (mark schemes, best fit schemes, level descriptors, etc).

But how to determine reliability for teacher assessment? Where teacher assessment and external tests are intended to be measuring the same constructs then it would be legitimate to compare the two. If the test were considered to be a valid assessment instrument with an established and acceptable degree of reliability then any discrepancies between teachers and tests could be interpreted as indicating that the teachers were not assessing as expected. This assumes, however, that test results and teacher assessments are independently arrived at. Lack of independence was noted in Chapter 2 as a problem for the interpretation of classification consistency data in national curriculum assessment in England (Reeves et al 2001). It has also been an issue for the interpretation of classification consistency rates between A level grades and teachers' grade predictions, the latter being used for higher education admission in the absence of the actual A level results. Studies have consistently shown grade agreement rates to be moderate to low, with variation in the strength of relationship related to the subjects of the qualification (for a comprehensive review see Dhillon 2005). A tendency for teachers to underestimate grades for weaker candidates and to overestimate those of the most able candidates has also recently been exposed (Snell et al 2008).

Whatever the levels of agreement, a disturbing factor for interpretation must be the interdependence of the two assessments. As noted in Chapter 4, teachers' grade predictions form one part of the complex set of information that grade awarders use when deciding critical grade boundaries, so that to some extent actual grade distributions are influenced by the teachers' predictions. On the other hand, although provided ahead of the actual grade awards for their current students, teachers' predictions are made with the benefit of knowledge of actual grade results for several presumably similar previous cohorts of students. Actual grades and predicted grades are therefore not truly independent. And even if independence could be assumed the impact of strategic prediction on the part of teachers (Dhillon 2005; Snell et al 2008) clouds the value of grade agreement rate as an indicator of reliability for either type of assessment.

Where, as in the GCSE, teacher assessment is explicitly included in high-stakes examinations to embrace those aspects of the taught curriculum that tests cannot address, then comparing teacher assessment with test results would not be useful. In such cases the only valid way to investigate the reliability of teacher assessment is to carry out designed inter-rater agreement studies of the kind much practiced currently in the medical and health sciences (Murphy et al 2009).

Some types of teacher assessment are less amenable to reliability investigation than others, particularly where ephemeral outputs are judged. But where the assessment involves the application of marking or rating schemes to tangible physical outcomes, such as samples of students' writing, portfolios or video clips of oral performances, then reliability investigation is clearly feasible. And yet there seem to have been just two relevant inter-rater studies in the GCSE in the last 20 years or so, both having been undertaken at or soon after the launch of the GCSE itself. One of these is the relatively small-scale published study of the reliability of the assessment of oral skills

in French (Good 1988), described in Harlen (2004). The other is the larger-scale internal awarding body study by Taylor (1992) described in Chapter 4 of this report, that, with the addition of generalizability analysis, could serve as a model for future reliability investigation. This would be for the direct exploration of the reliability of internal assessment and for moderator training where moderation is shown to be necessary.

Wood (1991), Wilmut et al (1996) and others have long been recommending the adoption of generalizability theory for reliability assessment in external examinations, and indeed the TGAT report made a similar recommendation for its use in national curriculum assessment (TGAT 1987, Appendix J). Within Ofqual's reliability programme the potential of generalizability analysis in this respect has now been demonstrated, but only for written papers within the GCCE/GCE system (Johnson & Johnson 2011). The methodology is extendable to cover investigation of the reliability of teacher assessment, particularly where this is performance assessment. It is timely to plan such investigations, and to attempt to embed the methodology within ongoing awarding body quality assurance practice.

References

AQA (2008). *GCSE Drama. 2011 Examination. Specimen Controlled Assessment Material.* Manchester: The Assessment and Qualifications Alliance.

AQA (2009a). *General Certificate of Secondary Education. English Literature 4710. Specimen. Controlled Assessment Tasks*. Manchester: The Assessment and Qualifications Alliance.

AQA (2009b). *GCSE Chinese (Mandarin)/French/German/Italian/Spanish/Urdu. Additional Exemplar Tasks. Controlled Assessment Writing and Speaking.* Manchester: The Assessment and Qualifications Alliance.

AQA (2010). *Uniform marks in A-level, GCSE and Functional Skills exams and points in the Diploma*. Version 3.1. Manchester: The Assessment and Qualifications Alliance.

Baker, E.L., Ayres, P., O'Neil, H.F. Choi, K., Sawyer, W., Sylvester, R.M. & Carroll, B. (2008). *KS3 English Test Marker Study in Australia*. London: National Assessment Agency.

Bennett, R.E., Gottesman, R.L., Rock, D.A. & Cerullo, F. (1993). Influence of behaviour perceptions and gender on teachers' judgements of students' academic skill. *Journal of Educational Psychology,* 85, 347–356.

Black, P and Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education,* 5:1, 7-74

Black, P., Harrison, C., Hodgen, J., Marshall, B. & Serret, N. (2010). Validity in teachers' summative assessments*, Assessment in Education,* 17:2, 215-232.

Brennan, R.L. (1992). *Elements of Generalizability Theory* (Second edition). Iowa City: ACT Publications (First edition: 1983).

Brennan, R.L. (2001). *Generalizability theory*. New York: Springer-Verlag.

Burgess, S. & Greaves, E. (2009). *Test scores, subjective assessment and stereotyping of ethnic minorities*. Working paper 09/221, University of Bristol, Centre for Market and Public Organization.

Cale, A. & Burr, K. *Foundation Stage Profile Moderation 2006-07*. Surrey County Council.

CCEA (2010). *GCSE Qualifications. Controlled Assessment Guide*. Belfast: Council for the Curriculum, Examinations and Assessment.

Clarke, S. & Gipps, C. (1998). The role of teachers in teacher assessment in England 1996-1998. *Evaluation and Research in Education*, 14(1), 38-52.

Cronbach, L.J., Gleser, G.C., Nanda, H. & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Crooks, T.J. (2002). Educational assessment in New Zealand schools. *Assessment in Education,* 9:2, 237-254.

Daugherty, R. (2007). National Curriculum assessment in Wales: evidence-informed policy?. *Welsh Journal of Education* 14, 62-77.

Dhillon, D. (2005). Teachers' estimates of candidates' grades. Curriculum 2000 Advanced Level Qualifications. *British Educational Research Journal,* 31:1, 69-88.

DCELLS (2008). *Key Stages 2 and 3 Statutory Assessment Arrangements for school Year 2008/09.* Cardiff: Welsh Assembly Government.

Edexcel (2008). *Sample Assessment Materials. Edexcel GCSE in French (3FR0S)(3FR0W)(2FR0W)*. London: Edexcel.

Elton, L. & Johnston, B. (2002). *Assessment in universities: a critical review of research.* LTSN Generic Centre.

Elwood, J. (2009). The English national curriculum assessment system: a commentary from Northern Ireland. *Educational Research*, 51:2, 251-254.

Falk, B., Ort, S.W. & Moirs, K. (2007). Keeping the focus on the child: Supporting and reporting on teaching and learning with a classroom-based performance assessment system. *Educational Assessment*, 12:1, 47-75.

Gardner, J. (2006, ed). *Assessment and Learning*. London: Sage.

Good, F.J. (1988). Differences in marks awarded as a result of moderation: some findings from a teacher assessed oral examination in French. *Educational Review*, 40, 319-331.

Greatorex, J. (2005). Assessing the evidence: different types of NVQ evidence and their impact on reliability and fairness. *Journal of Vocational Education and Training,* 57:2, 149-164.

Hayward, E.L. (2007). Curriculum, pedagogies and assessment in Scotland: the quest for social justice. 'Ah kent yir faither'. *Assessment in Education*, 14:2, 251-268.

Harlen, W. (2004). A systematic review of the evidence of the reliability and validity of assessment by teachers for summative purposes. In *Research Evidence in Education Library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Harlen, W. (2005). Trusting teachers' judgements: research evidence of the reliability and validity of teachers' assessment used for summative purposes. *Research Papers in Education*, 20:3, 245-270.

Harlen, W. (2007) *Assessment of Learning*. London: Sage.

Harth, H. & Van Rijn, P. (2011). *On the reliability of results in vocational assessment: the case of work-based certification*. Coventry: Office of the Qualifications and Examinations Regulator.

Hauser-Cram, P., Sirin, S.R. & Stipek, D.J. (2003). When teachers' and parents' values differ: teacher ratings of academic competence in children from low-income families. *Journal of Educational Psychology*, 95:4, 813-820.

Hoge R.D. & Coladarci, T (1989). Teacher-based judgements of academic achievement: a review of literature. *Review of Educational Research,* 59, 297-313.

Hutchison, D. & Schagen, I. (1994)(eds), *How reliable is national curriculum assessment?* Slough: National Foundation for Educational Research.

Isaacs, T. (2010. Educational assessment in England. *Assessment in Education*, 17:3, 315-334.

JCQ (2003a). Moderation of GCSE centre-assessed units/components (including centre-assessed units in GCSEs in vocational subjects). Joint Council on Qualifications.

JCQ (2003b). *Moderation of GCSE and FSMQ centre-assessed units/components.* Joint Council on Qualifications.

JCQ (2006). *Setting Standards – The Awarding Process*. Joint Council on Qualifications.

JCQ (2009). *Notice to centres. Moderation arrangements for controlled assessments of Speaking in GCSE Modern Foreign Languages from 2010.* Joint Council for Qualifications, on behalf of AQA, City & Guilds, CCEA, Edexcel, OCR, SQA and WJEC.

Jessup, G. (1991). *Outcomes. NVQs and the Emerging Model of Education and Training*. London: The Falmer Press.

Johnson, M. (2006**).** A review of vocational research in the UK 2002-2006: Measurement and accessibility issues. *International Journal of Training Research*, 4:2, 48-71.

Johnson, M. (2008). Exploring assessor consistency in a Health and Social care qualification. *Journal of Vocational Education and Training*, 60:2, 173-187.

Johnson, R.L., Penny, J. & Gordon, B. (2000). The relation between score resolution methods and interrater reliability: an empirical study of an analytic scoring rubric. *Applied Measurement in Education*, 13:2, 121-138.

Johnson, S. & Cohen, L. (1984). Cross-Moderation: a useful comparative technique. *British Educational Research Journal,* 10, 89-97.

Johnson, S. (1988). Comparability in degree awards: implications of two decades of secondary level examinations research. *Studies in Higher Education,* 13, 177-187.

Johnson, S. (2007). National Assessment informing policy and practice – the Scottish Survey of Achievement. Paper presented at the annual conference of the *Association for Educational Assessment – Europe*.

Johnson, S. (2008). The versatility of generalizability theory as a tool for exploring and controlling measurement error. *Mesure et Evaluation en Education*, 31:2, 55-73.

Johnson, S. (2009). *The reliability of writing in the 2009 survey*. Internal report produced for the Scottish Government.

Johnson, S. & Johnson, R. (2009). *Conceptualising and interpreting reliability*. Ofqual/10/4706. Coventry: Office of the Qualifications and Examinations Regulator.

Johnson, S. & Johnson, R. (2010). *Component reliability in GCSE and GCE.* Ofqual/10/4780. Coventry: Office of the Qualifications and Examinations Regulator.

Johnson, S. & Munro, L. (2008). Teacher judgements and test results: Should teachers and tests agree? Paper presented at the Annual Conference of the *Association of Educational Assessment – Europe.*

Lafontaine, D. & Monseur, C. (2009). Les évaluations des performances en mathématiques sont-elles influences par le sexe de l'élève? *Mesure et Evaluation en Education,* 32:2, 71-98.

MacCann, R.G. & Stanley, G. (2010). Classification consistency when scores are converted to grades: examination marks versus moderated school assessments. *Assessment in Education*, 17:3, 255-272.

Mansell, W., James, M. & the Assessment Reform Group (2009). *Assessment in Schools. Fit for purpose? A commentary by the Teaching and Learning Research Programme*. London: ESRC TLRP, Institute of Education London. Downloadable from: http://www.tlrp.org/pub/commentaries.html.

Martinez, J.F., Stecher, B. & Borko, H. (2009). Classroom assessment practices, teacher judgments, and student achievement in mathematics: evidence from the ECLS. *Educational Assessment*, 14, 78-102.

Massey, A.J. & Raikes, N. (undated). *Item-level Examiner Agreement.* Cambridge: Cambridge Assessment.

Meadows, M. & Billington, L. (2005). *A review of the literature on marking reliability.* London: National Assessment Agency.

Mobley, M., Emerson, C., Goddard, Y., Goodwin, S. & Letch, R. (1986). *All about GCSE*. London: Heinemann.

Morgan, C. & Watson, A. (2002). The interpretive nature of teachers' assessment of students' mathematics: issues for equity. *Journal of Research in Mathematics Education*, 33:2, 78-110.

Murphy, D.J., Bruce, D.A., Mercer, S.W. & Eva K.W. (2009). The reliability of workplace-based assessment in postgraduate medical education and training: a national evaluation in general practice in the United Kingdom. *Advances in Health Sciences Education*, 14, 219-232.

Newton, P. (2009). The reliability of results from national curriculum testing in England. *Educational Research*, 51:2, 181-212.

Newton, P. (2010). Educational Assessment – Concepts and Issues: The Multiple Purposes of Assessment. In E. Baker, B. McGaw, & P. Peterson (eds), *International Encyclopedia of Education. Third Edition.* Oxford: Elsevier.

Ofqual (2008). *Regulatory arrangements for the Qualifications and Credit Framework*. Ofqual/08/37/26. Coventry: Office of the Qualifications and Examinations Regulator.

QCA (2006). *A review of GCSE coursework*. London: Qualifications and Curriculum Authority.

QCA(2007). *Controlled assessments.* London: Qualifications and Curriculum Authority.

QCA (2009a). *Research into marking quality: studies to inform future work of national curriculum assessment*. London: Qualifications and Curriculum Authority.

QCA (2009b). *Changes to GCSEs. Including controlled assessment.* London: Qualifications and Curriculum Authority.

QCDA (2009). *Changes to GCSEs and the introduction of controlled assessment for GCSEs.* London: Qualifications and Curriculum Development Agency.

QCDA (2010a). *Teacher assessment and reporting arrangements. Key Stage 3.* London: Qualifications and Curriculum Development Authority.

QCDA (2010b). *Assessment and reporting arrangements. Key Stage 2.* London: Qualifications and Curriculum Development Authority.

QCDA (2010c). *Assessment and reporting arrangements. Key Stage 1.* London: Qualifications and Curriculum Development Authority.

Reeves, D.J., Boyle, W.F. & Christie, T. (2001). The relationship between teacher assessments and pupil attainments in standard test tasks at key Stage 2, 1996-98. *British Educational Research Journal*, 27:2, 141-160.

Robinson, C. (2007). Awarding examination grades: Current processes and their evolution. In Newton, P., Baird, J-A., Goldstein, H., Patrick, H. & Tymms, P. (eds), *Techniques for monitoring the comparability of examination standards,*.97-123. London: Qualifications and Curriculum Authority.

Rose, J. (1999). *Weighing the Baby. The Report of the Independent Scrutiny Panel on the 1999 Key Stage 2 National Curriculum Tests in English and Mathematics.* London: Department for Education and Employment.

Sainsbury, M. (1994). The structure of national curriculum assessment. In Hutchison, D. & Schagen, I. (eds), *How reliable is national curriculum assessment?* Slough: National Foundation for Educational Research.

Scott, D. (1991). Issues and themes: coursework and coursework assessment in the GCSE. *Research Papers in Education*, 6:1, 3-19.

Scottish Government (2010). Curriculum for excellence Building the curriculum 5 a framework for assessment: executive summary. Edinburgh: Scottish Government.

Shavelson, R.J., Baxter, G.P. & Pine, J. (1992). Performance assessments: political rhetoric and measurement reality. *Educational Researcher,* 21:4, 22-27.

Shavelson, R. & Webb, N. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

Snell, M., Thorpe, A., Hoskins, S. & Chevalier, A. (2008). Teachers' perceptions and A-level performance: is there any evidence of systematic bias?. *Oxford Review of Education*, 34:4, 403-423.

Stanley, G., MacCann, R., Gardner, J., Reynolds, L. & Wild, I. (2009). *Review of teacher assessment: evidence of what works best and issues for development*. Oxford: University of Oxford centre for Educational Assessment.

Stobart, G. (2009). Determining validity in national curriculum assessments. *Educational Research*, 51:2, 161-179.

Taylor, M. (1992). *The reliability of judgements made by coursework assessors.* Associated Examining Board internal report.

Taylor, M. (2005). *Teacher Moderation Systems.* Report commissioned by the National Assessment Agency. Manchester: The Assessment and Qualifications Alliance.

Taylor, M. (2009). *Sample sizes for moderation from summer 2009*. Draft paper for JCQ adoption.

TGAT (1987). *National Curriculum. Report of the task group on assessment and testing.* London: Department of Education and Science.

Thomas, S., Madaus, G. E., Raczek, A. E. & Smees, R. (1998). Comparing teacher assessment and standard task results in England: the relationship between pupil characteristics and attainment. *Assessment in Education*, 5:2, 213–246.

Murphy, R., Burke, P., Content, S., Frearson, M., Gillispie, J., Hadfield, M., Rainbow, R., Wallis, J. & Wilmut, J. (1995). *The reliability of assessment of NVQs*. Report presented to the National Council for Vocational Qualifications. School of Education, University of Nottingham.

Wheadon, C. & Béguin, A. (2010). Fears for tiers: are candidates being appropriately rewarded for their performance in tiered examinations? *Assessment in Education*, 17(3), 287-300.

Whetton, C. (2009). A brief history of a testing time: national curriculum assessment in England 1989-2008, *Educational Research*, 51:2: 137-159.

Wikstrom, C. (2006). Education and assessment in Sweden. *Assessment in Education*, 13:1, 113-128.

Wiliam, D. (2001). Validity, reliability and all that jazz. *Education 3-13*, 29:3, 17-21.

Wiliam, D. (2003). National curriculum assessment: how to make it better. *Research Papers in Education,* 18:2, 129-136.

Willmott, A.S. & Nuttall, D.L. (1975). *The reliability of examinations at 16+*. London: Macmillan.

Wilmut, J. (2005). *Experiences of summative assessment in the UK*. London: Qualifications and Curriculum Authority.

Wilmut, J., Wood, R. & Murphy, R. (1996). *A review of research into the reliability of examinations*. Nottingham: University of Nottingham.

Wilmut, J. & Tuson, J. (2005). *Statistical moderation of teacher assessments*. London: Qualifications and Curriculum Authority.

WJEC (2010). *Key Stages 2/3 Cluster Group External Moderation, Pilot 2010.* Cardiff: Welsh Joint Education Committee.

Wolf, A. (1995). *Competence-based assessment*. Buckingham: Open University Press.

Wood, R. (1991). *Assessment and Testing: A Survey of Research.* Cambridge: Cambridge University Press.

Wyatt-Smith, C. & Castleton, G. (2005). Examining how teachers judge student writing: an Australian case study. *Journal of Curriculum Studies*, 37:2, 131-154.

We wish to make our publications widely accessible. Please contact us if you have any specific accessibility requirements.

First published by the Office of Qualifications and Examinations Regulation in 2011.

Office of Qualifications and Examinations Regulation
Spring Place
Herald Avenue
Coventry Business Park
Coventry
CV5 6UB