

# Grades of uncertainty

Reviewing the uses and misuses  
of examination results



The Association of Teachers and Lecturers exists to promote the cause of education in the UK and elsewhere, to protect and improve the status of teachers, lecturers and non-teaching professionals directly involved in the delivery of education, and to further the legitimate professional interests of all members.

For a free copy of the Association of Teachers and Lecturers' publications catalogue, please call the ATL publications despatch line on 0845 4500 009.

© Association of Teachers and Lecturers 2004  
All rights reserved. Information in this book may be produced or quoted with proper acknowledgment to the Association.

To receive the text of this booklet in large print, please contact ATL on 0207 7930 6441

<b>Introduction</b>	<b>6</b>
<b>The dependability of individual examination grades</b>	<b>8</b>
<b>Comparing grades between different subjects</b>	<b>12</b>
<b>Comparing grades between different years</b>	<b>14</b>
<b>Comparing grades between different Awarding Bodies</b>	<b>17</b>
<b>The implications of grading uncertainties</b>	<b>19</b>
<b>The growth of a media circus? The annual release of GCSE and A-Level results</b>	<b>21</b>
<b>Could examination grades be made more dependable?</b>	<b>23</b>
<b>Living with uncertainty: the way ahead?</b>	<b>25</b>

## Foreword

This publication tackles some of the important issues about examination results, their uses and misuses. It offers cautionary reminders about the variability that is inherent in the current system and about the risks that can arise from drawing conclusions from examination grades alone.

We commissioned this report from Professor Roger Murphy knowing that he would make a key contribution to what is an important debate. We want teachers to be aware of the dangers of too slavish a reliance on results. ATL believes in maximising teaching to learn, and minimising teaching to the test. By furthering the debate, it is my hope that this publication will bring us closer to ensuring that learners have an assessment system that does full justice to their achievements.

Dr Mary Bousted  
General Secretary, ATL

## Biography

Professor Roger Murphy is an acknowledged expert in the field of educational assessment and examinations. He has written a large number of books and articles on this subject and has undertaken research and consultancy work in the UK, for DfES, QCA, and the Awarding Bodies, as well as a number of overseas consultancies. He is a former President of the British Educational Research Association and has been a Professor of Education at the University of Nottingham since 1989. His previous books have included *The Changing Face of Educational Assessment*, *Educational Evaluation: Issues and Methods*, *The Impact of Graded Tests*, *Changing Educational Assessment: International Perspectives and Trends* and *Effective Assessment and the Improvement of Education*. At the University of Nottingham he founded and co-directs the Centre for Developing and Evaluating Lifelong Learning (CDELL), which is well known for its pioneering work in the field of educational assessment and examinations.

## Introduction

Most people remember for the rest of their lives the public examination results they get when they are at school. They are for most of us a very big deal. Much depends upon them and the difference of a single grade can quite literally make a world of difference to an individual who needs a particular result to progress to another course or job.

Increasingly, such grades are also used to judge the performance of teachers, schools, LEAs, and even the effectiveness of government policies about education – and many parents understandably see them as the most important outcome of the whole of their children's school education. Is all of this justified? Should such grades be used as the sole basis for making such important decisions? In this publication I am going to deliver a strong warning about the limitations of public examination grades, due to a range of uncertainties that are necessarily associated with them. Examination grades are undoubtedly only approximate, rather than exact, measures, and as such I will argue that they need to be used sensibly, and with a full understanding of their limitations as indicators of educational achievements.

I am not going to adopt the extreme view which states that all assessments are therefore evil and should be excluded from education altogether. I will argue for a more mature and better informed attitude towards assessment processes and results. This will in many cases involve recognising the assessment processes as compromises, and the assessment results as approximate estimates of achievements in relation to an arbitrarily defined cluster of educational endeavours. Neither of these things will be easy for those of us who like such things to be more straightforward and dependable. But they are, I think, both long overdue within an education system and society, which, I believe, have a distorted view of the value of examinations and the grades that result from them.

A simple system of letter grades resulting from students sitting examinations attempts to place a straightforward set of values onto the diverse achievements of students in hugely different areas of learning. All such judgements are made on the basis of restricted information, much of which is collected through contrived one-off assessments. Furthermore, whatever approach is taken in assessing those achievements will depend upon decisions about which learning is valued most and which learning is valued least. Also, trade-offs have to be made between the amount of precision and detail included within assessment results and the time and resources that are dedicated to assessment processes and procedures.

In the next few sections I will describe in a little more detail why I think that it is appropriate for us to be less certain about the meaning of examination grades than most people are at the moment. In doing this, I will look at three different areas of uncertainty that exist in relation to the meaningfulness of public examination grades, in relation to both how they are produced and how they can be used.

These uncertainties relate to the dependability of individual grades as:

- An ultimate statement of the achievements of individual students in individual subjects.
- A way of comparing the performance of students in different subjects.
- The basis for comparing standards across extended time periods.

### **In conclusion**

- Examination grades are always approximate.
- All decisions about assessment involve compromises.
- It is only possible to use examination grades sensibly if they are seen as approximate indicators.
- There is a risk of giving examination grades unwarranted significance in forming judgements about individuals or about the success of teachers, schools or education systems.

## The dependability of individual examination grades

Sensational debate on the value of public examination grades tends to reveal extreme, polar opposite positions. Thus grades can be presented as worthless, unreliable and misleading pieces of misinformation, or as infallible, unquestionably precise measures of educational achievement - the last word on the matter of the individual's standing in relation to their stage of learning in relation to a particular subject. The reality is, of course, that public examination grades are neither useless nor are they the last word on a pupil's attainment in a particular subject.

In England, Wales and Northern Ireland we are fortunate in having an elaborate and well respected public examination system. Over the years our public examination system has undoubtedly become more sophisticated, even though in recent years it has also experienced increased pressures from educational reforms, and escalating numbers of examinations and student entries. And yet the system necessarily has its own limitations, and these are compounded both by a widespread ignorance of what those limitations are, and by the unrealistic expectations that some people have of the dependability of the grades themselves.

There is still, I think, quite a high level of ignorance about the procedures that the Awarding Bodies follow in the conduct of public examinations. In years gone by that might have been explained by a somewhat secretive approach that tended to obscure the operational details from the public gaze. Nowadays the Awarding Bodies are much more open about their procedures and full information is available through their websites and freely distributed publications – further details are provided at the end of this report. The fact that various myths continue is difficult to understand, even though some of the detailed procedures are so complicated that they are hard to convey to the wider public. Am I alone in wondering whether another reason why many people don't seem to know about the limitations of public examinations is that life is much simpler if you can convince yourself that they don't exist?

Among those who have been closely involved in public examinations, whether as markers, examiners, administrators or researchers, it is well known that huge amounts of care and attention go into making examination grades as accurate and fair as is humanly possible. Nevertheless, those involved in setting, marking and grading examinations also know about the many limitations of the procedures used, and how easily students can end up with better or worse than expected grades depending upon somewhat arbitrary aspects of the assessment processes used. All of this is easy to understand rationally, if you are involved in the processes, but it is a hard message to convey to others within a society which appears to want to hold on to a degree of certainty about the value and dependability of examination grades. Ultimately, one may be trying to convey two apparently contradictory messages. Firstly, we have a public examination system that is highly sophisticated and very professionally organised, and which is the envy of many other countries around the world. Secondly, even such a highly rated system can only produce examination grades which are approximate measures of

educational achievement and which need to be interpreted with great care because there is much uncertainty in relation to what they might mean.

Another notable feature of our public examination system is the considerable amount of research evidence which has been collected about it over the last forty years or so. A major contribution to this has been research conducted by the Awarding Bodies themselves, all of whom have consistently invested in an ongoing programme of research investigations. Alongside that, the Qualifications and Curriculum Authority (QCA) and other educational research bodies have provided significant funding for relevant research investigations, which now give us a secure knowledge base from which to assess the strengths and weaknesses of certain aspects of our public examinations. Books such as Wood (1991), Murphy and Broadfoot (1995) and Stobart and Gipps (1997) will provide individuals who want to delve into this detailed research evidence a much fuller overview than is possible in this publication.

Such research has, among many other things, shown that the same student responding to the same examination paper on two different occasions (perhaps just a few days apart) can quite commonly produce work that will lead to the award of quite different marks and grades. We all have good days and bad days, and one-off examinations have no way of accommodating for that. Increasingly, public examination results depend on the aggregation of marks from different assessment events, and that approach can help to mitigate the worst effects of one-off assessments. Nevertheless, we are still left with dilemmas about how much assessment information needs to be collected on different occasions before a conclusion can be drawn about students' typical levels of achievement. If we go for a small number of assessments, then students may receive grades that reflect their best or worst performances. On the other hand, if we go for large numbers of assessment occasions, we face the danger of producing assessment schedules that are hugely costly to administer and start to erode the time available for learning as opposed to assessment. Again, we are in the realm of difficult trade-offs. How much precision do we need, and how much uncertainty are we prepared to cope with in order to keep assessment as an adjunct to learning, rather than allowing it to dominate all that goes on in education?

Further research investigations have shown that the same student responding to equivalent, but different, sets of assessment tasks, will typically achieve different grades depending upon the particular set of tasks which is used as the basis for determining their result. Assessment tasks that are regarded as equivalent by examiners rarely produce precisely equivalent results when given to students, as student performance is known to be influenced by even slight changes in wording or task context. Attempts to remove gender, ethnic and other kinds of cultural bias from educational assessments have frequently been well intentioned but ultimately doomed to failure. Following a very careful exploration of this issue, Gipps and Murphy (1994) concluded that:

*In an assessment which looks for best rather than typical performance, the context of the item should be the one which allows the pupil to perform well; this would suggest different contexts for different pupils or groups, an awesome task.*

What is being stated here is that the best performance of an individual will occur when their assessment is situated within their ideal context, and assessing them in other contexts may give a misleading impression of what they are capable of. Even though some types of assessment are becoming more flexible and amenable to individual choice through developments such as on-demand testing, the possibility that individuals might be offered assessments tailored to their own preferences, in these terms, still seems rather hard to achieve. However, if we have large numbers of assessments and vary the contexts used within which to present assessment tasks, then we can hope to mitigate against the worst kinds of bias in the assessment results. When only a more restricted set of assessments is involved, all we can hope to do is to avoid the most extreme kinds of bias in the choice of assessment tasks and the contexts within which they are situated. This is far from ideal and remains a significant further cause of uncertainty in interpreting the results of large-scale national examinations.

A further major source of uncertainty in examination grades has been shown to arise from processes used to mark and grade the work that students produce in response to the tasks set. Here we know very well that the responses of any individual student, when marked and graded by different examiners, can result in different grades being awarded. Awarding Bodies on the whole take extensive care to recruit well qualified examiners, give them plenty of guidance and training, and carefully monitor and adjust their marking where necessary in the interests of consistency. Despite all of these steps, the assessment of student work in many areas of the curriculum depends a great deal on 'examiner judgement'. The human judgmental elements in such assessment processes bring with them further elements of variability and uncertainty in the judgements made.

Public examination grades are approximate. They depend upon the judgements of fallible human beings. They depend upon snapshots of student performance under certain conditions, at a certain point in time, in response to a certain set of assessment tasks.

***It would be possible to produce far more dependable grades if:***

- it were possible to spend several weeks and unlimited resources on having a large number of examiners assess each individual student
- the assessments were taken under a variable range of conditions
- a wide range of assessment tasks was always included.

Thankfully, it is deemed unnecessary to invest so heavily in assessment as that, and we still prefer instead to invest more of our available resources in teaching students and promoting their learning than we do on assessing them. If anything, there is still an argument that our highly elaborate public examination and national assessment system is too big a drain on the resources that are available to provide good quality education for all students.

### **In conclusion**

- The same student taking the same assessment on two different occasions may well achieve quite different results.
- The same student taking two apparently equivalent versions of the same assessment may well achieve quite different results.
- The work of a student graded by different examiners may be awarded different results.
- Increasing the range of assessment tasks can mitigate but not entirely avoid variations in student performance.
- Spending more time and resources on assessment can reduce some of the uncertainties related to results, while increasing costs.

## Comparing grades between different subjects

An even bigger challenge in the world of examination grading is the question of how the achievements of students can be compared between different areas of the curriculum. Both GCSE and A-Level examinations are based around common scales of letter grades, regardless of the subject being examined. This level of simplicity is in many ways attractive to the users of public examination grades as they are provided with a currency that they can handle relatively easily. Five or more subject grades A\*-C at GCSE is one of a number of requirements that can be used as a criterion for entry into more advanced courses or other areas of training and employment. Used as a rough and ready measure to indicate something about an individual, or group of students, aggregated grades are quite effective.

However, the difficulty comes when we look more closely into the basis for equating grading standards between subjects. Let us consider the achievements of two students both with a GCSE Grade B result. Let us imagine that one of these students has obtained a Grade B in mathematics and the other in French. In what sense can these achievements be seen as similar or different? Arguing that they are different is fairly easy as we all know that studying mathematics is quite different from studying French, and we all realise that the examination tasks used to assess achievements in these two different areas are bound to have nothing in common. The student with the Grade B in mathematics could if faced with the GCSE French examination perform extremely well or extremely poorly, and their grades in mathematics will give us no clue as to what grade they might get in French. The same, of course, applies to the student with the Grade B in French if confronted with a GCSE mathematics paper.

The only basis upon which the achievements of these two students are equivalent is that we are operating within a socially defined system within which they are deemed to be roughly the same. The assumption is that based on our knowledge of learning in different curriculum areas, the achievements of these two students is roughly equivalent. We choose to treat them as equivalent in order that we can operate a system of examination grading, where we use the same simple grading scale for all areas of the curriculum examined through GCSE and A-Level examinations.

Other areas of the curriculum can have more in common than mathematics and French, and where such subjects cover similar types of learning then it can become possible to draw more meaningful comparisons between assessment tasks and examination grading standards. Thus French examinations and their grading standards may be compared more easily with other foreign language examinations, such as German, and the main Awarding Bodies do all they can to maintain reasonably comparable standards across all subjects at GCSE and A-Level. Nevertheless, despite all of the attempts that are made to make such a system fair and

reasonably equivalent, such a system cannot begin to approach the levels of accuracy required when high-stake decisions are made on the basis of small variations in grades arising from non-equivalent public examinations – such as history and biology.

In other areas of life we live quite happily with the idea that chalk is different from cheese, and apples are quite different from pears. When it comes to education we also know that the study of some areas of the curriculum is totally different from other areas, and appropriate methods of assessment for different areas of the curriculum need to be quite different in order for them to be appropriate to those different areas. For most purposes it doesn't matter too much that grading criteria in different subjects can only be loosely related to each other. For most purposes we only need a rough idea of whether a student is on top of their learning in a particular area of the curriculum or still at a less developed stage. Public examination grades give us a viable and simple language in which to discuss educational achievement in different subject areas. They are a shorthand that avoids the need to give detailed descriptions of individual achievement within the different elements that go to make up individual areas of the curriculum. This shorthand language, however, only works as long as we remember that it is only a shorthand language developed to help us gain approximate snapshot impressions of the achievements of students within sub-areas of the commonly studied complexity of the school curriculum.

### **In conclusion**

- Examination grades are approximate indicators of achievements within defined areas of the curriculum.
- Comparisons between examination grades obtained in different subjects may also carry very limited absolute meaning.

## Comparing grades between different years

Increasingly these days there is understandable public interest in the question of how the achievements of students in one year compare with students in previous years. As government, teachers and all associated with educational provision strive for improvements in facilities, learning resources, training, staff-pupil ratios, etc there is an ever increasing clamour for evidence to indicate whether such changes are making a difference or not. The greater the concern about wanting things to get better, the greater the anxiety can become about whether they are indeed improving.

Despite the attempts by Ofsted (and their equivalent in Wales and Northern Ireland) and others to provide publicly available data on the performance of LEAs, schools, teachers and students, it is natural that the major annual release of public examination results is seen by many as a barometer by which the ultimate success of the system can be judged. Again, public examination results appear seductively simple. The grading scales remain fairly standard over long periods of time and the Awarding Bodies, DfES, QCA and others provide helpful summaries of performance by subject, by school, by LEA, by year, etc.

What appears on the face of it to be a useful national barometer of educational change, however, never quite delivers the clear-cut conclusions that many are hoping for. The overall profile of results at both GCSE and A-Level have shown a steady improvement over the last 15 years or so with proportionately more students entering for these examinations and with more students proportionately achieving higher grades. This steady trend of improved entry patterns and grade profiles is regularly subjected to close scrutiny and critique. Many are asking how we can know whether things are really getting better as opposed to the examinations perhaps becoming a little easier. Forty years ago, only 20 per cent of the age cohort achieved two or more good grades in A-Level examinations - now such results are achieved by well over 40 per cent of the age cohort. Similarly at GCSE there has been a year-on-year improvement in grades achieved ever since the first year of results for that examination in 1988.

If we are uncertain about the extent to which grades in different subjects at GCSE and A-Level are in any meaningful sense strictly equivalent, perhaps we also need to own up to the fact that grades in GCSE and A-Level examinations in different years are very hard to equate with a high level of precision. Again, despite all of the best efforts of our Awarding Bodies, exact equivalence of grading standards for their examinations over considerable periods of time is not something that can be achieved with a high level of certainty. The Awarding Bodies keep examples of examination scripts, so that comparisons can be made, and they instruct their Grade Awarding Committees to do all they can to make grading standards as equivalent as they can from year to year. All involved do the best that they can. However, this isn't always enough, as we saw graphically in 2002 with the national outcry about the grading standards for A-Level. Public confidence in the Awarding Bodies' ability to maintain standards from year to year is based much more on trust in a poorly understood set of procedures, than it is on a system that is robust enough to survive close public scrutiny.

The reality of public examination standards is that they are tailored to specific contexts. Grade Awarding Committees attempt to follow procedures that lead to a fair set of grades in relation to evidence submitted by students following a defined curriculum in particular subject areas. Over time, curriculum specifications and approaches to teaching and learning inevitably change and so therefore must grading standards change as well. If we tried to apply 1984 grading standards in 2004 it would be a complete nonsense. The curriculum necessarily has to move with time and public examinations are designed to reflect the nature of that curriculum as it exists at any point in time. GCSE and A-Level examinations in 2004 have been designed to assess achievement against the goals of the curriculum being taught in 2004, and not the goals that were appropriate in 1984 or 1964.

Gradually, we move back to the 'chalk and cheese' scenario that we faced when we tried to think how achievements in GCSE mathematics could be compared with achievements in GCSE French. How can we possibly compare achievements in public examinations subjects in 2004, with achievements in those same subjects 10, 20, 30, 40 or even 50 years before?

### ***A sporting analogy***

All of this is like trying to compare the achievements of Olympic athletes competing in the 2004 summer games in Athens with those who competed in the Olympics in, say, the 1950s. The modern day athletes on the basis of the records kept clearly out-perform those who participated in similar events 50 years ago. Olympic records in most events continue to improve gradually over time. That process is understood to be the result of multiple factors such as the equipment used, diet, coaching techniques, physique, and the opportunities modern professional athletes have to devote large amounts of time to preparing themselves for their event. It is a process that is expected to continue into the future, and it makes any straightforward comparison between athletes competing in different eras hugely problematic. Sir Roger Bannister famously broke the world record for the one mile race in 1954, achieving the first ever time below four minutes. He was undoubtedly a very high achiever and received great acclaim throughout the world for that achievement. Set alongside the times achieved by athletes running a similar distance in the 2004 Olympic Games his time appears quite mediocre. Thus the same achievement seen in quite different contexts can be judged quite differently. What we can say is that Sir Roger Bannister and the winner of the 1,500 metres race at the 2004 Olympics have both performed at a very high standard indeed within the context that applied when they were running middle distance races. Having said that, it is a matter of pure speculation how they would have fared if given the opportunity to race against each other with access to similar facilities, training techniques, equipment, and diets. Exactly the same argument applies to students sitting public examinations in different years. Their performances are in some abstract sense comparable, but they cannot in any highly sophisticated and specific way be directly compared or equated.

## In conclusion

- Absolute comparisons between examination results obtained in different years can be very misleading.
- The most dependable comparisons are made between results achieved in the same subject at the same level in the same year.
- There is a quite proper sense in which educational standards are constantly changing in line with changes in knowledge and in society. This is one of the major reasons why comparisons between examination results in different years are particularly hazardous.
- Absolute universal examination grading standards do not exist in a way that allows them to be applied consistently to all examinations in all subjects. This is not the fault of teachers or Awarding Bodies, but is a necessary fact of life.

## Comparing grades between different awarding bodies

The fact that there are three large Awarding Bodies in England (AQA, OCR and Edexcel), one in Wales (WJEC) and one in Northern Ireland (CCEA) is a further source of uncertainty for the users of the public examination results. Do all of these Awarding Bodies set different types of examinations and operate with different grading standards, people might wonder? Between boards, variations in examination procedures, grading practices and standards have undoubtedly diminished as the number of examination boards has dwindled since the 1970s when there were well over 25 CSE and GCE boards each operating as quite separate and independent organisations. Equally, the enhanced role of the examinations regulators, the introduction of a Statutory Code of Practice for GCSE, AS and A-Level examinations, and the introduction of a National Curriculum, have all contributed to a situation where there are now far fewer inter-board variations, whether in syllabuses, styles of assessment, or grading standards and practices.

Nevertheless, the fear undoubtedly exists that increased competition between the three big English Awarding Bodies, which each seek to attract entries from the same schools and colleges, might encourage them to try to make their examinations more attractive to their 'customers'. Providing a good service to your customers is fine when it involves effective procedures and a good response to queries, but becomes more problematic when there is a perceived pressure to produce a profile of grades that will prove popular as well. Despite all the assurances that Awarding Bodies give that they try individually and collectively to maintain similar grading standards from one year to the next, this further dimension adds another piece to our jigsaw of the degree of uncertainty that has to be associated with public examination grades. Even though a candidate who gets a Grade C from one Awarding Body might well also have got a Grade C from another Awarding Body, that cannot be guaranteed. For all of the reasons set out in previous sections, different Awarding Bodies, by setting different assessment tasks and employing different examiners, are bound to introduce further elements of uncertainty into the examination grading processes.

The position set out above has from time to time led to calls from some quarters for the introduction of a single national Awarding Body in England to replace the three that exist at present. That is a far from straightforward argument as there would undoubtedly be costs associated with such a move as well as possible benefits. It is much easier for Awarding Bodies to remain reasonably independent from the influence of government than it would be for one such body. In an age of increasing numbers of governmental targets, with many of those set for education depending on improvements in assessment results, there is an advantage in keeping examination grade awarding well away from any perceived interfering hand of politicians. Also, schools, colleges and students all appear to like having a choice between the different syllabuses and forms of assessment that the various Awarding Bodies offer. So up until now the single national Awarding Body solution has not been adopted.

In the meantime, the examination regulators are left with the job of ensuring that these different Awarding Bodies apply similar grading standards.

### **In conclusion**

- There is no guarantee that a student getting an examination grade from one Awarding Body would get the same grade from a different Awarding Body.
- The Awarding Bodies do, however, follow many similar procedures which are regulated by QCA, and conduct regular 'comparability exercises' to attempt to keep their grading standards as close as possible.
- The introduction of a single national Awarding Body in England has been resisted because there are perceived disadvantages as well as advantages in removing choice.

## The implications of grading uncertainties

The fact that public examination grades are not highly precise measures is nothing new. This has been acknowledged in numerous educational reports and in the 1980s was emphasised through Schools Council guidance which stated:

*...research has suggested results on a six- or seven- point grading scale are accurate to about one grade either side of that awarded.*

Schools Council, 1980

Such an acknowledgement of a standard error is quite common in other systems where the measures are known to have a degree of uncertainty associated with them. There is, in my view, no reason at all why this guidance shouldn't apply equally well to current GCSE and A-Level examinations as it did to the examinations being taken in 1980. All of the research evidence that exists would certainly support the idea that such examination grades are no more accurate than + or – one grade on either side of the grade awarded. This provides us with a clear way of thinking about the degrees of uncertainty associated with examination results and should serve as a caution to those who are tempted to use them as though they have a much greater degree of precision and dependability. This estimate relates to uncertainties with any single examination grade, and means that someone who is awarded a Grade B in a GCSE or A-Level subject could just as easily have been awarded a Grade A or C. This is helpful in giving us a feel for the level of uncertainty that should be attached to such grades. Such uncertainties become even greater if we then want to compare such grades across subjects or years.

Anyone closely associated with educational assessment and examinations is likely to be aware of these grading uncertainties. Nevertheless, there is no reason at all why public examination grades cannot continue to perform a useful role within our education system.

For many purposes, I want to argue, examination grades can still provide a sensible source of information, even if they do not represent a highly calibrated measure of educational precision.

Nevertheless, some uses of exam results are more legitimate than others. It is unwise to put results to inappropriate uses, tempting as it may be to do so. What then are the low-risk, as opposed to high-risk, uses of such examination results?

### **Low-risk uses include the following:**

- A useful target to focus students, and teachers, efforts in relation to definable levels of achievement within specific areas of the curriculum.
- A broad indication of the current level of achievement of individual students in individual subjects, taking due regard of the syllabus and assessment procedures in use.

- A useful bit of background information to be used by those wanting to assess the potential of a student to engage in further learning on more advanced courses.
- A useful approximate indicator of the relative achievements of students in a specific subject when compared with their examination grades in other subjects.
- An indication of the success of a teacher in teaching the relevant syllabus, when due account is taken of the potential of the students in their group.
- An indication of the success of a school/LEA, when due account has been taken of the potential abilities of their students, before comparisons are made with other schools/LEAs.

High-risk uses follow fairly easily from ignoring the caveats associated with many of the proposals listed above. They include:

- Assuming that public examination grades tell you how intelligent/able a particular student is, without taking any account of the learning and teaching opportunities that they have experienced.
- Assuming that public examination grades can be used by themselves to predict accurately future performance in other educational courses/employment situations.
- Using raw examination results/data to judge the performance of teachers/schools/LEAs/government initiatives.
- Using examination results as the sole basis for denying students the opportunity to continue with their studies.
- Using examination grades to conclude whether certain subjects are being better taught than other subjects.
- Using year-on-year comparisons of aggregated results of whole cohorts of students to decide whether educational standards are rising/falling and/or whether government reforms are working.
- Using aggregated results as a basis for deciding whether one Awarding Body's examinations are easier or more difficult than another's, without paying attention to differences in entry patterns.
- Comparing percentage pass rates on a year-on-year basis, without paying any attention to changes in the proportion of student cohorts who enter for the examinations.

There are sufficient examples here to illustrate the ways in which examination grades can be added together and used on the basis for all sorts of ill-founded claims. Indeed, their sensible use as a motivator of individual student learning can all too easily be overshadowed by the clamour to turn them into something much more sensational. That leads us into a further consideration of how the media use this data.

## The growth of a media circus? The annual release of GCSE and A-Level results

*...A-Level stories, and their subtext, are predictable. By the time you read this, the highest-ever proportion of good A-Level passes will have been recorded; ignorant commentators will have trumpeted that A-Levels and GCSEs have been 'dumbed down'; ministers will have piously responded that we should 'celebrate the achievements' of our children.*

*New Statesman, 18/8/03.*

One of the more disturbing modern developments in relation to the wider public exposure to debates about public examination results and possible conclusions that can be drawn from them is the media circus that now predictably accompanies the release of the GCSE and A-Level results each year in August. The month of August has become known to some as the 'media silly season': the time of the year when Parliament is in recess and many of the other newsworthy elements of public life go particularly quiet. In order to keep people interested the media have to turn to stories that would not usually command much attention into something more appealing. This is now acknowledged as one of the main reasons why public examination results are regularly presented as major news, even if they are in fact not very noteworthy.

### ***The majority of such examination story headlines contain assertions that:***

- educational standards are falling
- examinations are getting easier
- more students doing well must necessarily be a bad thing
- the increased participation and success of students in schools, colleges and universities will lead to poorer standards.

A recent research study conducted at the University of Nottingham (Warmington and Murphy, 2004) highlighted how a very small set of standard 'media templates' are wheeled out each August in such a way that they can be adjusted to fit whatever pattern of results emerges. Thus the media reporting of public examination results has become quite sterile, ritualised, and polarised, as similar 'media pundits' are called upon to say very similar things to what they said the previous year. Studio debates are staged between those who want to rubbish the examination system and those prepared to defend it. It all makes fine entertainment, which is of course one of the requirements for news reporting. However, it can also have a damaging effect on students, teachers, and parents, whose sense of achievement can quickly be tarnished by such media reports; undermining the very grades that, for such a long time, have been the preoccupation of those who have been working to obtain them.

## In conclusion

- The media reporting of the release of examination results each August tends to be organised around certain set stories, which vary very little from one year to the next.
- During the mid-summer 'media silly season' the release of examination results is a major diary event, which is relied upon to create sensational headlines, even if the pattern of results is very similar to previous years.
- Even tiny fluctuations in percentages of students achieving particular grades in particular subjects may become the focus for sensational reporting.

## Could examination grades be made more dependable?

Many people faced with the uncertainties of public examination grades tend to imagine that there must be obvious ways in which the system can be improved to rule out many, if not all, of the inconsistencies that we have been considering in the previous sections. At the current time a greater use of computers in so-called E-assessment might be seen by some as the way to eliminate so much human judgement and variation in grading practices. Unfortunately this isn't likely to be the case. Although computers can greatly assist the complicated processes involved in issuing literally millions of public examination grades each summer, they cannot turn this process into a highly scientific one, within which all doubts and uncertainties are removed. Computers can be used for swift transmission of information, and E-assessment may involve students and examiners working mostly on-screen, so that student responses can be transmitted directly to the examiners rather than using postal systems for moving large quantities of exam scripts around. Despite the efficiencies that such a system can offer it will not remove the intricate human judgements that individual examiners will have to make when assessing work in different examination subjects. So although E-assessment may bring many benefits and increased efficiencies it won't by itself eliminate grade uncertainties.

Another popular response to the issue of variable human judgements in marking and grading public examinations, is to think of styles of assessment that eliminate the need for any kind of human judgement. This usually means going for assessment formats such as multiple-choice tests, where there is only one correct answer to each question. This approach became very popular in the 1960s and 1970s, especially in the USA. It undoubtedly has advantages as student responses can be entered on machine-readable response sheets, or directly onto a computer, and the marks gained can be computed more or less instantly. However, the difficulty here is that there are only certain kinds of things that can be assessed effectively through multiple-choice questions, and such things do not adequately map onto the wide curriculum coverage that is expected through GCSE and A-Level examinations. In the technical language of assessment you can get very high reliability through the use of multiple-choice questions, but this is usually at the expense of high validity. What this means in everyday language is that multiple-choice tests produce scores that are quite robust, and which could be expected to be repeated on other occasions, etc. However, the scores obtained usually only represent an accurate measure of certain aspects of the subject being examined, and generally do not adequately represent achievement across the broad range of skills and abilities covered in GCSE and A-Level public examination syllabuses.

The search for high validity in examination grades is a vital one, especially in high-stakes situations where one is making decisions about the suitability of students to go on to more advanced courses in higher education, etc. Such a search often causes assessors to use a wide variety of assessment processes to ensure that particular skills and abilities can be

demonstrated and observed in situations that are authentic to the curriculum aims for that particular area of the curriculum. Here we come to the strong case that can be made for including coursework and other extended opportunities for students to demonstrate skills that can never adequately be assessed in short written examinations. Ultimately, there is the argument that the teachers, who have been teaching individual students, will often tend to have a much fuller picture of the achievements of those students than any external examiners can ever gain through viewing their much more restricted examination responses. Thus breadth of response from students through normal classroom work is often viewed as the best basis upon which to assess them, even though this takes us about as far as we can get from multiple-choice tests set under highly controlled timed conditions.

In Section 5, I discussed the other popular solution to the grades of uncertainty problem, that is, the introduction of a single national public examination board. Like the multiple-choice test solution this so-called solution, in taking away one set of difficulties, brings with it an equally problematic new set of potential examination grading problems.

The harsh reality is that there is no simple solution to the grades of uncertainty problem. If there was one it would certainly have been adopted years ago, as a result of the very many reviews that have been conducted into public examinations in the UK and elsewhere. Education is in fact a very complex process, and understanding and summarising the educational progress of individual students will always be a complex challenge. Teachers are well used to the work that needs to go into assessing the potential as well as the current achievements of students, whom they are teaching. This complex reality isn't something that can be easily summarised in the form of a small set of letter grades. Such summaries may have their uses, but those uses will, of necessity, always be limited.

## Living with uncertainty: the way ahead?

In this analysis I have confronted what I regard as an uncomfortable given. Inherent in the use of educational assessments are varying levels of uncertainty. To pretend otherwise is to live in an unreal world.

Progress will come if we can introduce a greater air of reality into debates that are provoked each year by the release of GCSE and A-Level results. Regardless of what reforms are introduced in the future, we will continue to need to accommodate such uncertainties into the uses we make of public examinations.

Education is a highly valuable commodity in our society, but its value cannot be measured in the same way as can gold, silver and other fixed and highly regulated commodities. Educational achievements are varied and difficult to fit neatly into boxes. We can try to summarise them through ever more sophisticated public examinations, but clear descriptions of what individuals have really got out of educational opportunities will always remain elusive. The educational progress of an individual is a unique phenomenon that is very difficult to summarise simply. As with so many areas of life, simple summaries are both superficially attractive and maddeningly frustrating at one and the same time.

## References

Gipps, C. and Murphy, P. (1994) *A Fair Test? Assessment, Achievement and Equity*. Open University Press. Buckingham.

Murphy, R. and Broadfoot, P. (1995) *Effective Assessment and the Improvement of Education – A Tribute to Desmond Nuttall*. The Falmer Press. London.

Schools Council (1980) *Focus on Examinations*, No 5. Schools Council. London.

Stobart, G. and Gipps, C. (1997) *Assessment – A Teacher's Guide to the Issues*. Hodder and Stoughton. London.

Warmington, P. and Murphy, R. (2004) *Could do better? Media depictions of UK educational assessment results*. *Journal of Education Policy*, 19(3): 285-299.

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

## Where to find out more about public examinations

Readers who want to find out more about public examinations and the detailed ways in which they are set, marked and graded, can get access to much information through the websites of the five Awarding Bodies, which operate in England, Wales and Northern Ireland. Additional information is also available from the three regulatory authorities, who in each of these three countries monitor standards and the quality of examinations provided by the awarding bodies.

## Awarding Bodies

AQA (Assessment and Qualifications Alliance) [www.aqa.org.uk](http://www.aqa.org.uk)

Edexcel (The Edexcel Foundation) [www.examzone.co.uk](http://www.examzone.co.uk) or [www.edexcel.org.uk](http://www.edexcel.org.uk)

OCR (Oxford, Cambridge and RSA Examinations) [www.ocr.org.uk](http://www.ocr.org.uk)

CCEA (Northern Ireland Council for the Curriculum, Examinations and Assessments) [www.ccea.org.uk](http://www.ccea.org.uk)

WJEC/CBAC (Welsh Joint Education Committee) [www.wjec.co.uk](http://www.wjec.co.uk)

## Regulatory authorities

QCA (Qualifications and Curriculum Authority) [www.qca.org.uk](http://www.qca.org.uk)

ACCAC (Qualifications, Curriculum and Assessment Authority for Wales) [www.accac.org.uk](http://www.accac.org.uk)

CCEA (Northern Ireland Council for the Curriculum, Examinations and Assessments) [www.ccea.org.uk](http://www.ccea.org.uk)  
(In Northern Ireland this one body covers both the awarding and regulatory functions)



## Association of Teachers and Lecturers 2004

ATL members	FREE
Non-members	£9.99
ATL product code	PR21
ISBN	1902466373

The examination grades students obtain can affect their whole lives and schools and colleges are often judged on their results.

Written by acknowledged expert in the field, Professor Roger Murphy, this study dispels the myths and emphasises just how much our current assessment system is surrounded by degrees of uncertainty. It demonstrates how results can be correctly used – but cautions readers about some of the most frequent misuses of examination results. This clear analysis looks at the issues all teachers should be aware of when considering student performance in terms of external assessments, revealing the uncertainty that lingers behind every grade.