The (un)reliability of public examination grades: evidence, explanation, solutions

Dennis Sherwood, 21st June 2019

Acknowledgements

Over the several years during which I have been working on the issues associated with grade reliability, I have interacted with very many people, all of whom have contributed to my thinking, and to all of whom I express my thanks. That said, may I mention in particular, and thank in a very special way, Bernard Minsky, Michael Buchanan, Neil Sheldon, Nick Hillman, Rob Eastaway and William Richardson.

The data I have used is in the public domain, and I have noted the corresponding sources; much of the analysis, and certainly all the opinions, are mine alone.

Summary

In November 2018, Ofqual published the results of a study measuring the reliability of GCSE, AS and A level examination grades for 14 subjects – Ofqual's definition of 'reliability' being the answer to the question "*A script is marked by an ordinary examiner and awarded a grade accordingly; what is the probability that the same grade would have been awarded had the same script been marked by a senior examiner?*".

Ofqual's research shows that the average grade reliability varies by subject, from about 96% for Mathematics to about 52% for a combined qualification in English Language and Literature. An inference from Ofqual's findings is that the average grade reliability, over all subjects, is about 75%: on average across all subjects, and across GCSE, AS and A level, about one grade in every four as originally awarded is wrong.

These various subject-dependent levels of grade (un)reliability are attributable not to poor marking, but to the fact that the marking of essay-style examination scripts is not precise: different, equally qualified and equally conscientious examiners can, and do, give (slightly) different marks to the same answer. Marking is intrinsically 'fuzzy', and when a fuzzy mark straddles a necessarily hard-edged grade boundary, the resulting grade is unreliable.

The fuzziness associated with a particular subject examination can be measured, and this measure correlates with the reliability of that examination subject's grades. Some subjects, such as Mathematics and Chemistry, are less fuzzy, and have more reliable grades; others, such as English Language and History, are more fuzzy, and have less reliable grades.

To improve grade reliability, a number of approaches could be adopted, for example:

- To change the structure of examinations to a sequence of unambiguous multiple-choice questions.
- To continue with essay-style examinations, and change the mark schemes to reduce the scope for an individual examiner to exercise discretion.
- To continue with essay-style examinations, and change the way in which examiners are appointed and trained, and by which quality control is exercised, so that the range of different marks given by different examiners is reduced, ideally to zero.
- To continue with essay-style examinations, and reduce the number of grades, so increasing grade widths, thereby reducing the likelihood that grade boundaries will be straddled.
- To continue with essay-style examinations, and have differing numbers of grades for different subjects, such that the more reliable subjects have more grades, and the less reliable subjects, fewer grades.
- To continue with essay-style examinations, and change both the policy by which the assessment as shown on the candidate's certificate is determined from the original mark, and also the policy for appeals.

For a script given an original mark represented as m marks, submitted in an examination for which the fuzziness is measured as f marks, this last approach has a number of variants, including:

- Award three grades, determined by each of the marks $m - f$, m and $m + f$.
- Award two grades, determined by each of the marks m and $m + f$, or m and $m - f$, or $m - f$ and $m + f$.
- Award one grade, determined by the mark $m + f$.
- Award one grade, determined by the mark $m - f$.
- Award one, two or three grades, determined by marks of the general form $m + \alpha f$, where α is a number between -1 and $+1$, determined according to a policy defining how reliable grades should be (noting that the policy in force at the time of writing is $\alpha = 0$).
- No longer show a grade on a certificate, but show the original mark m, and also the measure f of the subject examination's fuzziness.

Each of these last six possibilities offers the benefit of significantly improving grade reliability, in principle towards 100%.

None of these six is 'perfect', for it is impossible to achieve both total accuracy and also total reliability simultaneously for examinations structured as questions inviting candidates to demonstrate their knowledge and understanding by writing their own answers, often in the form of essays.

Each possibility has its own combination of benefits and potentially adverse consequences. A wise policy choice therefore requires that each policy is thoroughly assessed, and compared to the default policy of maintaining the *status quo*. This assessment is not the purpose of this paper. Rather, this paper:

- examines the evidence for the measures of grade reliability;
- explains how grade reliability relates to an examination subject's intrinsic fuzziness; and
- presents the various solutions.

This then sets the scene for the subsequent evaluation of the possible solutions as discussed – and indeed additional solutions that might be identified.

The (un)reliability of grades: evidence

From Edgeworth to Ofqual

That examination is a very rough, yet not wholly inefficient, test of merit is generally admitted.

This is the opening sentence of an academic paper by the eminent Victorian statistician, F Y Edgeworth, published in September 1888. Edgeworth's ground-breaking work examines the statistics of marking, and includes sections entitled *Deviations caused by the idiosynchrasies of examiners*, *Errors caused by the negligence of examiners*, and *Amount of displacement due to error: number of candidates who may have been placed in the wrong class.*

Familiar territory. And rather prescient for 1888.

Another study of the statistics of examinations contains these words:

The probability of receiving the definitive grade or adjacent grade is above 0.95 for all qualifications, with many at or very close to 1.0 (*ie suggesting that 100% of candidates receive the definitive or adjacent grade in these qualifications*).

This extract is from a report published in November 2018 by the examination regulator, Ofqual, and the qualifications referred to are GCSE, AS and A level public examinations. What those innocuous-looking words '*or adjacent*' actually mean is that most GCSE, AS and A level grades are right only to one grade either way: a grade B, as appearing on a certificate, might be an A. Or a C. Furthermore, for some qualifications (unspecified), even that might not be true, for there is a 5% chance – that means 5 grades in every 100 – that the B might be an A*. Or a D.

On reading Ofqual's report, Edgeworth would nod in agreement – yes, these results are indeed '*rough, but not wholly inefficient*': the possibility that a B might be an A or a C, or even an A* or a D, is rough indeed, but it is good that we are confident that it is not an E.

But he would probably also be shaking his head in disbelief. As will be shown on page 16, an inference from Ofqual's research is that, of the 6,627,978 GCSE, AS and A level grades awarded for the summer 2018 examinations in England, Wales and Northern Ireland, about one quarter – more than 1.6 million – were wrong. How is it that, 130 years after Edgeworth's study, the '*number of candidates who may have been placed in the wrong class*' – or rather awarded the wrong public examination grade – is so huge?

The purpose of this paper is therefore:

- to explore the evidence concerning the reliability of public examination grades;
- to identify the fundamental reason why grades can be unreliable; and
- to suggest some possible solutions, so that the examination assessments awarded to our young people are no longer, as Edgeworth would say, '*rough*', but are fair and reliable.

Does grade reliability matter?

In the author's opinion, yes. Very much.

Those who use grades do so in good faith, rightfully expecting that a B is indeed a B, and taking important decisions accordingly. But if a B might have been a C or an A had that script been marked by a senior examiner, that faith is broken, and different decisions would have been taken.

In higher education, for example, grades often define minimum entry requirements for courses, and universities competing for the best talent will use grades as a determinant for admission to their most popular and prestigious programmes. An offer of A*, A, A means that a candidate with A, A, A will not be admitted. But if, as Ofqual states, grades are right to only one grade either way, then it as quite possible that a (rejected) candidate actually awarded A, A, A might have been awarded A*, A*, A had the scripts been marked by a senior examiner. Likewise, the (accepted) candidate actually awarded A*, A, A might have been awarded A*, A, B.

"Quite possible", though, is rather vague, and begs the question "in practice, how likely is it that a candidate is awarded the wrong grade?" As will be discussed on pages 13 and 22 to 24, the likelihood of being awarded the wrong grade depends on both the subject and also the mark given. Two inferences from Ofqual's research, however, are that:

- for every 100 candidates sitting A level Mathematics, Further Mathematics and Physics, about 81 receive a certificate on which all 3 grades are right, and about 19 are awarded at least 1 wrong grade; and
- for every 100 candidates sitting A level English Language, English Literature and History, about 20 receive a certificate on which all 3 grades are right, and about 80 are awarded at least 1 wrong grade.

Also, at the time of writing, a debate is raging as regards the suggestion that student loans should be available only to those who have been awarded at least grades DDD at A level, or the equivalent in other Level 3 qualifications. Whether or not there should be an attainment threshold is a debate in itself; but if there is, then such a threshold is meaningless if the grades are unreliable.

Yet another debate concerns contextual admissions: should candidates from disadvantaged backgrounds be eligible for admission to university or college courses with lower grades than those awarded to the more privileged? This is indeed an important question. But if a candidate's grades – whether 'raw' or adjusted for context – are to play any role in admissions, then those grades must be reliable. Unfortunately, they are not.

Furthermore, those who achieve grade 3 in GCSE English Language or Mathematics are obliged to re-sit, and many places at Further Education Colleges are taken up accordingly. In the summer 2018 GCSE English Language examinations, approximately 168,000 students were awarded grade 3. According to the author's simulation (see pages 40 to 43), more than 26,000 of these students would have been awarded grade 4, 5 or even 6 had their scripts been marked by a senior examiner, yet they are being forced to re-sit. Similarly, some 27,000 students were awarded grade 4, but would have been down-graded. The numbers of students originally under-graded (~26,000) and over-graded (~27,000) are about the same, and so awarding reliable grades would not have a significant impact on the number of re-sit places and the costs. But it would have a significant impact on the specific candidates, with the places and resources being allocated to the correct students, rather than being (unknowingly) squandered on the wrong ones.

For an individual candidate, being awarded the wrong grade can have life-changing consequences. If the awarded grade is erroneously low, doors are shut, and opportunities denied. If the awarded grade is erroneously high, the candidate might embark on a further programme of education for which he or she is under-qualified. The candidate might succeed; but alternatively might struggle and either decide to, or be forced to, drop out, having wasted time and resources, having possibly denied a place to someone else, and perhaps having also suffered a damaging blow to self-esteem. Similarly, the candidate might be offered, and accept, a job, and once again struggle, resulting not only in the loss of the job (and all that entails), but also causing the now-frustrated employer to lose confidence and trust in the educational system.

There could be other, more subtle, consequences too. Might an erroneously low grade set up a self-fulfilling prophecy of failure? And what is the effect of an erroneously low GCSE grade on a teacher's forecast of expected A level grades?

Yes. Grade reliability does matter.

Grades reliability – the evidence

Different examiners can give different marks

In November 2018, Ofqual published measures of the reliability of GCSE, AS and A level qualifications for 14 subjects (see Figure 12 on page 21 here). These measures were derived from a comprehensive study of a (very) large number of blind double marked school examinations, in which one mark was given by an 'ordinary' examiner drawn randomly from the community of examiners who normally mark scripts; the other mark was given by a ‘senior’ examiner.

As is well-known, except for questions expressed in terms of unambiguous multiple choice, and especially for those inviting candidates to write essays in which they express their own opinions, it is possible for two different examiners to give the same answer (slightly) different marks. This is explicitly recognised by Ofqual: for example, a posting on their website dated 3 June 2016, point 5 of which states:

> *There is often no single, correct mark for a question. In long, extended or essay-type questions it is possible for two examiners to give different but appropriate marks to the same answer. There is nothing wrong or unusual about that.*

Furthermore, this variability is explicitly recognised, and accepted, within the quality control processes used by the examination boards. As marking takes place, a senior examiner will, from time to time, review the work of an ordinary examiner to determine whether or not the ordinary examiner’s mark is within a defined ‘tolerance’ of their own, presumably expert, mark - where the tolerance is a pre-defined number of marks, dependent on the nature of the particular question. If the ordinary examiner’s mark is within tolerance, then the quality of that marking is confirmed; if the mark is beyond the allowed tolerance, then the senior examiner will intervene accordingly.

Suppose that, in Ofqual’s study, a script was marked 59 by an ordinary examiner, and 60 by a senior examiner. These two marks are, to use Ofqual’s own words, ‘*different but appropriate*’; the difference of just 1 mark is well within tolerance; and ‘*there is nothing wrong or unusual about that*’.

The certificate that the candidate receives, however, does not show the candidate’s mark: it shows a grade, as derived from a subject-dependent scheme by which all marks are mapped onto a scale of grades: for A level, A* (top), A, B, C, D, E, U; for AS, A, B, C, D, E, U; for some subjects at GCSE, A*, A, B, C, D, E, F, G, U, and for others 9 (top), 8, 7, 6, 5, 4, 3, 2, 1, U.

Suppose that for a particular subject, grade C is defined as all marks from 55 to 64 inclusive. In this case, a mark of 59, as given by the ordinary examiner corresponds to grade C, as does the mark of 60 given by a senior examiner. The grades corresponding to the marks of both examiners are the same.

If, however, grade C is defined as all marks from 50 to 59 inclusive, and grade B, all marks from 60 to 69 inclusive, then the grade corresponding to the ordinary examiner’s mark 59 is grade C, but the grade corresponding to a senior examiner’s mark 60 is grade B. **Even though the marks are just one mark apart, that single mark results in the award of different grades** - despite both marks being, once again to use Ofqual’s own words, ‘*different but appropriate*’.

The lottery-of-the-first-mark

This grade difference creates a problem. Which grade is a fair measure of the candidate's attainment? Which grade is right? This rather problematic question will be explored in more depth on pages 49 to 50, and 98 to 100; for the moment, suppose that the senior examiner's mark of 60 is deemed to be 'right', this being in accordance with Ofqual's terminology whereby the mark given by a senior examiner is described as the 'definitive mark', corresponding to the 'definitive grade' (see the caption to Figure 1 on page 6 here) – or, more strongly, the 'true mark', corresponding to the 'true grade' (see pages 2 and 7 here, and the captions to Figures 12 and 13 on pages 23 and 24 here). For the example just given, this implies that the right grade for this candidate's script is grade B, and all other grades are wrong.

In reality, any candidate's script is marked just once, either by a single examiner, or by a team of examiners, with each examiner marking just one question, so avoiding any systematic bias that might arise if a single examiner marks all the questions. It is this first, and only, mark that is used to determine the grade.

If the first examiner is an ordinary examiner who gives the script 59 marks, the certificate would show (the wrong) grade C; but if the first examiner is a senior examiner – or another ordinary examiner who happens to give the same mark as a senior examiner – the script would be given 60 marks and the grade on the certificate would be (the right) grade B. That same script could therefore be awarded a C or a B, but there is no information as to which examiner marked the script, and whether the grade on the certificate is right or wrong. And despite Ofqual's assertion that both marks are '*different but appropriate'* – with the implication that the two marks are therefore equivalent and indistinguishable – there is one, important, feature of these marks that distinguishes between them very much: one mark results in the right grade; the other, the wrong one.

Ofqual's study therefore highlights an important weakness in the way examination grades are determined: the grade that appears on the candidate's certificate depends on the lottery of which examiner (or team of examiners) happened to mark the script first, and on the location of the grade boundaries. This lottery-of-the-first-mark runs deep, and will feature throughout this paper: as will be discussed in various contexts, as long as different examiners can give different marks to the same script, this lottery must exist. The key policy issue is therefore how best to prevent this lottery from damaging any candidate's life chances.

This all casts a deep, dark, shadow over all the grades on all certificates. Which grades, if any, are reliable? How can any candidate know whether any grade is right or wrong? Is this right/wrong dilemma rare, applying to just a few candidates every year? Or are many candidates affected? And, very importantly, why does this happen and what can be done to improve reliability?

The last two questions – "why?" and "what can be done?" – will be explored on pages 16 to 21, 44 to 48, and 58 to 60; the first task is to present the evidence concerning the measurements of grade reliability, and estimates of how many candidates are affected.

Ofqual's measurements of average grade reliability

Ofqual's study is important, for it is the first public statement of the reliability of public examination grades, and the numbers of candidates to whom unreliable grades have been awarded. The design of the research is described in detail in Ofqual's report: in summary, for each of 14 subjects, Ofqual's researchers analysed a data set derived from:

- the blind double marking of the scripts submitted by a full subject cohort (say 100,000) ...
- ... such that one examiner was ordinary ...
- ... whilst the other was a senior examiner, whose mark was deemed 'definitive'.
- The grades corresponding to these two marks were then compared ...
- ... so determining, for any single script, whether the two grades are the same ...
- ... or different ...
- ... thereby allowing, for the whole cohort, the total number of scripts awarded the same grades to be counted (say, 75,000), as well as the total number of those awarded different grades (say, 25,000).
- The average grade reliability for that subject was then calculated as the ratio of the number of scripts awarded the same grades (75,000) to the total cohort (100,000), expressed as a decimal (75,000/100,000 = 0.75), or as a percentage (75%).

A significant feature of this research is that it was based on the double marking of full subject cohorts, embracing all candidates. The results are therefore not biased towards any particular subset of candidates, for example, those candidates given marks just below a grade boundary – as happens for statistical analyses based on appeals. That said, this research does have relevance to appeals, for it answers the question "If an entire subject cohort were to appeal, and if every script were to be fairly re-marked by a senior examiner, for what percentage of scripts would the originally-awarded grade be confirmed?"

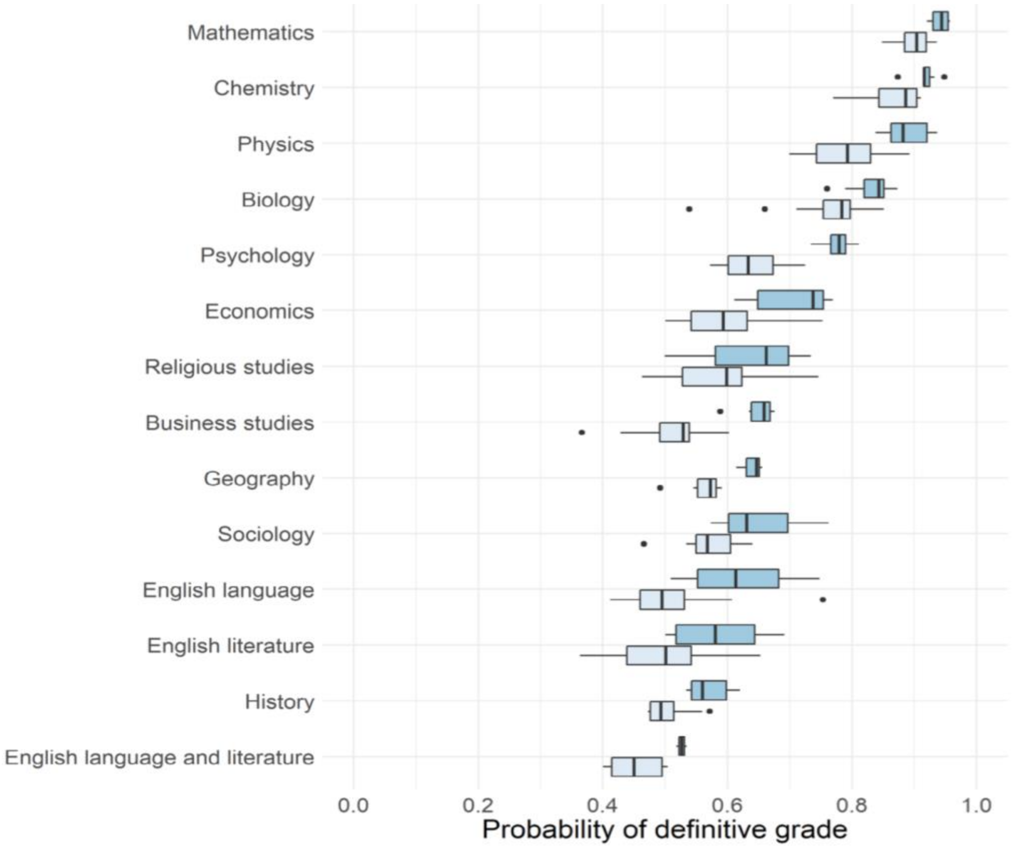
If the grades for a particular subject were fully reliable, then, for every script in the cohort, the grade corresponding to the mark given by an ordinary examiner would be the same as that corresponding to the mark given by a senior examiner. For a cohort of, say, 100,000 scripts, the number of 'same grades' would be 100,000; the number of 'different grades' would be zero; and the reliability would be 100,000/100,000 = 1.00 or 100%.

Likewise, in the (hopefully hypothetical) case of total unreliability, for cohort of 100,000, the number of 'same grades' would be zero; the number of 'different grades' would be 100,000; and the reliability would be 0/100,000 = 0 or 0%.

For any subject, the closer the reliability to 1.00 or 100%, the more reliable the grades; the further from 1.00 or 100%, the less reliable the grades.

With those thoughts in mind, the key results of Ofqual's study are shown in Figure 1, which reproduces Figure 12 on page 21 of Ofqual's report:

Figure 1: Ofqual's measurements of the average reliability of GCSE, AS and A level examination grades for 14 subjects



Source: Ofqual, *Marking consistency metrics: An update*, 2018, Figure 12, page 21

This chart is rather cluttered, but for each of the 14 subjects, the important feature is the heavy vertical black line within the darker blue box: this identifies the average reliability of the GCSE, AS and A level grades awarded for that subject.

As the chart shows, the average reliability of the grades for (all varieties of) Mathematics is about 96% (expressed on the horizontal axis as a probability of 0.96), for Economics, about 74%; for History, about 56%.

Taking the example of Economics, about 74% of originally-awarded grades correspond to those that would have been awarded had a senior examiner marked the scripts, and are therefore right. Accordingly, the remaining 26% of grades must be wrong. Mathematics is better: 96% right, 4% wrong. But History is worse – about 56% right, 44% wrong. Ofqual's report does not give an overall average measure of grade reliability across all subjects, but when the numbers shown in this chart are weighted by the corresponding subject cohort sizes, the average comes to about 75% right, 25% wrong – or, more simply, about one grade in every four is wrong (my own calculation, available on request).

Importantly, 'wrong' in this context refers to grade errors in both directions – some awarded grades are lower than the grades that would have been awarded had a senior examiner marked the corresponding scripts, some are higher. There is no reason why there might be a bias in either direction, and so, if, on average across all subjects, about 25% of all awarded grades are wrong, then about 12.5% of all awarded grades are too low, and 12.5% too high. Also, 'too high' and 'too low' do not necessarily imply just a single grade adrift – as will be shown, for example, in Figure 12, discrepancies of two grades can happen, and even three are possible, albeit rare.

The inference that about 25% of awarded grades are wrong is based on data for only 14 subjects, not including, for example, French, Spanish, music and art. The 14 subjects studied by Ofqual, however, represent over 60% of the total number of grades awarded, and even if all the remaining subjects were as reliable as Mathematics (96% right, 4% wrong), the author's calculations (available on request) indicate that the average reliability would be about 82% right, 18% wrong. Since the subjects not included in Ofqual's research surely cannot all be as reliable as Mathematics, an average reliability of about 75%/25% across all subjects is likely to be a reasonable estimate.

Some consequences of grade unreliability

If, on average, about one grade in four is indeed wrong, there are some important consequences. For example, rather crudely:

- A candidate taking four A levels is quite likely to be awarded one wrong grade.
- A candidate taking eight GCSEs is quite likely to be awarded two wrong grades.

These assertions are 'rather crude' in that they are vague as regards what 'quite likely' means; furthermore, the actual number of wrong grades on any particular A level or GCSE certificate depends on the specific subjects, and also the actual marks awarded. These 'headlines' are, however, worth bearing in mind for they illustrate the scale of the problem; much less crudely, but nonetheless vividly, three direct inferences from the data shown in Figure 1, are:

- For every 100 candidates sitting A level Mathematics, Further Mathematics and Physics, about 81 receive a certificate on which all three grades are right, and about 19 are awarded at least grade wrong grade.
- For every 100 candidates sitting A level English Language, English Literature and History, about 20 receive a certificate on which all three grades are right; about 80 are awarded at least one wrong grade.
- For every 100 candidates sitting GCSE in the 8 subjects Mathematics, Chemistry, Physics, Religious Studies, Geography, English Language, English Literature, and History, about two candidates receive a certificate on which all eight grades are right; about 23, a certificate on which only one grade is wrong; and about 75, a certificate on which at least two grades are wrong.

Unfortunately, it is impossible to identify which particular grades are wrong, and by how much.

As already noted, the overall extent of the (un)reliability of school examination grades was described in the Ofqual report by these words, which appear in the Executive Summary on page 4:

> *The probability of receiving the definitive grade or adjacent grade is above 0.95 for all qualifications, with many at or very close to 1.0* (*ie suggesting that 100% of candidates receive the definitive or adjacent grade in these qualifications*).

What this means is that for ‘many’ qualifications, the grade, as appearing on the certificate, *or an adjacent grade*, is fully reliable. Ofqual’s words also state that for 5% of qualifications (unspecified), a grade B might not perhaps be just an A or a C, but possibly an A* or a D.

Perhaps this will encourage universities, for example, to make offers in a form such as ‘A requirement for admission to [this course] is a minimum of a grade B in [this subject] if your script was marked by a senior examiner, but grades C or D are acceptable if your script was marked by somebody else’.

Unfortunately, no one knows who marks any particular script – perhaps this should be declared.

The (un)reliability of grades: explanation

Why are grades unreliable?

Ofqual's measurements, as shown in Figure 1, imply that the incidence of unreliable grades is not at all rare: on the contrary, very many unreliable grades are awarded each year. For the summer 2018 examinations, for example, 5,470,076 GCSE grades were awarded in England, Wales and Northern Ireland; 346,126 AS grades; and 811,776 A level grades: 6,627,978 grades in total. If about 25% of these were wrong, the total number of wrong grades awarded was more than 1.6 million. As an example for a single subject, in summer 2018, 733,085 grades were awarded in GCSE English language (683,838 graded 9, 8, 7... and 49,247 graded A*, A, B...). According to Figure 1, about 39% of those are wrong – in excess of 275,000.

Why are so many grades wrong?

The 'obvious' answer is as a result of 'marking error' – for example, the failure an examiner to comply with the examination's mark scheme, an administration error, or a failure in quality control.

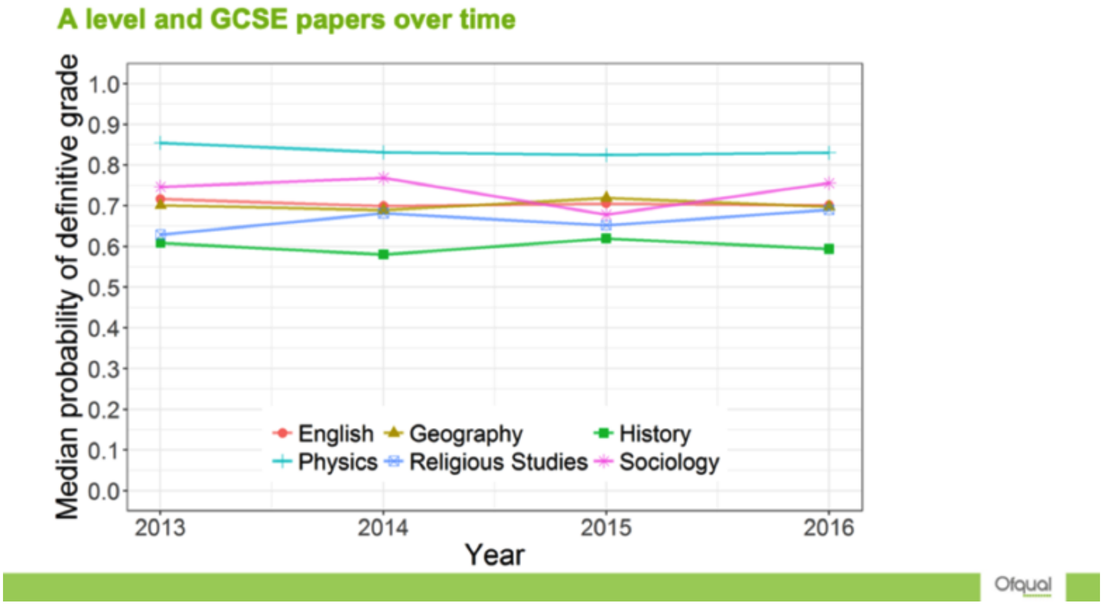
In fact, marking error is not the cause, however 'obvious' this explanation might be.

Although it is not hard, scientific, evidence, an appeal to 'common sense' is informative: is it plausible that the incidence of marking error can be so high as to explain the observed data? Examiners are professional and trained, and the vast majority appreciate the significance of their work, and strive to do a good job. Furthermore, the examination boards all have extensive quality control procedures. Certainly, with over 6.6 million scripts marked in the summer of 2018, there will be some mistakes – but is it plausible for the number of mistakes to exceed 1.6 million? Furthermore, if marking error were indeed the cause, then the number of actual marking errors would in fact exceed 1.6 million by a substantial amount, for 1.6 million is the number of marking errors that result in a grade change – in addition, there must be an unknown number of marking errors that correspond to different marks within the same grade width, and so do not trigger a grade change, but are marking errors none the less.

Perhaps summer 2018 was a particularly bad year, with an exceptionally high number of marking errors. That this is most unlikely is suggested by the statement, on page 4 of Ofqual's November 2018 report, that '*...marking consistency over time* (*between 2013 and 2017*) *appears to be relatively stable – it has neither deteriorated nor improved*'. This implies that the measures of grade reliability, as shown in Figure 1, apply to each of the five years 2013 to 2017, and quite probably to 2018, and any number of years prior to 2013, too.

The stability of the measures of grade reliability is confirmed by the chart shown in Figure 2, showing measures of reliability for each of six subjects over four successive years:

Figure 2: Grade reliability for six subjects over time



A level and GCSE papers over time

Source: Slide 7 from a presentation entitled *Quality of marking: confidence and consistency*, Ofqual Summer Series Symposium, 2017

Figure 2 reproduces a powerpoint slide accompanying a talk given at an Ofqual symposium held in June 2017. The author was not present, and so did not hear the speaker's description, and the slides that can be downloaded from the Ofqual website have no explanatory notes. The chart can therefore only be interpreted from the information as displayed, and that gives rise to three problems.

Firstly, the heading '*A level and GCSE papers over time*' is somewhat ambiguous: does the word *papers* imply '*qualification*' – in which case is the 'definitive grade' (as referred to on the vertical axis) the grade that would appear on the candidate's certificate? Or does '*papers*' refer to individual papers, units or components within a multi-paper qualification? This is not clear, and the slide pack contains no notes; but whatever level of aggregation is reported in this chart, the underlying data set comprises the marks given to individual examination answers. Accordingly, the fact that each subject has nearly the same reliability year-on-year (with 2016 Sociology as an exception) is direct evidence that marking has been stable over the four years shown, and confirms that '*marking consistency over time* (*between 2013 and 2017*) *appears to be relatively stable – it has neither deteriorated* *nor improved'*. If there were more than 1.6 million marking errors across all subjects in 2018, then this number is representative of the incidence of marking errors in previous years too.

Secondly, the chart gives no indication of the uncertainty, or measurement error, associated with each data point. This is of particular relevance to the result for 2016 Sociology, as already noted: is the dip real, indicating a departure from the otherwise essentially flat lines, across the page, for each subject? Or is the dip an artefact associated with the uncertainty inevitably associated with all measurements? The data points as shown imply a high degree of precision, and therefore of accuracy – but unless something special happened in relation to 2016 Sociology that would explain the dip, that data point appears to be somewhat suspect.

Thirdly, the data shown in this chart is somewhat inconsistent with the data shown in Figure 1: although in both charts Physics is the most reliable of the six subjects, and History the least, the sequence of the other subjects is not quite the same. Furthermore, the specific numbers associated with the reliabilities of each subject are similar, but they are not identical. Perhaps these differences can be explained by the fact that Figure 2 relates to GCSE and A level, whereas Figure 1 incorporates AS too; it is hard to tell.

But regardless of all this, the key message portrayed by Figure 2 remains valid: grade reliability, whether of the qualification as awarded, or of a unit, cannot be attributable to marking error.

Further insight as to why this must be the case may be gained by recognising that, strictly speaking, marking error is a misnomer: 'marking' does not make mistakes; mistakes are made by examiners. What is commonly referred to as 'marking error' is in fact a euphemism for 'examiner error': errors made by examiners who are indolent or slovenly; errors made by examiners who fail to comply with mark schemes; errors made by examiners who are inadequately supervised by more senior examiners who themselves are failing to exercise suitable quality control.

Consider the hypothesis that grade reliability is primarily attributable to 'examiner error'. If this is true, then what are the implications of the data in Figure 2? The key inference must be that consistently, every year, Physics examiners are systematically more conscientious, thorough and careful than History examiners, who, by comparison, are lazy and careless; examiners in the other subjects are more professional than their historian colleagues, but rather less so than the physicists.

Such a conclusion must be preposterous. Why, every year, should the examiners in any one subject be consistently more (or less) conscientious than the examiners in any other subject? If examiner error were the primary cause of grade unreliability, it would be much more likely that the rank order of subject reliability would differ year-on-year: in one year, the sociologists might be the best and the geographers the worst, in another, the historians the best and the physicists the worst. This clearly is not what is happening, and so the premise – that grade reliability is primarily attributable to examiner error – must be false.

But if examiner error is not the cause, what is? Figure 2 provides an important clue: the correlation of reliability with the subject. Physics is consistently the most reliable subject, and History the least; Geography and English Language are about the same, and both rather more reliable than Religious Studies, which in turn is more reliable than History; and – discounting the 2016 data point – Sociology is not as reliable as Physics, but more reliable than the other subjects. Is there a feature of the *subject*, rather than of the *examiner*, that might explain this? Yes, there is. 'Fuzziness'.

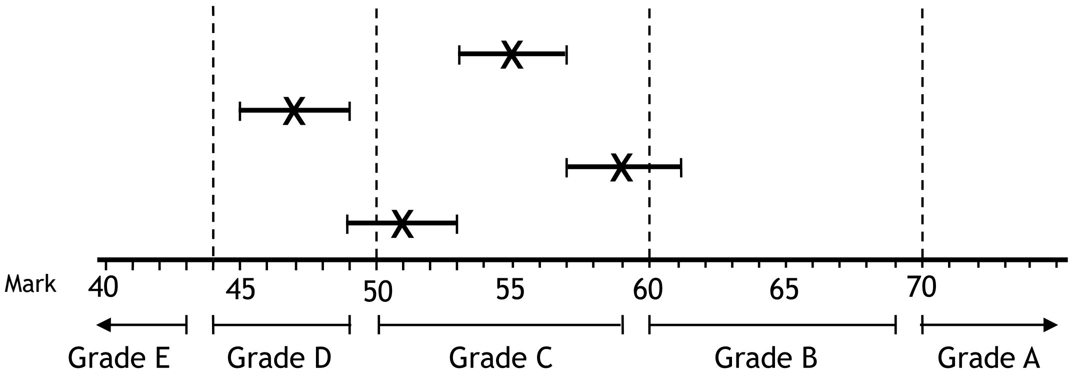
Fuzziness

Fuzzy marks, and grade reliability

As has already been noted, a script does not have a single, precise, mark of, say, 59. Rather, different examiners can legitimately give the same script (slightly) different marks, and so the marks for that script might be 58, 59 or 60. If these three marks lie within the same grade width, then all result in the same grade. But if the C/B grade boundary is 60, then marks of 58 and 59 are grade C, while 60 is grade B. One mark can, and does, make all the difference.

A fundamental truth is that all marking is 'fuzzy'. And when a fuzzy mark straddles a grade boundary, the corresponding grade must be unreliable, as depicted in Figure 3.

Figure 3: Fuzzy marks and grade reliability



Source: Author's graphic

Suppose that a particular script is given a mark X, as shown for four different scripts in Figure 3. Suppose further that each script is fairly re-marked by a senior examiner, who might give the same mark, or a mark or two higher, or a mark or two lower. The range of possible re-marks – in this example two marks either way – is shown by the 'whiskers' symmetrically associated with each original mark X.

The likelihood of any particular re-mark within this range can be estimated from a statistical probability distribution, determined from an appropriately valid sample, as described in detail on pages 108 to 116 in the Appendix; this likelihood has a maximum at the original mark X, and in general decreases symmetrically and progressively on either side (see page 121). The end-to-end range of possible re-marks is one possible way of measuring 'fuzziness', but however measured, as will be discussed shortly, it is the author's contention that fuzziness is an attribute of the examination subject, and not of the mark, or of the candidate. For any given examination subject, the same value for the fuzziness may therefore sensibly be associated with all marks, and hence all candidates – as shown by the same lengths of the whiskers for each of the four marks X in Figure 3.

As illustrated in Figure 3, a candidate given 55 marks is awarded grade C. Since the range of this mark's 'fuzziness' – from 53 marks to 57 – lies completely within grade C, this grade is reliable. A candidate given 59 marks is also awarded grade C, but since the corresponding fuzziness straddles the C/B grade boundary, there is a possibility that a senior examiner might give 60 or 61 marks, grade B. The grade awarded to this candidate is therefore unreliable – as is the grade awarded to a candidate originally given 51 marks.

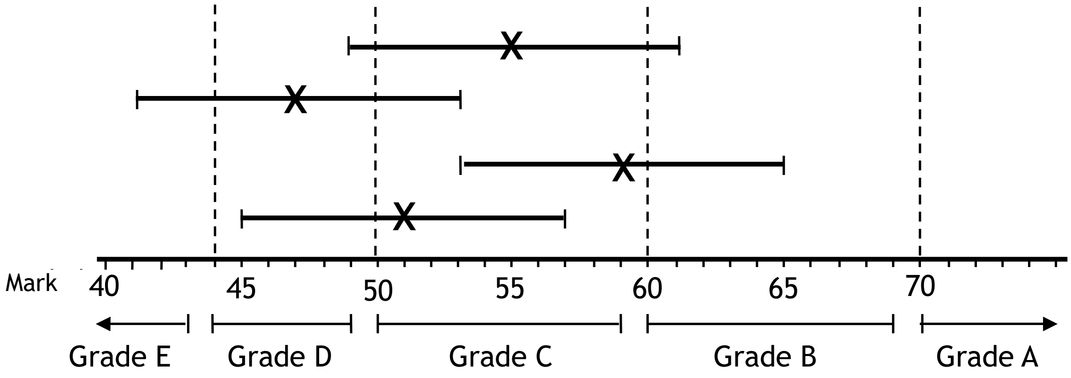
Only the grades corresponding to all marks within the unshaded zones in Figure 3 are fully, 100%, reliable; the grades for all marks within the shaded zones are unreliable, increasingly so as the mark approaches a grade boundary.

As can be seen in Figure 3, grade D is narrower (6 marks) than each of grades B and C (10 marks), and within grade D only the grade corresponding to a mark of 47 is fully reliable. This implies that, for any given subject associated with a specific fuzziness, the wider the grade width, the more reliable the corresponding grades; conversely, the narrower the grade width, the less reliable the grades.

Some subjects are fuzzier than others

All teachers – and many non-teachers too – know, intuitively, that some subjects are inherently fuzzier than others. A question in a Physics paper, for example, might have the majority of marks being given for a mathematically correct answer, with perhaps just a mark or two given at the examiner's discretion for, say, the clarity of the accompanying explanation. In contrast, an open-ended essay question in a History examination might have a range of several marks. History is intrinsically 'fuzzier' than Physics, and this has a significant impact on the reliability of grades, as illustrated in Figure 4

Figure 4: A fuzzier, and more unreliable, subject



Source: Author’s graphic

In Figure 4, the grade boundaries are the same as in Figure 3, as are the marks X given to each of the four candidates. The only difference is that the fuzziness is now six, rather than two, marks either way, and so this diagram depicts a fuzzier subject than that shown in Figure 3. In this example, the 12-mark end-to-end fuzziness is greater than each of the grade widths. As a consequence, all marks in the range shown in Figure 4 straddle at least one grade boundary, and no mark results in a fully reliable grade. Furthermore, some marks straddle two grade boundaries, notably the marks at the centres of each grade.

The grades associated with the subject depicted in Figure 4 are therefore consistently less reliable than those associated with the subject depicted in Figure 3, and it is the intrinsic fuzziness of each subject that explains the observed data, as shown in Figures 1 and 2.

The sequence of the subjects shown in Figure 1, for example, is consistent with intuitions regarding fuzziness: Mathematics is intuitively the least fuzzy subject and has the most reliable grades, with Chemistry, Physics and Biology close behind; it is no surprise that English Language, English Literature and History, as intuitively the most fuzzy subjects, have the least reliable grades; subjects such as Geography and Business Studies are more fuzzy than the pure sciences but less so than the humanities, and are somewhere in-between. A similar explanation applies to Figure 2 – recognising, as already noted, that the sequence of subjects in Figure 2 is slightly different from the sequence shown in Figure 1, perhaps because of the inclusion of AS results in the data associated with Figure 1.

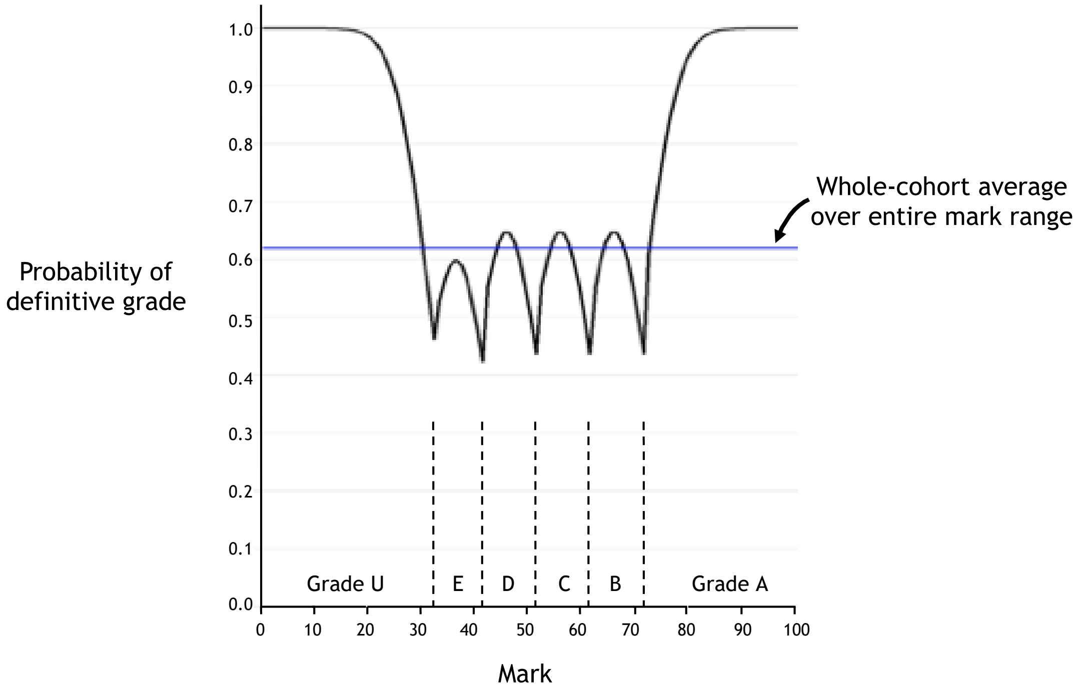
Grade reliability by mark

A particular case – AS History

Figure 1 shows, for each subject, the average grade reliability, across GCSE, AS and A level, for the entire subject cohort. For an individual candidate, however, the reliability of the grade associated with a specific mark in a specific subject is much more important.

An example of an analysis of grade reliability by mark is shown in Figure 5, which reproduces (with some additional annotations) Ofqual's measures of the reliability of the grades, for each mark, for AS History.

Figure 5: Grade reliability by mark, AS History



Source: Ofqual, *Marking consistency metrics: An update*, 2018, Extract from Figure 11, page 20

The horizontal axis represents the range of all marks, standardised from 0 to 100. This exam is graded A, B, C, D, E and U, and the grade boundaries are as shown.

The vertical axis represents grade reliability, expressed as a probability, on a scale from 0.0 to 1.0, corresponding to percentages from 0% to 100%.

The horizontal line at about 0.62 (or 62%) shows the whole-cohort average over the entire mark range. This value is different from the value of 56% for the average reliability for History, as shown Figure 1, but Ofqual give no explanation of this – perhaps this discrepancy is attributable to the fact that 56% is an average over the total cohorts of GCSE, AS and A level, but 62% is the average for AS only.

For any specific mark, the corresponding point on the 'wiggly line' answers two (equivalent) questions:

- What is the probability that a script given a specific mark is awarded the 'definitive' grade?
- What percentage of all scripts given that specific mark will have the original grade confirmed if all those scripts were re-marked by a senior examiner?

If grading were fully reliable, this chart would show a straight horizontal line, at a probability of 1.0 (100%) for all marks. As can be seen, the only marks for which the corresponding grades are fully reliable are those greater than about 85 (good As) or less than about 20 (poor Us). All other marks are unreliable to a greater or lesser degree.

This chart shows a number of important features, many of which are consistent with History being a more, rather than less, fuzzy subject, as validated by Figure 1 and illustrated in Figure 4:

- The grade boundaries are clearly visible, and for marks at, or very close to, any grade boundary, the probability of being awarded the right grade is about, or less than, 50%. This implies that it is more likely that the wrong grade will be awarded than the right one. Tossing a coin would be more fair.
- For many of the marks associated with the intermediate grades B, C, D and E, corresponding to most of the candidates, the probability of being awarded the right grade is less than the whole-cohort average.
- The intermediate grades B, C, D and E are most reliable at the centre of the grade width, as shown by each local peak, and progressively less reliable towards either grade boundary, as is consistent with Figure 4.
- The maximum reliabilities of grades B, C, D and E are all significantly less than 100%, implying that no mark – not even the mark at the centre of the grade width – corresponds to a fully reliable grade. The fuzziness associated with this central mark therefore cannot lie wholly within the grade width but must straddle both grade boundaries simultaneously.[*] The overall end-to-end fuzziness of any mark is therefore greater than each of the B, C, D and E grade widths – as illustrated in Figure 4.
- If a script marked at the middle of, say, grade B, is fairly re-marked, there is a probability of about 35% (roughly 1 chance in 3) that the grade will be changed – even though the mark is as far from a grade boundary as possible. Because the fuzziness straddles both the B/A and C/B grade

[*] The fuzziness associated with each of the marks shown in Figures 3 and 4 is shown as being symmetrical, with the whiskers extending equally on either side. As discussed on page 121 in the Appendix, this symmetry is highly likely, but there are some circumstances in which the fuzziness can be asymmetrical.

boundaries, the grade change might be to an A. But it is just as likely to be a C. This is consistent with the Ofqual statement that grades are reliable one grade either way: '*The probability of receiving the definitive grade or adjacent grade is above 0.95 for all qualifications, with many at or very close to 1.0* (*ie suggesting that 100% of candidates receive the definitive or adjacent grade in these qualifications*)'.

- Grade A approaches 100% reliability at about 85 marks, some 13 marks higher than the B/A grade boundary at about 72 marks. This provides an estimate of AS History's fuzziness, suggesting that the overall fuzziness associated with any mark – say, 59 – might be expressed as 59 plus or minus 19, represented mathematically as 59 +/– 13 or 59 ± 13. This estimate is confirmed by the number of marks between the U/E grade boundary (about 33 marks), and the point at which grade U becomes 100% reliable (about 20 marks). It is also consistent with the inference that the end-to-end fuzziness (estimated as 2 x 13 = 26 marks) is greater than each of the B, C, D and E grade widths.
- The left-right symmetry of each 'arch' is strong evidence that fuzziness is symmetrical about a script's mark, as illustrated by the symmetrical 'whiskers' in Figures 3 and 4. If this were not the case, the 'arches' would not have their centres at the middle of each grade width, but would be skewed either to the left or the right, as evidenced by the author's computer simulations (available on request).
- The fact that the 'arches' associated with grades B, C and D are identical is strong evidence that the value of the fuzziness is the same for all marks associated with grades B, C, D. This supports the assertion made on page 20 that fuzziness is a property of an examination, and not of a particular mark or script. If the fuzziness were to vary for different marks for the same examination, the shape of the 'wiggly line' would be much less regular, as also evidenced by the author's computer simulations (available on request).
- The maximum grade reliability associated with each of grades B, C and D is the same, and greater than the maximum grade reliability of grade E. Careful measurement will show that grades B, C and D are of the same width (10 marks), but grade E is narrower (9 marks). This is a particular instance of a general rule that the wider the grade width, the more reliable the grade; the narrower the grade width, the less reliable the grade.

Grade reliability and grade width

This last point has particular relevance to the change in the grading structure of GCSE from letters (A* to G, plus U) to numbers (9 to 1, plus U), as first implemented for three subjects in summer 2017, and extended to a further 20 in summer 2018. By policy, the old C/D boundary is pegged to the new 4/3 boundary, implying that the six new grades 9, 8, 7, 6, 5, 4 occupy the same total width as the four old grades A*, A, B, C. The average grade width has therefore become narrower. As a consequence, the higher GCSE grades have become more unreliable.

That this change in the grading structure would have this effect was known by Ofqual before the new grading structure was implemented. In a report published in November 2016, before the new grades took effect, these words will be found on page 21: '*Thus, the wider the grade boundary locations, the greater the probability of candidates receiving the definitive grade. This is a very important point*'. What this does not, explicitly, say is the converse: the narrower the grade widths, the less reliable the grades – but this is surely implied. The reduction in GCSE grade reliability has therefore happened, if not deliberately, at least knowingly. This begs the question: what is the point of increasing the number of grades knowing that those grades will become (even) more unreliable?

The relationship between grade width and grade reliability also has an implication as regards the interpretation of Figure 1, which, as has been seen, shows, for each of 14 subjects, the average grade reliability across GCSE, AS and A level. All current A levels are assessed according to seven grades (A*, A, B, C, D, E and U); all AS examinations according to six (A, B, C, D, E and U); some GCSE examinations according to nine (A*, A, B, C, D, E, F, G, U); and other GCSEs according to ten (9, 8, 7, 6, 5, 4, 3, 2, 1 and U). All share the same overall standardised mark scale from 0 to 100, and so for any subject examinable at each level, and assuming that the grades at each level are of equal width, AS grades are (marginally) more reliable than A level grades, which in turn are (less marginally) more reliable than GCSE grades – with those graded A*, A, B... being more reliable than those graded 9, 8, 7... .

The specific issue concerning the top grades A*, A, B, C and 9, 8, 7, 6, 5, 4 has already been discussed; the issue here concerns the averages shown in Figure 1. This average is the average of the reliability at each of the three levels, with each level's reliability weighted by the corresponding cohort size. As a consequence, for each subject offered at each level, the average reliability of the A level and AS examinations must be *greater* than as shown in Figure 1, and the average reliability *less* than as shown.

Taking English Language as an example, Figure 1 shows the average grade reliability over all levels to be about 61%. The average reliability of A level English Language is therefore greater than this number, and the average reliability of GCSE English Language, graded either way, will be less than this number.

How much greater, and how much less, the author does not know; however, given that the GCSE cohorts (2018, 49,247 graded A*, A, B... and 683,838 graded 9, 8, 7...) are significantly greater than the cohorts for AS (4,581) and A level (18,049), the average reliability for GCSE will be closer to, and rather less than, 61%; the average reliabilities for each of AS and A level will be further from, and rather greater than, 61%.

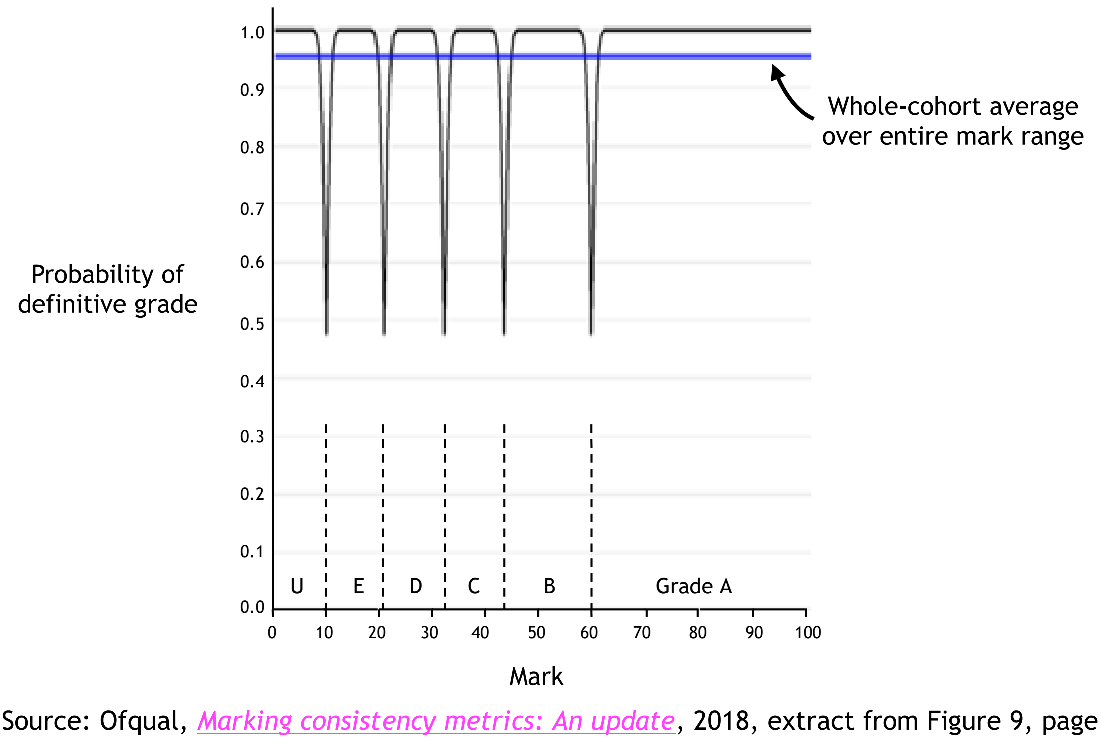
This uncertainty is unhelpful. More informative data should be routinely published: certainly the average grade reliability for each subject, at each level, for each board, and by mark.

Grade reliability by mark – the general case

Figure 5 refers to AS History, and, according to Figure 1, History is one of the least reliable subjects. But the features identified in Figure 5 apply to measures of grade reliability by mark for all subjects, at GCSE, AS and A Level, and for any number of grades of any widths.

Accordingly, all charts will show 'wiggly lines', with minima at the grade boundaries and maxima at the centre of each intermediate grade. The values of the probabilities associated with those minima and maxima, however, will vary according to the subject: for Mathematics, for example, the mid-grade maxima will reach 100% but the minima will still dip towards 50% at the grade boundaries, as illustrated in Figure 6, which shows the grade reliability by mark for foundation tier reformed GCSE mathematics, graded A, B, C, D, E, and U.

Figure 6: Grade reliability by mark, foundation tier reformed GCSE mathematics

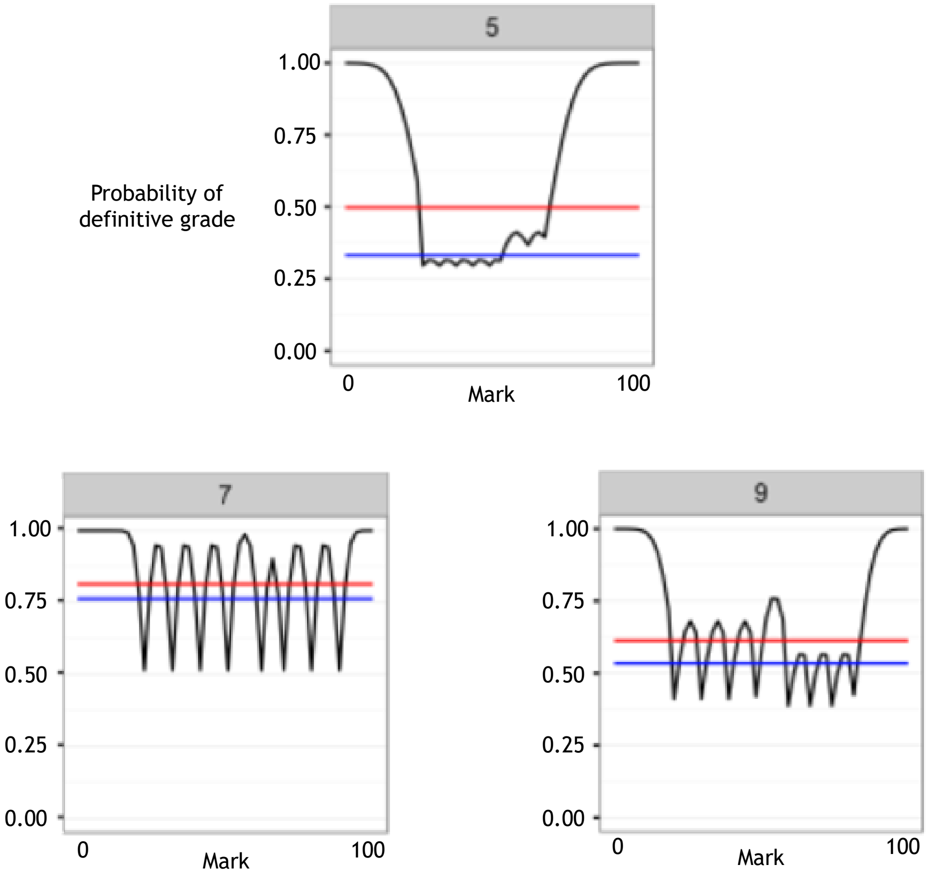


Source: Ofqual, *Marking consistency metrics: An update*, 2018, extract from Figure 9, page 19

Mathematics is, of course, a less fuzzy subject, and so Figure 6 is consistent with Figure 3: marks that are not close to a grade boundary are fully reliable.

Other examples are to be found in Ofqual's November 2016 report, a particularly dramatic instance being that reproduced as Figure 7, which shows the reliability of three GCSE units, graded A*, A, B, C, D, E, F, G, U, within an (unidentified) humanities subject.

Figure 7: Grade reliability by mark, units 5, 6 and 7 within an unidentified GCSE humanities subject



Source: Ofqual, *Marking consistency metrics*, 2016, extract from Figure 13, page 24

The two horizontal lines show two different averages: the upper, red, line is the whole-cohort average, across the entire mark range and calculated on the same basis as the whole-cohort averages shown in Figures 1, 2, 5 and 6; the lower, blue, line is an average excluding those candidates awarded the top grade A* and the bottom grade U. For the current purposes, these averages are less important than the 'wiggly lines'. As expected, these show the same features as those discussed in connection with Figure 5. For unit 7, other than towards the grade boundaries, the reliabilities of the grades are between about 75% and 95%. But for unit 5, for every candidate awarded grades A or B, the reliability is about 40% (6 grades in every 10 wrong), and for every candidate awarded grades C, D, E, F and G, about 30% (7 grades in every 10 wrong).

The appeals process is failing

It might be thought that post-results appeals (Ofqual uses the term 'challenges') would resolve all these grading errors. It does not.

As an illustration, consider the summer 2017 GCSE examinations in England, Wales and Northern Ireland. According to Ofqual's official statistics, a total of 5,470,385 grades were awarded, of which 277,960 grades were challenged; 50,875 grades were subsequently changed, 50,680 up and 195 down.

Based on these figures, the number of grade changes (50,875), expressed as a percentage of the total number of grades awarded (5,470,385), is 0.9%. At first sight, this appears to imply that the remaining 99.1% of grades were right-first-time, and a claim of this nature has been made, in public, by one of the examination boards.

In fact, the examination board's inference that 99.25% of their grades were right-first-time is false. Grades can be changed only if a challenge has been made, and the fact that the number of up-grades is very much greater than the number of down-grades indicates that appeals are – quite understandably – made by candidates with marks just below a grade boundary, in the hope of an award of an up-grade. If, however, a grade is not challenged, there is no knowledge as to whether or not the grade would have been changed, had an appeal been made. Using once again Ofqual's overall statistics for 2017 GCSE, since 277,960 grades challenged and 50,875 grades were changed, it is true to state that those 50,875 grades were originally wrong, and that the 277,960 – 50,875 = 227,085 grades that were confirmed were originally right. But it is false to state that the 5,470,385 – 277,960 = 5,192,425 grades that were *not* challenged were all right: it is, in principle, possible that all 5,192,425 were in fact wrong, with the grading errors remaining undetected simply because no corresponding appeals were made.

Which leads to the question "how many undetected grade errors might there be?".

The publication of Ofqual's November 2018 report allows this question to be answered with confidence. Of the GCSE 5,470,385 grades awarded, about 25% – about 1,360,000 – were wrong, of which about one-half, say, 680,000, are grades that would be re-graded downwards, and the other half, also about 680,000, would be re-graded upwards. The number of candidates that would have been awarded an up-grade as the result of a challenge is therefore about 680,000. The official statistics, however, state that 50,680 up-grades were in fact awarded, implying that some 680,000 – 50,680 = 629,320 candidates missed being awarded an up-grade because they did not appeal. 92.5% of those eligible for an up-grade did not receive one. If an objective of the appeals process is to resolve grading errors, it is failing.

Why do so few candidates appeal? Up until 2016, there were three primary reasons.

Firstly, a fee has to be paid in advance, and although this fee is refunded if a grade change is made, the initial payment is a disincentive, especially for state-funded schools who are under increasingly burdensome financial pressure. In this regard, the author notes that a 2017 judgement of the Supreme Court ruled that the fee for bringing a claim to the Employment Tribunal is unlawful on the basis that this fee is a barrier to justice. Might the same be said of the examination appeal fee?

The second reason is trust. Most students and teachers trust 'the system'. So when a grade is awarded, many candidates will say, 'Oh dear, I haven't done as well as I had hoped', and their teachers will commiserate accordingly. How many will say, 'The system is totally unreliable – that grade must be wrong!'?

The third reason is 'ignorance'. Suppose that the candidate, or the teacher, does indeed think that the awarded grade is wrong. On what basis can that judgement be made? There is no 'second opinion', and neither the candidate nor the teacher can know, with any confidence, that a grading error has been made. As with all professional opinions, the layman is obliged to accept what the expert says.

But since 2017, there has been another reason why candidates do not appeal. They are not allowed to.

In the summer of 2016 – shortly before the publication of the November 2016 report – Ofqual announced some changes to the rules under which candidates can, and cannot, appeal their awarded grades. In particular, to use Ofqual's own words: '*It is not fair to allow some students to have a second bite of the cherry by giving them a higher mark on review, when the first mark was perfectly appropriate*'.

Ofqual fully recognise that '*it is possible for two examiners to give different but appropriate marks to the same answer*', yet they also insist that the grade corresponding to the first-given mark must stand, and cannot be appealed, for this is '*having a second bite of the cherry*', which Ofqual deems '*unfair*'.

It is indeed true that '*it is possible for two examiners to give different but appropriate marks to the same answer*'. But although the marks are '*different but appropriate*', as noted on page 11, those different marks can be of very different qualities. One might be given by an ordinary examiner, and the other by a senior examiner. One might be wrong. And the other right.

But the changes introduced by Ofqual in 2016 deny the opportunity to right this wrong. Furthermore, Ofqual are asserting that any appellant who attempts to do so is being '*unfair*'. Perhaps the unfairness is not in the appellant's after-the-event appeal, but in the lottery of happening to have been marked by an ordinary examiner whose mark, '*appropriate*' and within 'tolerance' as it indeed is, just happens to be on the wrong side of a grade boundary, as compared to a senior examiner's mark.

The appeals process is failing.

In this context, it is appropriate to note that Section 22 of the Education Act 2011 places an obligation on the Chief Regulator of Qualifications and Examinations '*...to secure that regulated qualifications give a reliable indication of knowledge, skills and understanding ...*' . Perhaps it is not just the appeals process that is failing.

A statistical explanation of Ofqual's results

Quantifying fuzziness and reliability

A central feature of the discussion so far, and the key to the explanation of Ofqual's results as shown in Figures 1, 2, 5, 6 and 7, has been fuzziness – the recognition that different examiners can give the same script '*different but appropriate marks*', to quote Ofqual once again.

The purpose of this section is to present a more analytical explanation of Ofqual's results, based on the key features of the statistics that underpin fuzziness. A detailed and mathematically rigorous discussion is given in the Appendix; this section features the numerical and graphical results derived from a computer simulation, undertaken by the author (from whom full details are available on request), of 2018 GCSE Geography, graded 9, 8, 7....

Each of the cohort's 243,392 scripts is assumed to be marked (once) by a fully qualified and conscientious examiner, drawn randomly from the pool of all examiners, and who is therefore not necessarily, but might be, a senior examiner; a further assumption is that there are no marking errors. The number of candidates awarded a given mark, over the range from 0 to 100, is determined according to the bell-like symmetrical shape that mathematicians refer to as a 'Gaussian' or 'normal' distribution. Some candidates are given a high mark, 70 or above; some a low mark, 30 or below; most are in the range 31 to 69.

According to the simulation, 5,565 scripts were given 59 marks, all of which were awarded grade B. Suppose that one of these scripts is then fairly re-marked by a senior examiner. The re-mark might be the same as the original mark, 59, but it might be a few marks higher or lower – so suppose further that, in this particular case, the re-mark is 62.

A second, different script, also originally given 59 marks, is re-marked, and given 58 marks. And a third, and a fourth... until the 5,565 scripts, each originally given 59 marks, have been fairly re-marked (once) by a senior examiner, giving a total of 5,565 original mark/re-mark pairs. The results of the author's simulation of this re-marking are shown in Table 1.

Table 1: Author's simulation of the fair re-marking, by a senior examiner, of 5,565 2018 GCSE Geography scripts, each originally marked 59 and originally awarded grade 7, and re-graded according to the grade boundaries given on page 32

Re-mark	Re-grade	Number of scripts given re-mark	% of scripts given re-mark	Difference between re-mark and original mark
45	4	0	0.00	– 14
46		0	0.00	– 13
47		0	0.00	– 12
48		0	0.00	– 11
49	5	0	0.00	– 10
50		1	0.02	– 9
51		4	0.07	– 8
52		14	0.25	– 7
53	6	43	0.77	– 6
54		109	1.96	– 5
55		235	4.22	– 4
56		426	7.65	– 3
57		652	11.72	– 2
58	7	841	15.11	– 1
59		915	16.46*	0
60		841	15.11	1
61		652	11.72	2
62		426	7.65	3
63	8	235	4.22	4
64		109	1.96	5
65		43	0.77	6
66		14	0.25	7
67		4	0.07	8
68		1	0.02	9
69		0	0.00	10
70	9	0	0.00	11
71		0	0.00	12
72		0	0.00	13
73		0	0.00	14
Total		5,565	100.00	14

* Rounded up to ensure that the percentages, as shown, add to 100%

As can be seen, the re-marks range from 50 to 68, and the third column shows the simulated number of scripts, drawn from the 5,565 scripts originally marked 59, given each particular re-mark; the fourth column shows the corresponding percentages. The third and fourth columns also show that the distribution of re-marks is symmetrical on either side of the original mark, 59; this is a consequence of the use of a (symmetrical) Gaussian function to simulate the distribution of re-marks..

If, for this particular examination, grade 7 is broad – say, a range of 20 marks from 50 to 69 inclusive – then each of the 5,565 scripts originally given 59 marks is awarded grade 7. But since all the re-marks – which are all within the range from 50 to 68 marks – are still within grade 7, the original grades of all 5,565 scripts would be confirmed. The grades corresponding to all scripts originally marked 59 would therefore all be 100% reliable.

Suppose, however, that the grade boundaries are such that:

- grade 9 – all marks from 70 to 100 inclusive
- grade 8 – 63 to 69 inclusive
- grade 7 – 58 to 62 inclusive
- grade 6 – 53 to 57 inclusive
- grade 5 – 49 to 52 inclusive
- grade 4 – 45 to 48 inclusive.

Reference to Table 1 will verify that the 5,565 candidates, all of whom were originally given 59 marks and all originally awarded grade 7, are re-regraded such that:

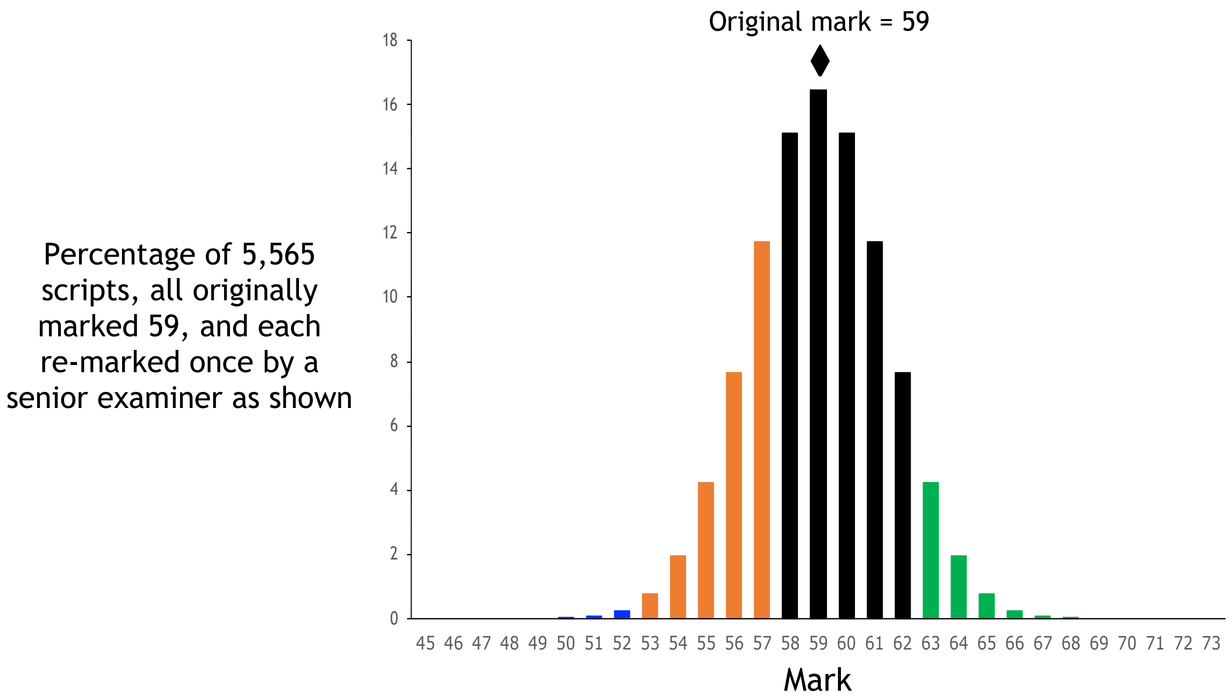
- 406 candidates (about 7.3% of the total) are up-graded to grade 8
- 3,675 candidates (about 66.1% of the total) have the originally-awarded grade 7 confirmed
- 1,465 candidates (about 26.3%) are down-graded to grade 6 (about 3%)
- 19 candidates (abut 0.3%) are down-graded to grade 5.

Overall, of the 5,565 candidates originally given 59 marks and awarded grade 7, 3,675 candidates (about 66%) have their grades confirmed, and 1,890 candidates (about 34%) have their grades changed. This implies that, for these grade boundaries, the reliability of the grades associated with an original mark of 59 is about 66%, so determining the point on this examination subject's wiggly line (as exemplified by Figure 5) corresponding to the original mark of 59.

In principle, the process described can be carried out for all other marks, so determining the grade reliability for all marks, but the amount of work required to do this is huge, and totally impracticable – fortunately, as will be shown shortly, there is a much easier approach.

The percentages shown in the fourth column of Table 1 may be represented graphically as illustrated in Figure 8.

Figure 8: Author's simulation of 2018 GCSE Geography, showing the distribution of re-marks, by a senior examiner, for each of 5,565 scripts all originally marked 59, with grade boundaries as in the text.



As discussed on pages 108 to 113 in the Appendix, this distribution is known as the ***special re-mark distribution***, where the adjective 'special' indicates that the re-marks are 'special', in that they have been given by a 'special' person, a senior examiner, whose mark is 'definitive'.

Importantly, the distribution illustrated in Figure 8, which represents the results of re-marking of 5,565 *different* scripts, all originally given the same mark, is different from the distribution of marks resulting from marking a *single* script 5,565 times.

The distribution shown in Figure 8 corresponds to scripts given a specific original mark, in this case 59. As also discussed in the Appendix, the shape of the special re-mark distribution, and in particular its width, is a property of the subject examination, and is likely to be the same, or for practical purposes very similar, for all marks. One measure of the width of this distribution is the overall end-to-end range, as represented by the whiskers in Figures 3 and 4. For the example illustrated in Figure 8, and for which the corresponding numbers are shown in Table 1, this range is from 50 marks to 68 marks, an end-to-end range of 68 – 50 = 18 marks, 9 marks either side of the original mark, 59. It is this measure that has been referred to throughout this paper as the subject's fuzziness. The constancy of this measure for all marks was an important feature of Figures 3 and 4, and some important evidence that this measure is the same for all marks is given by the regularity of the shapes of the 'arches' shown in Figures 5, 6 and 7.

Within the simulation, this value (18 marks) for the fuzziness was used to determine all the re-marks for all original marks, and resulted in an estimate of the whole-cohort average grade reliability for 2018 GCSE Geography as 65%, in agreement with Ofqual's research as shown in Figure 1.

Two key drivers of grade reliability

The example discussed in this section is a particular instance of a more general principle: the reliability of the grades originally awarded to scripts, all of which were given the same original mark, and all of which have a fair re-mark, depends on two factors:

- Firstly, the shape, and in particular the overall end-to-end width, of the distribution of re-marks, as exemplified by the data shown in Table 1 and depicted in Figure 8. As has been discussed, the shape and the end-to-end width are properties of the examination subject.
- Secondly, the locations of the grade boundaries for the examination subject, these locations being the result of a policy decision, the primary impact of which is to determine the percentage of the subject cohort awarded each grade.

Accordingly, this analysis confirms the inferences drawn on page 24 that, for any subject examination, the narrower the grade widths, the more unreliable the grades, and *vice versa*; similarly, for any grade boundaries, the greater the subject's fuzziness, the more unreliable the grades, and, once again, *vice versa.*

A pragmatic way to measure fuzziness

Finally in this section, an important practical point. The earlier discussion on how to determine the reliability of the grades associated with the mark of 59 described a process in which all 5,565 scripts originally marked 59 were fairly re-marked by a senior examiner. In principle, the reliability of the grades associated with all other marks could be determined by repeating the process accordingly. This implies that the entire subject cohort is double marked, firstly by an ordinary examiner, and secondly, by a senior examiner. Such a process is, of course, wholly impracticable: not only is it a prodigious amount of work, but also half of that work is unnecessary – if all the scripts are to be re-marked by a senior examiner, and if that mark is, by definition 'definitive', then a senior examiner might as well mark the scripts first, and there is no need for any scripts to be marked by any ordinary examiners at all.

Fortunately, as will be described on pages 56 and 57, there is an easier, much more pragmatic, approach – an approach that relies upon the property of the re-mark distribution that is has the same shape, and importantly the same end-to-end width, for all marks[2].

The significance of this is that it implies that for any examination subject, the shape of the distribution needs to be assessed just once (or perhaps a very few times to verify the results) – for example, by choosing scripts which have a particular mark, and then fairly re-marking each of those scripts. It is not necessary for all the scripts given the same original mark to be re-marked, for a suitably selected statistically meaningful sample will do; also, the re-marking does not have to be carried out by a senior examiner – indeed, there is a strong argument that it is preferable for the re-marking to be carried out by randomly chosen ordinary examiners. Further details will be found on pages 56 and 57.

Collectively, the three concepts of determining the shape of the re-mark distribution for just a few original marks, of using statistically valid samples, and of carrying out the re-marking by ordinary examiners – as opposed to double marking the entire subject cohort by a senior examiner – makes the process of measuring grade reliability, in practice, feasible.

Two examples – simulations of 2018 A level Biology and 2018 GCSE English Language

This section presents the results of the author's computer simulations of 2018 A level Biology, and 2018 GCSE English Language, graded 9, 8, 7..., 2, 1, U. The key results are the calculation, for each subject, of the grade reliabilities by mark, reproducing Ofqual's figures for the whole-cohort average grade reliabilities of 85% for Biology, and 61% for English Language, as shown in Figure 1. Details of the simulations are available on request.

The simulations use the published values of the cohort sizes – 63,819 candidates for 2018 A level Biology and 683,838 for 2018 GCSE English Language – and assume that the distributions of marks, standardised on a scale from 0 to 100, are Gaussian. For each subject, the mean and standard deviation of this distribution, in combination with the locations of the grade boundaries, are determined within the simulation so that the percentage of candidates awarded each grade is as close to the published value as possible. The actual values, and simulated values, for 2018 A level Biology are shown in Table 2.

[2] Not quite: for very low marks, and very high marks, the end-to-end width is narrower, but as discussed on pages 56 and 57, for pragmatic operational purposes, these 'outliers' may safely be ignored.

Table 2: Author's simulation of 2018 A Level Biology

Actual cohort size: 63,819

Simulated mean mark: 50.5

Standard deviation of simulated normal distribution of marks: 14.2

Standard deviation of simulated normal distribution of re-marks: 2.57

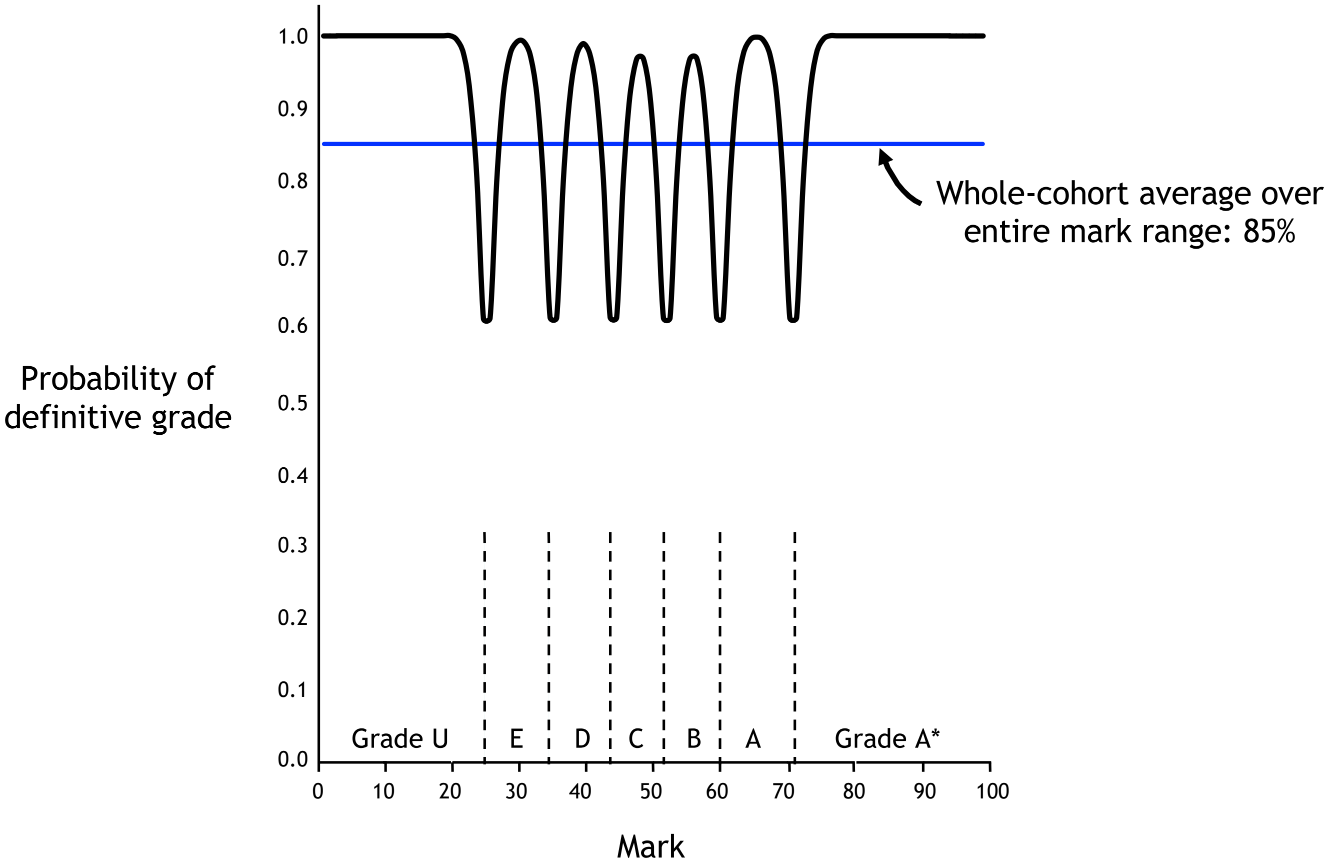
Grade	Grade boundary		Actual population		Simulated population	
	Lower	Upper	Number	%	Number	%
A*	71	100	4,850	7.6	5,058	7.9
A	60	70	11,679	18.3	11,723	18.4
B	52	59	13,785	21.6	13,334	21.0*
C	44	51	14,231	22.3	13,861	21.7
D	35	36	11,105	17.4	11,563	18.1
E	25	26	5,999	9.4	6,152	9.6
U	0	26	2,170	3.4	2,128	3.3
Total			63,819	100.0	63,819	100.0

* Rounded up to ensure that the percentages, as shown, add to 100%

The simulation is by no means perfect, but it is quite close.

The special re-mark distribution for each mark is simulated as a Gaussian, of mean the mark in question, and of standard deviation 2.57 for all marks, this being the mathematical equivalent of the qualitative statement that fuzziness is a property of the examination, and has the same value for all marks. The resulting values of grade reliability by mark are shown in Figure 9: as can be seen, the whole-cohort average grade reliability is 85%, in agreement with Ofqual's findings.

Figure 9: Author's simulation of the grade reliability by mark for 2018 GCSE Biology

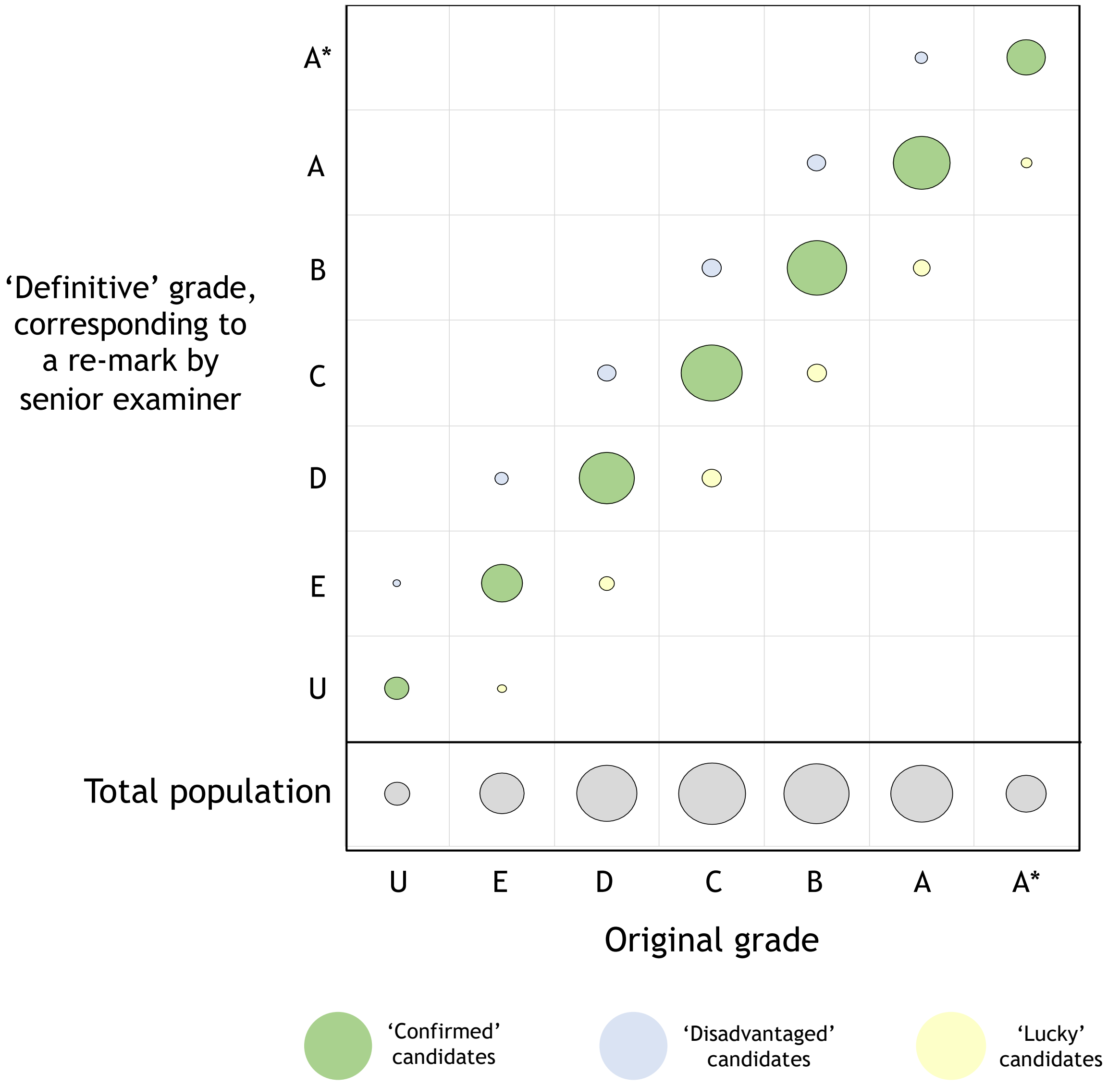


Although the style of this chart is somewhat different from that of Ofqual's charts (compare Figures 5, 6 and 7), the messages are the same – in particular, the 'dip' at all the grade boundaries, and the relationship between grade width and grade reliability (grades B and C, which have the same grade width, are the narrowest, and have the lowest reliability).

Figure 9 also shows that the 'arches' resulting from the simulation have regular symmetrical shapes, very similar to those shown in Figures 5, 6 and 7, which reproduce charts from Ofqual's research using actual marks. Since the simulation assumes that the measure of fuzziness is the same for all marks, the regularity of these shapes across the whole mark range is evidence that, in reality, the measure of fuzziness is a property of the examination subject, and not of the mark or the individual candidate, as discussed on pages 20 and 24. Furthermore, all the simulated and actual 'arches' are left-right symmetrical. For the simulation, this symmetry is the result of using a symmetrical Gaussian to define the underlying distribution of re-marks; the fact that the actual 'arches' show the same symmetry is evidence that the underlying actual distribution of re-marks is also symmetrical, as discussed on page 24.

Another, rather different, visualisation of the simulated data is shown in Figure 10.

Figure 10: Bubble chart for the author's simulation of 2018 A level Biology



Each grey 'bubble' along the bottom row represents the simulated number of candidates originally awarded the corresponding grade, as shown in the sixth column of Table 2, where the area of each bubble is proportional to the corresponding population. When the entire cohort's scripts are fairly re-marked by a senior examiner, each script is given a second mark, corresponding to a second grade. The results of the simulation of this re-marking are shown in each column.

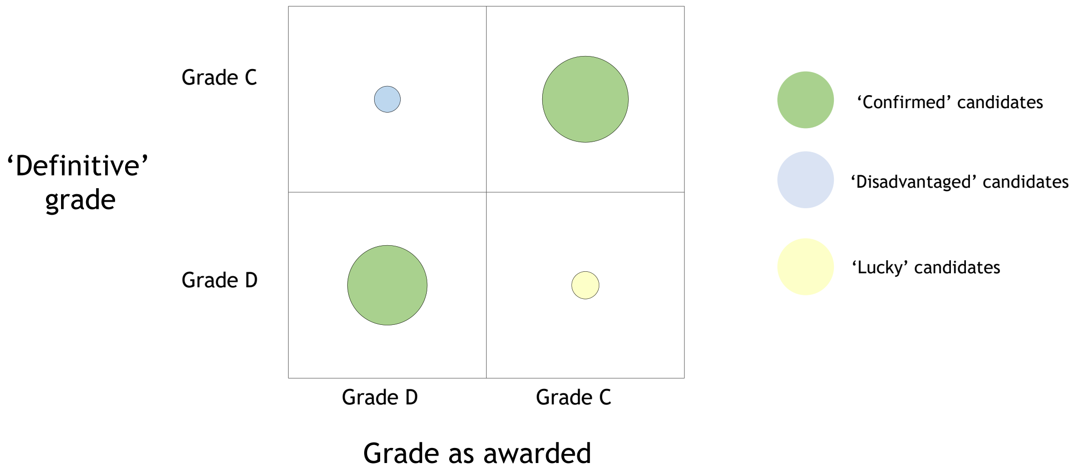
Taking the candidates originally awarded grade C as an example, most have that grade confirmed, as shown by the green bubble at the intersection of the grade C column (original grade) and the grade C row (grade following re-mark). Some candidates, however, are up-graded to grade B (as shown by the blue bubble). All these up-graded candidates may be regarded as 'disadvantaged' in that they were originally awarded a grade lower than the grade that would have been awarded had their scripts been originally marked by a senior examiner, whose mark, by definition, is 'definitive'.

Furthermore, as a result of the senior examiner's re-mark, some candidates originally awarded grade C are down-graded to grade D, (as shown by the yellow bubble). All these candidates may be regarded as 'lucky', in that their original grade is higher than that resulting from a re-mark.

In Figure 10, of the total of 63,819 candidates, the simulation computed that 54,237 candidates[3] have their original grades confirmed, this being 85% of the cohort, in accordance with the overall average grade reliability for Biology as shown in Figure 1. The total number of 'disadvantaged' candidates was simulated as 4,760, 7.4% of the cohort; the simulated total number of 'lucky' candidates is very similar, 4,822, 7.6% of the cohort.

Figure 11 shows a 'close-up' view of the grade D/grade C boundary for 2018 A level Biology.

Figure 11: The D/C grade boundary: author's simulation of 2018 A level Biology



[3] The number 63,819, is precise, and is the actual cohort for 2018 A level Biology. In contrast, the number 54,237, the number of candidates whose grades are confirmed after a re-mark by a senior examiner, is not real, for no such re-marks have actually happened. Rather, it is the result of a computer simulation of what is likely to have happened, had all the scripts been re-marked. '54,237' implies an unwarranted precision: 'about 54,000' is more sensible. Within the simulation, however, all the candidates need to be accounted for, and so the simulation calculates precise numbers, as reported here. Full details are available from the author on request.

The lower-left quadrant represents the simulated number (9,761) of candidates who were originally awarded grade D, and whose grade is confirmed after a re-mark by a senior examiner. The upper-right quadrant shows the simulated number (11,432) whose original grade C is also confirmed. The upper-left quadrant represents the simulated number (1,069) of 'disadvantaged' candidates who were originally awarded grade D, but who would have been awarded grade C had their scripts been marked by a senior examiner. The lower-right quadrant, the simulated number of 'lucky' candidates (1,163) who were originally awarded grade C, but who would have been awarded grade D. Similar diagrams apply to all grade boundaries.

For a totally reliable subject, for which all original grades are confirmed, a chart of the form of Figure 10 would show an upward-sloping diagonal line of green bubbles, each of a size identical to that of the grey bubble at the base of the corresponding column, and there would be no blue or yellow bubbles. For a subject of high reliability, such as Biology, each column shows a large green bubble, accompanied by one small blue bubble in the grade above (corresponding to up-grades), and one small yellow bubble in the grade below (down-grades), as exemplified by Figure 10. As the subject reliability decreases, the green bubbles become progressively smaller, and the blue and yellow bubbles progressively larger. For the most unreliable subjects, small further bubbles can appear two or more grades above and below any green bubble, corresponding to grade changes of more than one grade. In terms of the representation shown in Figure 4, this corresponds to very fuzzy marks, for which the whisker on either side of the given mark X straddles more than two grade boundaries.

An example of a bubble chart for a fuzzier subject is illustrated in Figure 12, which shows the results of the author's simulation of 2018 GCSE English Language, which, in accordance with Ofqual's research as reported in Figure 1, has an average grade reliability of 61%. The corresponding grade reliability by mark is shown in Figure 13, and data relating to the simulation is given in Table 3.

Figure 12: Bubble chart for the author's simulation of 2018 GCSE English Language

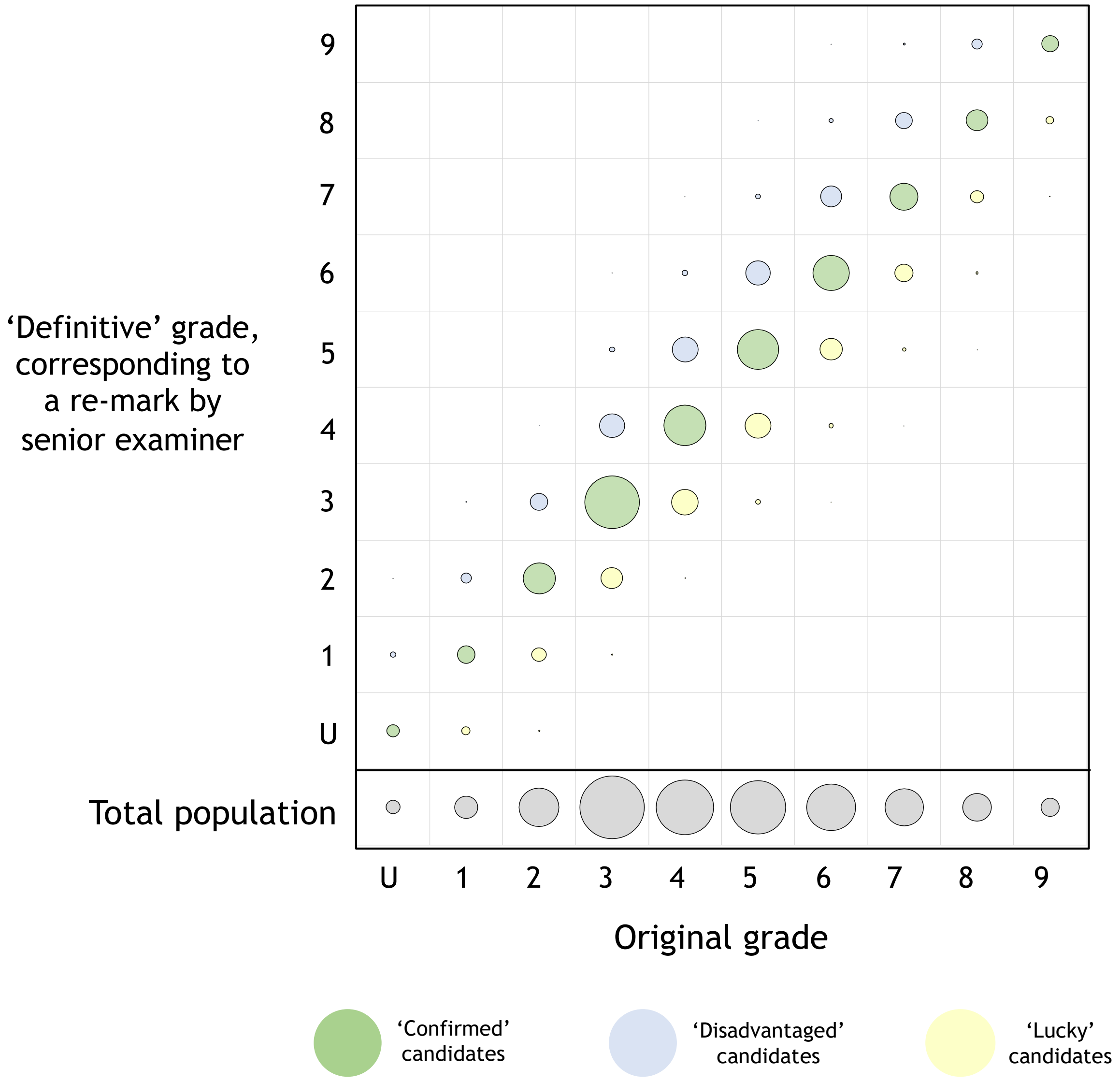


Figure 13: Author's simulation of the grade reliability by mark for 2018 GCSE English Language

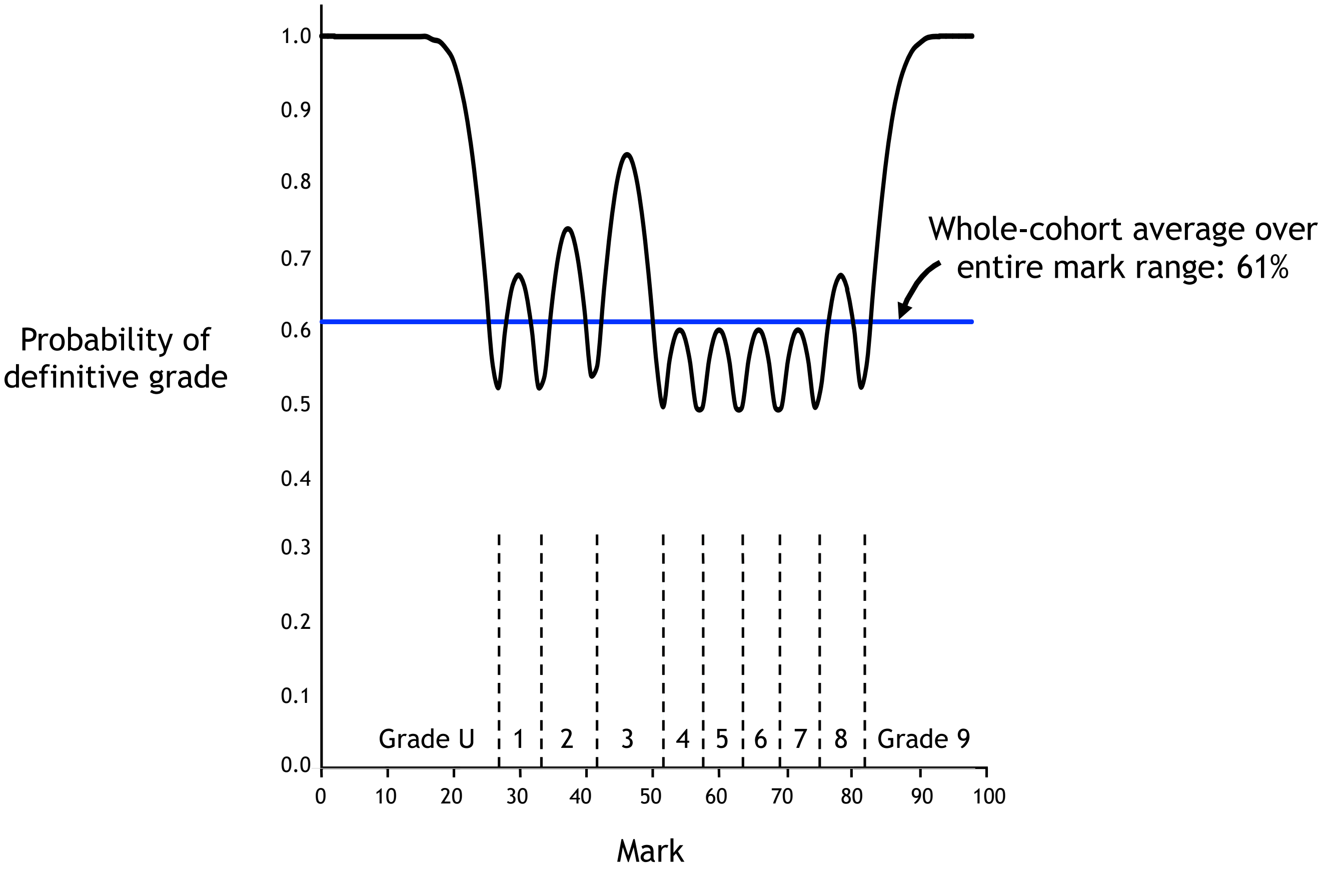


Table 3: Author’s simulation of 2018 GCSE English Language

Actual cohort size: 683,838

Simulated mean mark: 56.0

Standard deviation of simulated normal distribution of marks: 13.0

Standard deviation of simulated normal distribution of re-marks: 5.06

Grade	Grade boundary		Actual population		Simulated population	
	Lower	Upper	Number	%	Number	%
9	83	100	13,677	2.0	13,969	2.0
8	76	82	28,721	4.2	31,489	4.6
7	70	75	53,339	7.8	56,577	8.3
6	64	69	90,950	13.3	90,623	13.3
5	58	63	118,304	17.3	117,734	17.2
4	52	57	118,988	17.4	124,056	18.1
3	42	51	168,225	24.6	158,904	23.2*
2	34	41	62,229	9.1	61,958	9.1
1	27	33	21,883	3.2	20,590	3.0
U	0	26	7,522	1.1	7,938	1.2
Total			683,838	100.0	683,838	100.0

* Rounded down to ensure that the percentages, as shown, add to 100%

Figure 13 shows all the now-familiar features of a ‘wiggle chart’, somewhat accentuated by the combination of a relatively fuzzy subject (English Language is towards the bottom of the chart shown in Figure 1), and narrower grade widths (for GCSE graded 9, 8, 7..., the eight grades 1 to 8 span about the same mark range as the five A level grades A to E, implying that the average GCSE intermediate grade width is necessarily narrower than the average A level intermediate grade width).

The (un)reliability of grades: solutions

Some suggested solutions

Strategies for eliminating fuzziness

The evidence presented so far can be summarised very simply:

- School examination grades are, in general, unreliable.
- The degree of (un)reliability depends on the subject.
- Fundamentally, unreliability is caused by the inevitable fuzziness of marking...
- ...such that the fuzzier the subject, the more unreliable the grades...
- ...and, conversely, the less fuzzy the subject, the more reliable the grades.

Accordingly, any strategy that will eliminate fuzziness will result in more reliable grades. Given that fuzziness is attributable to the different marks given to the same script by different examiners, this suggests two different policy approaches:

- Policies by which examiners could become 'of the same mind'.
- Policies by which the examinations are designed so as to reduce the likelihood that different examiners might give different marks.

The former focuses on the examiner; the latter on the examination. As regards the former, the most obvious solution is for there to be only one examiner for each subject. This might be feasible for a subject with a cohort of only a very small number of candidates, such as some of the modern foreign languages, but for subjects taken by tens, if not hundreds, of thousands of candidates, this solution is impossible – as is the variant solution of a small team.

When the number of examiners becomes larger, the emphasis shifts to recruitment, training and quality control. Yet even if these were perfect, would the result be that each of, say, 100 examiners would all give the same mark to every History essay? Surely not. Certainly, policies that ensure that only suitably well-qualified people are appointed, that everyone is well-trained, and that quality control is vigilant, are to be applauded, and can always be even more rigorous. But they can never achieve the desired result of ensuring that all examiners will always give the same script precisely the same mark. This condition is both stringent and absolute – an apparently trivial difference of just one mark can make all the difference to the candidate: if the C/B grade boundary is 60, then a mark of 59 results in grade C but of 60 results in grade B.

A different policy approach is to change the structure of the examination so that different examiners just cannot give the same script different marks.

This eliminates fuzziness, and ensures that grades will be fully reliable. One way of achieving this to change the structure of examinations from questions inviting candidates to write free-form essays, to those requiring the identification of the right answers from sets of, say, four choices. Multiple-choice tests solve the grade reliability problem, and are used in various contexts around the world, notably the SAT tests used for college admissions in the USA, and the theory test used in the UK to qualify to drive a car. Multiple-choice examinations also have the benefit of being able to be marked by computers rather than by people, and by being taken by the candidate on-line rather than on paper – making the whole process much cheaper to operate.

It would be quite possible to change GCSE, AS and A level exams to multiple-choice format – but there are many consequences and implications, especially as regards the impact on teaching and learning.

Rather than using multiple choice outright, a more subtle approach is to maintain essay-style questions, but to define, in great detail, the mark scheme – this being the set of guidelines used by the examiners when marking. The more specific the mark scheme, the easier it is for the examiner, for rather than reading an answer for its overall sense and judging the quality of the response, the marking process becomes one of checking compliance of the response to the requirements of the mark scheme. This multiple-choice-by-stealth approach also offers the benefits of providing a defence should the marking be challenged. If, for example, a statement is made in an answer that is not explicitly specified by the mark scheme, then the examiner can claim that ''no mark was given for that statement because is not in the mark scheme – if it was regarded as important, it would have been included'. Even though the candidate's answer might be insightful, no mark is given; one of the dangers of compliance is that it enshrines the concept that the 'right' answer must be 'the answer the specifier of the examination has thought of first', thereby penalising any originality, imagination or creativity that the candidate might demonstrate. Compliance with the mark scheme can also work the other way – if it can be proven that an aspect of an answer does indeed comply with the mark scheme but that the appropriate marks were not given, then this is *prima facie* evidence of 'marking error'.

Fuzziness can indeed be eliminated, and grade reliability achieved, by adopting a policy to change the structure of school examinations to multiple choice, but this would have profound and important consequences. If essay-style examinations are to continue to be used, then ensuring, for example, that examiners are well-trained, and that the examinations are well-structured, are 'good things to do'. But some element of fuzziness will inevitably remain, and grades will continue to be unreliable to some degree. The grade reliability problem would still be present.

Why grades are a good idea

A consequence of the policy that examinations are structured around essay-style questions is that fuzziness is inevitable. That suggests a shift of focus: rather than striving, and continually failing, to eliminate fuzziness, surely it is wiser to accept that fuzziness exists and will always exist, and to adopt a strategy that ensures that fuzziness causes no damage. If this can be done, then the assessment awarded to any student will be fully reliable, even though the mark on which that assessment is based is intrinsically fuzzy.

Perhaps somewhat surprisingly, one solution to the fuzzy mark problem is to award grades. If everyone knows that a mark, say 54, given to a script is not precise, and that the same script, if marked by another examiner might have been marked 53 or 55, then a policy that states that ‘all scripts marked between 50 and 59 inclusive shall be awarded grade C’ is very sensible, for then it does not matter that a particular script might be marked 53, 54 or 55 – all are grade C.

This could well have been the thinking behind the first use of grades, at Yale University in 1785, where students were assessed according to one of four grades: *Optimi*, second *Optimi*, *Inferiores*, and *Perjores*.

But if a mark of 54 might be 53 or 55, then a mark of 59 might be 58 or 60. The possibility that the mark might be 58 is not a problem, for that, like 59, corresponds to grade C. The possibility that the mark might be 60, however, is a problem, for that corresponds to grade B.

At Yale, in 1785, and at many places thereafter, that problem was easily addressed. The team of examiners would assemble to review each border-line script carefully, and then form a wise and considered judgement as to which side of the grade boundary each script most fairly lies. The use of grades, associated with a fair review process for all border-line cases, is a sensible, effective and fair solution to the fundamental problem of fuzzy marks. Importantly, it results in assessments that are reliable – in that the same assessment would result from a fair re-mark (and indeed any number of fair re-marks). Grades are a good solution.

If, as is certainly the case, grades are a good idea, why are school examination grades so unreliable, as the evidence shown in Figure 1 proves so vividly?

The answer lies in that second aspect of the wise use of grades: the necessity for a fair review of all border-line cases. For this to happen, two important conditions must be fulfilled – there must be a sufficient number of suitably-qualified examiners available, and also sufficient time, to allow every border-line script to be reviewed fairly. At Yale in 1785, these conditions were honoured – the number of students whose fuzzy marks straddled a grade boundary would have been very few, and the three (or however many) professors could assemble in a comfortable room for an afternoon and do the job.

But in England, in the summer of 2018, with over one million GCSE scripts straddling grade boundaries, and under the time pressure of having to publish the results on that Thursday in August, it is impossible to carry out the required reviews. No reviews take place, with the consequence that, on average, about one grade in every four is wrong.

The problem is therefore not the use of grades *per se*; rather, it concerns the failure to review scripts whose marks straddle grade boundaries. Which immediately suggests some possible solutions, such as:

- Increase the number of examiners, so that the required number of reviews can be accomplished in the required time.
- Increase the time, so that a smaller number of examiners can do the required work.
- Reduce the number of grade-boundary-straddling scripts to a more manageable level, as can be achieved, for example, by reducing the number of grades, thereby increasing the average grade width, so reducing the likelihood of grade-straddling. A variant on this possibility is to use different grading structures for different subjects, according to each subject's intrinsic fuzziness – so, for example, Mathematics grades might be reliable when using, say, six grades; History, perhaps just three.

Each of these possibilities is, in principle, valid in its own right, and more powerful in combination. Each, however, has consequences, and powerful consequences too – from the cost of increasing the number of examiners (which assumes that the required number are potentially available) to the willingness (or otherwise) of policy-makers to make a U-turn, and reduce the number of grades just after the number of GCSE grades has been increased.

The difficulties associated with these possibilities are immense, if not overwhelming. Are there any other possibilities – possibilities that are more pragmatic?

Yes, there are.

Some other possibilities

These possibilities all have a common theme: a theme that changes the policy defining how the originally-given mark is used to determine the assessment that appears on a candidate's certificate. The current policy is that:

- The assessment on the certificate is a grade.
- The grade is determined by mapping the (assumed precise) mark, as given by the (single) examiner who marked the script (or the aggregate mark

if different marks are given by different examiners to different questions), onto a pre-defined grade scale: a mark of 59 is grade C.

As will be explained in detail shortly, there are many other possibilities, each of which uses a different policy for mapping the original mark onto a grade scale, or a different way of recognising a candidate's achievement on a certificate, other than a grade.

Importantly:

- All these possibilities continue to use the same examination structure, the same teaching, and the same learning, as at present – there is no change to the classroom experience, and so no impact on the day-to-day activities of teachers and learners.
- All these possibilities are pragmatic, and do not require any but the most modest increase in examiner resources (if any), and hence cost; nor will they require any increase in the overall time between the completion of any examination and the publication of the results as at present.
- All these possibilities require a change to the policy for appeals: as will be shown shortly, this change is easy to implement.
- All these possibilities result in the delivery of assessments that are as reliable as we would wish them to be – even though the underlying marks continue to be fuzzy.

There is, however, a word of caution: none of these solutions is 'perfect'; each has implications and consequences, some beneficial, others problematic. It is possible that some might consider (or perhaps portray) the problems associated with any particular solution as insuperable, and therefore reasons why the possibility should be rejected. But if the effort or costs required to overcome the problems are outweighed by the benefits of having reliable grades, then perhaps that solution is viable.

An important task not addressed in this paper is a thorough, professional, evaluation of these possible solutions, and to compare the benefits and associated problems of each against the benefits and associated problems of maintaining the *status quo* – which, though possessing the benefit of familiarity, is by no means problem-free, as Figures 1 and 5 bear vivid witness.

The possible solutions will be discussed shortly: to set the scene, the next section examines the important distinction between 'accuracy' and 'reliability', and the necessity of measuring each examination subject's fuzziness, as is required for all the solutions.

Accuracy and reliability

Accuracy

‘Accuracy’ requires a knowledge of ‘right’. A measure is accurate if it is verifiably correct, and can be confirmed as a truth. In general, this is a very demanding criterion, for what is the ‘truth’? Consider, for example, a 2019 BBC news report, from a very different field, that appeared shortly before this paper was drafted. The report concerned the number of people sleeping rough in England, and contains these words:

> *There were 4,677 people sleeping rough in England in autumn 2018, according to official estimates. The figure represents a slight fall of 74 on 2017, however rises were recorded in London, the Midlands, North-East and Yorkshire and the Humber. Numbers are still up 2,909 since the start of the decade, with charities calling for "fundamental action to tackle the root causes".*

The numbers stated – 4,677 people sleeping rough, 74 fewer than the previous year, 2,909 up over the decade – are quoted very precisely: for example, 4,677 people, rather than 4,676, or 4,678, or ‘around 4,700’ or ‘approximately 5,000’. Furthermore, the precision of these numbers, and the authority with which they are stated (‘*according to official estimates*’), imply that these numbers are accurate, the truth.

But are they? If someone else were to count the number of people sleeping rough in England, would the re-count be the same number, 4,677, precisely? Suppose, for example, that someone who is usually unfortunate enough to be obliged to sleep rough happens to be walking the streets, rather than sleeping, at the time the count was performed. That person should, presumably, be included within the true, accurate, number, but happened to have been missed when the researcher counted 4,677. And if 4,677 is not an accurate measurement, and that a more appropriate statement is ‘the number of people sleeping rough in England has been estimated at about 5,000’, then does the statement that there has been ‘*a slight fall of 74*’ have any meaning?

Accuracy, though desirable, is elusive, and especially so for examination marks and grades: for a particular script’s examination mark, and the corresponding grade, to be accurate, there must be an unambiguous, and trusted, definition of what the right mark is. In practice, this can never be the case.

As has been stated many times, different, equally qualified, equally conscientious, and equally committed examiners can give the same script different marks. As discussed on pages 97 to 100 in the Appendix, these marks form a statistical distribution, known as a ‘panel distribution’, raising questions such as:

- Is the ‘right’ mark the average, the ‘arithmetic mean’ of the distribution?
- Is it the ‘mode’ – the mark given by the greatest number of examiners?
- Is it the ‘median’ – the ‘half-way’ mark, such that half of the examiners give this mark or a smaller mark (and, by the same token, the other half give this mark or a higher mark)?
- Is it the mark given by a ‘special person’, such as a ‘senior examiner’, whose judgement, presumably, is superior to that of all other examiners? And, if there is more than one ‘senior examiner’ (as in practice is the case for the mainstream, large cohort, subjects), what are the consequences of the possibility that even the community of senior examiners might themselves not award the same script the same mark?

These are indeed problems. Problems that suggest one answer. That there is no ‘right’ mark. And that the search for the ‘right’ mark is a search for a chimera, a snark.

That said, to resolve the problem by an agreement that ‘right’ is defined by reference to a measurable statistic (such as the median mark of a defined statistical distribution), or to the mark given by a senior examiner, can be convenient. Indeed, ‘the right mark is that given by a senior examiner’ is the fundamental assumption throughout all Ofqual’s research, as published in the November 2016 and November 2018 reports. But it is an assumption; an assumption that can be useful, but not a truth.

Reliability, and two different statistical distributions

So much for ‘accuracy’; to turn now to ‘reliability’. The test of reliability is a second (or indeed third...) opinion, rather than a search for truth. This paper argues that an examination assessment – for example the grade appearing on a candidate’s certificate – is ‘reliable’ if the originally-awarded assessment is confirmed as the result of a second, fair, re-mark – where ‘fair’ implies that the re-mark is carried out under the same conditions as the original mark, and in the absence of any biases potentially associated with the an examiner’s tendency to be ‘harder’ or ‘softer’ given the knowledge that the script is being marked for a second time, perhaps as the result of an appeal.

Given that any particular original mark may result in a range of possible re-marks, the relationship between any original mark and any re-mark is a matter of statistics. The Appendix explores the statistics of marking and re-marking in some detail; two important results are that:

- There are two, different, re-mark distributions, depending on whether the re-marking is done by a senior examiner or by a second ordinary examiner.
- If the re-marking is done by a senior examiner, the resulting re-mark distribution is *narrower* and *sharper* than the re-mark distribution associated with re-marking by an ordinary examiner, which is *broader* and *flatter*.

This second point has particular relevance as regards how any specific examination subject's fuzziness can be measured and used.

An example of a re-mark distribution measured by reference to a re-mark by a senior examiner was shown as Figure 8: this chart, based on the author's simulation of the results of 2018 GCSE Geography, shows the distribution of the re-marks of all 5,565 scripts originally given 59 marks. As can be seen, nearly 17% of those scripts are re-marked 59, the same as the original mark; and the total range of re-marks is nine marks either way, this being one possible measure of this subject's fuzziness.

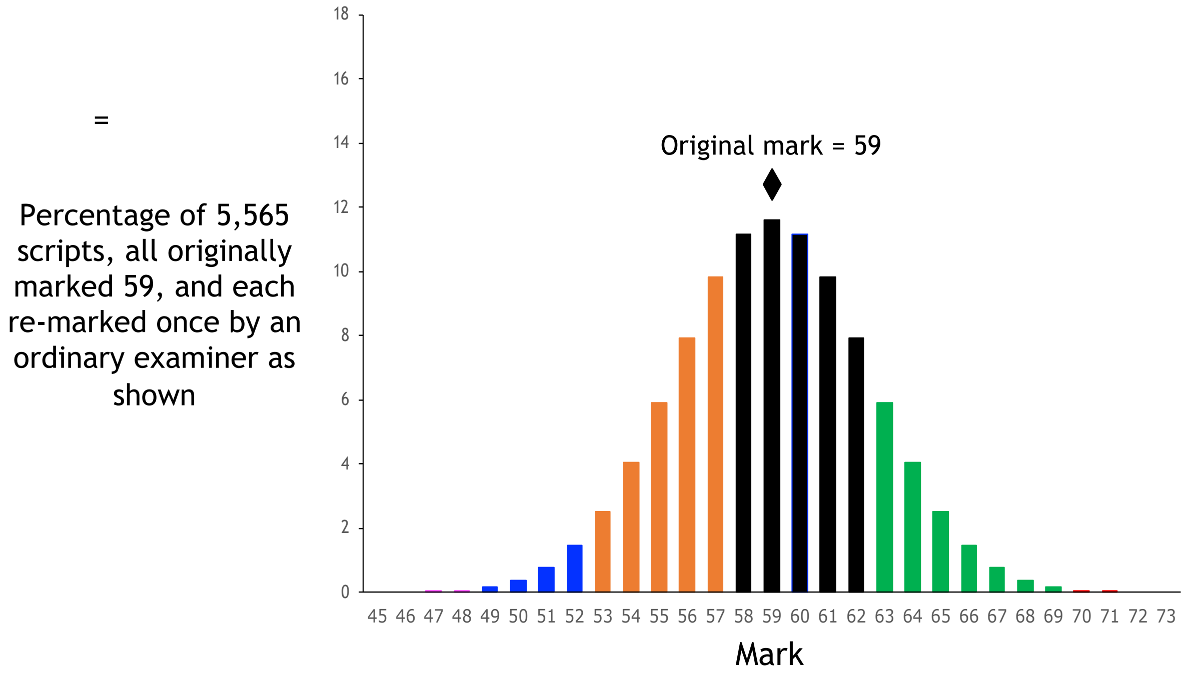
If, however, those same 5,565 scripts, all originally marked 59, are fairly re-marked by an ordinary examiner, the re-mark distribution is as shown in Table 4 and Figure 14.

Table 4: Author's simulation of the fair re-marking, by an ordinary examiner, of 5,565 2018 GCSE Geography scripts, each originally marked 59 and originally awarded grade 7, and re-graded according to the grade boundaries given on page 32

Re-mark	Re-grade	Number of scripts given re-mark	% of scripts given re-mark	Difference between re-mark and original mark
45	4	0	0.00	– 14
46		0	0.00	– 13
47		1	0.02	– 12
48		4	0.07	– 11
49	5	9	0.16	– 10
50		21	0.38	– 9
51		43	0.77	– 8
52		81	1.46	v7
53	6	140	2.52	– 6
54		224	4.03	– 5
55		328	5.89	– 4
56		442	7.94	– 3
57		546	9.81	– 2
58	7	620	11.14	– 1
59		647	11.62*	0
60		620	11.14	1
61		546	9.81	2
62		442	7.94	3
63	8	328	5.89	4
64		224	4.03	5
65		140	2.52	6
66		81	1.46	7
67		43	0.77	8
68		21	0.38	9
69		9	0.16	10
70	9	4	0.07	11
71		1	0.02	12
72		0	0.00	13
73		0	0.00	14
Total		5,565	100.00	

* Rounded down to ensure that the percentages, as shown, add to 100

Figure 14: Author's simulation of the re-mark distribution for re-marks by an ordinary examiner, for 2018 GCSE Geography, using exactly the same data for the original marks, and to the same scale as shown in Figure 8



Figures 8 (an example of a ***special re-mark distribution***) and 14 (an example of the corresponding ***ordinary re-mark distribution***) are based on the same underlying original marks, have the same scales, and are directly comparable. As can be seen from the peak in Figure 14, the percentage of scripts re-marked 59, the same as the original mark, by ordinary examiners is about 12% (as compared to about 17% as shown in Figure 8); also, the end-to-end range of re-marks extends to 12 marks either side of the original mark (as compared to nine marks either side in Figure 8). The distribution shown in Figure 14, corresponding to re-marking by an ordinary examiner, is broader and flatter than the distribution shown in Figure 8, corresponding to re-marking by a senior examiner, in accordance with the assertion made on page 50.

The examples shown in Figures 8 and 14 use exactly the same original marks, and demonstrate that, if the end-to-end range of the re-mark distribution is to be used as a measure of an examination subject's fuzziness, then it is important that the basis of the re-mark comparison is known, and, in particular, whether the examiners who carry out the re-mark are ordinary or senior. Since the examination's fuzziness correlates to that subject's grade reliability, the way in which a subject's fuzziness is measured has an impact on the corresponding measure of grade reliability, as shown in Table 5.

Table 5: Grade reliabilities as measured by reference to a re-mark by a senior examiner (in accordance with Ofqual's results as shown in Figure 1), and to an ordinary examiner (estimated by the author's simulations, available on request)

Subject	Reliability as measured by reference to a senior examiner	Estimate of reliability as measured by reference to an ordinary examiner
Mathematics	96%	~ 92%
Chemistry	92%	~ 88%
Physics	88%	~ 83%
Biology	85%	~ 78%
Psychology	78%	~ 69%
Economics	74%	~ 64%
Religious Studies	66%	~ 55%
Business Studies	66%	~ 55%
Geography	65%	~ 53%
Sociology	63%	~ 51%
English Language	61%	~ 49%
English Literature	58%	~ 47%
History	56%	~ 45%
Combined English Language and Literature	52%	~ 41%
Cohort-weighted average	75%	~ 66%

Sources: Senior examiner, Ofqual data as shown in Figure 1; ordinary examiner, author's simulations.

The second column in this table shows estimates of Ofqual's numerical values for the average grade reliability of each subject, as inferred from Figure 1. The third column shows the results of the author's computer simulations. For each subject, the simulation computed firstly a distribution of original marks, and then two measures of the average grade reliability: the first by reference to a re-mark by a senior examiner, and the second by reference and to a re-mark by an ordinary examiner. This enabled corresponding values of the two measures of reliability to be compared. As an example, as shown in Table 5, Ofqual's measurement, by reference to a re-mark by a senior examiner, of the average grade reliability for Economics is about 74%. The author simulated this result, and also the grade reliability by reference to re-marking by an ordinary examiner, resulting in an estimate of about 64% (better expressed as likely to be within the range 62% to 66%).

The mathematics and statistics of this are presented in the Appendix, and the simulations are available on request; the important conclusion is that grade reliabilities as measured by reference to a senior examiner are consistently higher numbers than as measured by reference to an ordinary examiner.

This difference between referencing a senior examiner and an ordinary examiner has especial significance in a particularly important context.

So far, the discussion has been based on a comparison between an original mark, and a subsequent re-mark, either by a senior examiner (resulting in a re-mark distribution such as that shown in Figure 8) or by an ordinary examiner (as in Figure 14). To take a concrete example, suppose that, for the examination represented in Figures 8 and 14, grade 7 corresponds to all marks from 58 to 62 inclusive, grade 8, all marks from 63 to 69 inclusive. Suppose further that a script is given an original mark of 59, grade 7, and then fairly re-marked.

If the re-mark is given by a senior examiner, it is Figure 8 that is relevant, which shows that there is a probability of about 17% that the re-mark will equal the original mark, 59, so confirming the original grade 7, and a probability of about 4% that the re-mark will be 63, resulting in an up-grade to grade 8. Any re-mark of 63 or above will result in an up-grade, and the probability of this can be estimated by adding the probabilities represented by the columns associated with each of these marks. When this is done, the probability of receiving up an up-grade resulting from a re-mark by a senior examiner is about 7%.

But if the script is re-marked by an ordinary examiner, it is Figure 14 that is relevant, and the sum of all the columns corresponding to marks equal to, or greater than, 63 is about 15%, including the probability of about 0.1% that the re-mark is 70 or 71, resulting in an up-grade jumping two grades, from 7 to 9. And although ‘only about 15%’ might by some be regarded as a small proportion, for a cohort of 5,565 candidates originally marked 59, ‘only about 15%’ represents 851 individual students, to each of whom the difference between a grade 7 and a grade 8 or a grade 9 could be very important indeed.

The interpretation of Figures 8 and 14 in terms of original marks and subsequent re-marks is, as has been discussed, important. Even more important, however, is a different interpretation – an interpretation addressing the question ‘Suppose a script had originally been marked not by the examiner who actually marked the script, but by someone else? Would the same grade have been awarded?’. This question directly addresses the lottery-of-the-first-mark, first mentioned on page 11, and is the most stringent test of grade reliability, for if there is a probability of 100% that the same grade will be awarded no matter which examiner marks a script, then that grade is truly reliable.

This question concerns a mark/re-mark pair; not quite as formally as an original mark followed by a fair re-mark resulting from an appeal, but a comparison for the same script nonetheless. The two circumstances are in principle identical, as are the corresponding mathematics and statistics. But since an original mark is far more likely to be given to a particular script by an ordinary examiner than a senior one, it is Figure 14, rather than Figure 8, that is more likely to correspond to reality.

A pragmatic way to measure fuzziness

Measuring the fuzziness of any examination subject is an essential requirement for awarding assessments that are more reliable than those currently awarded. Here is a suggestion for a pragmatic way to measure do this:

- Mark all the subject cohort scripts in the usual way.
- From all the scripts given the same mark (say, 59), randomly select, say, 100 scripts.
- Give each of these 100 scripts, all of which have the same original mark, to a randomly selected second examiner (necessarily different from the original examiner) who then fairly re-marks the script.
- This results in a 100 mark/re-mark pairs for scripts originally marked 59, so enabling the re-mark distribution to be determined.
- Measure the fuzziness as, for example, the end-to-end range of this distribution (possibly excluding any anomalous outliers).

Some further details:

- The choice of 100 as the number of scripts to be re-marked is illustrative only: the actual number to be used in practice should be determined so as to result in a statistically valid sample.
- Ideally, each of the 100 (or however many) scripts selected for re-marking should be re-marked by a different examiner. That assumes that there are at least as many examiners as there are scripts in the sample randomly chosen for re-marking. If the number of scripts to be re-marked is greater than the number of available examiners, then some examiners will need to re-mark more than one script.
- If all the re-marking is carried out by senior examiners, the result is a special re-mark distribution as represented by Figure 8; if by ordinary examiners, the ordinary re-mark distribution as represented by Figure 14. In general, the ordinary distribution is much more meaningful, and, in practice, much easier to carry out for there are many more ordinary examiners than senior ones. The ordinary distribution is therefore the preferred choice.
- As will be discussed in more detail on pages 78 to 82, the measure of the distribution’s width (for example, the end-to-end range), and also matters such as the determination of what is, and what is not, an

‘outlier’, are policy decisions that ultimately determine the resulting grade reliability: in general, the greater this measure, the more the grade reliability approaches 100% for all marks.

An important feature of this process is that it is operationally pragmatic, and the dependence on statistical analysis is *de minimis*: the only requirement for statistics is in determining the size of the samples to be used for re-marking, and in measuring the width of the resulting distribution. Neither of these require knowledge of, or depend on, the technical details examined in detail in the Appendix; nor indeed do these details need to be right – for example, an assumption that the distribution of the original marks is symmetrical, or follows a particular mathematical form such as a Gaussian. Pragmatically, if a sample of scripts all originally given the same mark are each individually fairly re-marked, the result will be an empirically-determined distribution, and that distribution will have a measure of width, such as the end-to-end range. The definition of the specific measure of width in any particular circumstances is a matter of policy; from a pragmatic standpoint, once that policy has been determined, the width can easily be measured.

As discussed on page 35, it is important that the measure of fuzziness is a property of the subject examination, and not of the mark, still less the individual script. This needs to be validated by an appropriate statistical study, but even if this assertion is, from a practical standpoint, valid, any single measurement of fuzziness is just a single measurement, and subject to measurement errors. It is therefore likely that, in practice, the process described would be carried out for several sets of scripts, for example, 100 scripts each originally marked 59, 100 scripts marked 74; 100 marked 67; and 100 marked 43. This will result in four measures of the examination subject’s fuzziness, from which an average measure can be determined.

The overall outcome is a measure of fuzziness, for example, the end-to-end range of the re-mark distribution for that examination subject. In the specific example of the simulation of 2018 GCSE Geography shown in Figure 14, this range extends 12 marks either side of the centre, implying that the total fuzziness of 24 marks – nearly one-quarter of the entire mark range from 0 to 100.

For the remainder of this paper, the measure of fuzziness that will be used will the end-to-end range of either the special re-mark distribution (for comparison with Ofqual’s results), or the ordinary re-mark distribution (for more general discussions). As discussed an page 121 in the Appendix, these distributions are very likely to be symmetrical about the original mark, and so the end-to-end range can be represented by the symbol $2f$. For the special re-mark distribution shown in Figure 8, the end-to-end range is 18 marks, implying that $2f = 18$ and so $f = 9$ marks; for the corresponding ordinary re-mark distribution shown in Figure 14, $2f = 24$ and so $f = 12$ marks.

Some possible solutions

An important policy choice

In this section, some different, pragmatic, solutions will be introduced, all of which result in assessments, as would appear on a certificate, that are more (and in principle much more) reliable than the current grades. There may be further solutions not mentioned here, so an important activity is to identify as many solutions as possible, so enabling the best to be selected.

None of the suggested solutions is ‘perfect’, in that none deliver accuracy. However, as discussed on pages 49 and 50, accuracy is impossible to achieve, and so the failure to deliver the impossible can hardly be a drawback. That said, all the suggested solutions have implications and consequences, of which some are likely to be considered beneficial, others more problematic. The following sections explore some of the more important consequences, but not exhaustively – the intention here is to lay the foundations for a further, more thorough, analysis which will provide comprehensive, insightful and balanced evidence on which a policy decision can be taken.

This policy decision is a choice – a choice between maintaining the current assessment policy, or to replace the current policy by an alternative that is believed to be better. It is therefore important that the current assessment policy is examined as regards its benefits and problems, so permitting a fair comparison.

The baseline – the current policy

Taking the current policy as the baseline for comparison, the key features of the current policy are:

- Each script is marked once, and given a single mark, symbolically represented as m.
- The assessment for each subject, as awarded on that candidate’s certificate, is a grade.
- The candidate’s grade is determined by mapping the mark m onto a pre-determined grade scale.
- If, as the result of an appeal, the original mark m is found to be ‘sound’, and there is no evidence of a marking error, then no re-mark is allowed, and the originally given mark m, and the corresponding grade, must stand.
- If, however, a marking error is identified, the error is corrected, resulting in a re-mark m^*. The candidate’s assessment is then changed from being based on the original mark m to being based on the re-mark m^*, using the same pre-determined grade scale. This may, or may not, result in a grade change, depending on the location of the grade boundaries.

The current policy has two elements: the first defining the rule by which the assessment, as shown on the candidate’s certificate, is determined from the original mark m; the second relating to appeals. As already discussed, the current policy for appeals allows for a re-mark m^* only if there was a marking error associated with the original mark m. The original mark m is therefore deemed to be precise, even though it is known, in reality, to be fuzzy. As will be seen, all the policies that deliver improved reliability recognise that marks are indeed fuzzy, and therefore take fuzziness into account in both the determination of the assessment as shown on the certificate, and also appeals.

With the current policy in mind, several other possible policies can be identified, all of which require measurement of each examination subject's fuzziness, as determined, for example by the end-to-end range of that subject's ordinary re-mark distribution. If that end-to-end range is represented as $2f$ marks, then since, as shown on page 121 in the Appendix, the ordinary re-mark distribution is always symmetrical about the original mark m, the lowest possible re-mark m^* is therefore $m - f$, and the highest, $m + f$. This can be verified by the example shown in Figure 14. The total range of marks is from 47 to 71, a range of 71 - 47 = 24 marks = $2f$, implying that f = 12 marks. The originally given mark was 59 = m, from which $m - f$ = 59 - 12 = 47, and $m + f$ = 59 + 12 = 71.

As expressed in terms of the parameters m and f, the possible policy solutions to be discussed are:

- Three grades, as determined by $m - f$, m, and $m + f$.
- Two grades, as determined by m and $m + f$; $m - f$ and m; or $m - f$ and $m + f$.
- One grade, as determined by $m + f$.
- One grade, as determined by $m - f$.
- As each of the preceding possibilities, but using 'adjusted' grades of the general form $m + \alpha f$, where α is a parameter that can take any value from -1 to $+1$ (including 0), as determined by policy.
- Do not award grades, but show on the certificate the mark m and also the subject examination's measurement of f.

All of these possibilities require measurement of the examination's fuzziness, as represented by the parameter f, which, in this paper, is assumed to be one-half of the total end-to-end range of the (ideally ordinary) re-mark distribution. Only in the last solution, however, is this measurement explicitly shown on the candidate's certificate; for all the other solutions, its value is used to determine the assessment, but is otherwise 'hidden'. Furthermore, as will be explained on pages 61 to 64, all these possibilities require a new policy for appeals - a policy which includes the suggestion that the fee currently levied should be abolished, so removing this significant barrier.

For completeness, two solutions, discussed earlier, are:

- Multiple-choice examinations (see page 44 and 45).

- As the current policy, but with grade boundaries and grade widths determined according to the subject, with a view to minimising grade unreliability (see page 47).

These two possibilities will not be examined further here, but should be included in a more comprehensive study. A further possibility – to double mark every script and use, for example, the average mark to determine the candidate's grade – is also not discussed: as shown on pages 122 to 125 in the Appendix, despite the widespread belief that 'two brains are better than one', double marking is not an effective solution.

The following sections now explore the suggested policy solutions in more detail.

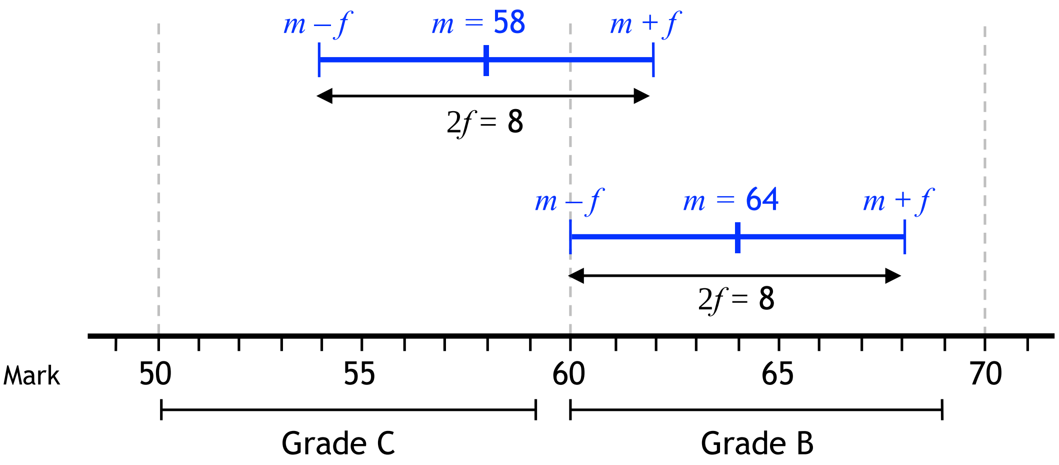
Three grades

The three grade solution

The 'three grade' solution is perhaps the most obvious, and certainly the easiest to explain, especially in the context of Figures 3 and 4, which show four different original marks (as indicated in the figures by each X), and each mark's corresponding fuzziness, as represented by the whiskers, which are symmetrical on either side.

Currently, the grade on the candidate's certificate is that corresponding to the original mark m. This solution proposes that, in addition, the grades corresponding to the two whiskers are also declared on the certificate, so indicating the range of marks, and the corresponding grades, that might have been awarded had the script been originally marked by a different examiner. In terms of measuring the end-to-end extent of the distribution of re-mark as $2f$ marks, this range extends from a lowest possible mark of $m - f$ to a highest possible mark of $m + f$. The certificate therefore shows the grades corresponding to each of the marks $m - f$, m and $m + f$, as determined by the mapping of those marks onto the pre-determined grade scale.

Figure 15: The three-grade solution: C, C, B and B, B, B



Accordingly, a certificate that shows C, C, B indicates a 'high C' that straddles (to an unspecified extent) the C/B grade boundary, as illustrated in Figure 15 for a script for which m = 58 marks and $2f$ = 8 marks. B, B, B is a 'solid' B, and is a fully reliable grade, corresponding to a subject for which the fuzziness for the candidate's mark lies fully within the B grade width, as also illustrated in Figure 15 for m = 64 marks. A certificate showing D, C, B probably says more about the subject than the candidate, indicating that the fuzziness associated with the given mark straddles two grade boundaries, as shown in Figure 4. A candidate's certificate showing E, C, A is alarming but quite real for the fuzziest subjects, such as History (and possibly some others, not included in the 14 subjects shown in Figure 1).

A new policy for appeals

Under the current policy, a script may be re-marked only if it can be demonstrated that there was a marking error associated with the original mark, such as the failure to comply with the mark scheme or an administrative error. If this is the case, the marking error is corrected, and the re-mark m^* takes precedence over the original mark m. If the re-mark m^* is within the same grade width as the original mark m, the original grade is confirmed; if the re-mark m^* lies on the other side of a grade boundary, the grade is changed accordingly.

Under all of the new policies, the policy for appeals is different, and in the author's opinion both fairer to all candidates, and also more effective. Importantly, all the new policies remove any barriers to appeal, including the fee.

Furthermore, all the new policies recognise the possibility that marking errors can occur. Accordingly, in all cases, if a marking error is detected, and then corrected, the resulting re-mark m^* replaces the original mark m as the basis of the candidate's assessment, as happens under the current policy. But since the new policies take fuzziness explicitly into account in determining the candidate's assessment, fuzziness also needs to be taken into account for appeals, even when there are no marking errors.

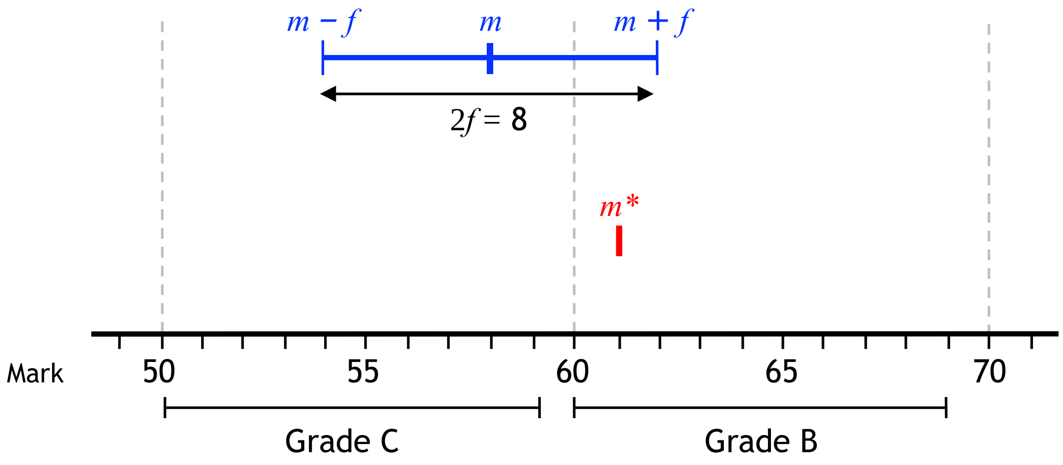
Under the three-grade policy, the three grades are determined by each of the three marks $m - f$, m and $m + f$ respectively. It might therefore be thought that, if the script is re-marked m^*, than the revised assessment should be based on the three marks $m^* - f$, m^* and $m^* + f$.

To do this, however, is problematic, for the policy for awarding the original assessment has anticipated the possibility (and indeed probability) that a re-mark m^* would be different from the original mark m, and within the range from $m - f$ to $m + f$. That a re-mark m^* is likely to be different from the original mark m is therefore not a 'surprise', nor evidence of a marking error; rather, it is simply the expected consequence of the fundamental fact that different examiners can give the same script '*different but appropriate*' marks. Furthermore, the value of f is a statistically valid measure of the likely range that those re-marks might take, and so by recognising f in the determination of both the 'low' grade and the 'high' grade of the three grade trio, all re-marks within the range from $m - f$ to $m + f$ have already been taken into account. If the re-mark m^* is within this range, there is therefore no reason to revise the original assessment based on $m - f$, m and $m + f$ to one based on $m^* - f$, m^* and $m^* + f$.

Accordingly, as noted in the brief descriptions on page 59, for any of the policy solutions which recognise, and incorporate, fuzziness as measured by f, a new policy as regards appeals and re-marks is appropriate, in this case:

- If the re-mark m^* is within the range from $m - f$ to $m + f$, the original assessment is confirmed.
- If the re-mark m^* less than $m - f$, or greater than $m + f$, the assessment is changed to one based on $m^* - f$, m^*, and $m^* + f$, which may or may not result in a grade changes, depending on the location of the grade boundaries.

Figure 16: A new policy for appeals: this re-mark m^* does not result in an up-grade



An example is illustrated in Figure 16. This figure is similar to Figure 15, and represents an examination for which $2f = 6$ marks. Under the current policy, a script marked $m = 58$ would be awarded grade C, and a re-mark $m^* = 61$ would result in an up-grade to grade B.

Under the three-grade policy, the certificate would show C, C, B, so anticipating that it is possible that a fair re-mark will be any mark between $m - f = 58 - 4 = 54$, and $m + f = 59 + 4 = 62$, a range of 62 - 54 = 8 = $2f$ marks. If an appeal is made and the script re-marked $m^* = 61$, this is within the originally anticipated range, and the original assessment C, C, B has already recognised the possibility that the candidate might be awarded grade B. The re-mark $m^* = 61$ is within expectations, and so the original assessment C, C, B is confirmed.

All marks in the range from $m - f$ to $m + f$ have been anticipated, and so any re-mark within this range confirms the originally-awarded assessment. This is the key feature that causes this policy (and the others too) to result in reliable grades. Only if a re-mark m^* is either less than the lower bound $m - f$, or greater than the higher bound $m + f$, is the assessment changed to one based on $m^* - f$, m^*, and $m^* + f$, with the possibility of corresponding grade changes. But if the statistical determination of the value of f has been done correctly, the likelihood that this will happen is very low, and if it does happen, it is more likely to resolve a true marking error in the original mark than an error in the statistics.

Finally, this policy poses no barriers to appeal, for it is important that no candidate is denied access to justice. Because this policy delivers much more reliable grades than the current system, the vast majority of appeals will result in confirmation of the originally-awarded grade, and those appeals that do result in a grade change will be correcting (a small number) of errors in the original marking. This will build trust in the new policy, and in the examination system overall, and the number of appeals will diminish, over the years, of its own accord.

Two grades

The 'two grades' solution is a variant in which the candidate's certificate shows two grades. There are three possibilities, the first being to combine the original mark m with the 'upwards-adjusted' mark $m + f$. Accordingly, for the examples illustrated in Figure 15, the assessments as shown on the candidate's certificate would be C, B for $m = 58$, and B, B for $m = 64$.

A second possibility is to show the grades corresponding to the 'downwards-adjusted' mark $m - f$ and the original mark m (with reference to Figure 15, C, C for $m = 58$, and B, B for $m = 64$); the third is to combine $m - f$ with $m + f$ (C, B for $m = 58$, and B, B for $m = 64$).

The policy for appeals is identical to that for the three grades solution: if a marking error is discovered, the assessment is changed to grades based on the re-mark m^* and the two 'adjusted' re-marks $m^* - f$ and $m^* + f$ as appropriate; if there are no marking errors and the re-mark m^* is within the range from $m - f$ to $m + f$, the original assessment is confirmed; if the re-mark m^* is less than $m - f$ or greater than $m + f$, the assessment is changed, and based on $m^* - f$, m^* and $m^* + f$.

One grade based on $m + f$

The benefit of the doubt

A further variant of three grade solution and the two grade solution is to award just one grade based on $m + f$.

In essence, this gives the candidate the 'benefit of the doubt'. This recognises that the mark as originally given was m, but it might have been a different mark, perhaps higher, perhaps lower. However, for any original mark m, the statistically valid determination of the parameter f implies that it is most unlikely that another examiner would give a mark greater than $m + f$. Basing the grade, as awarded on the certificate, on $m + f$ rather than on the original mark m therefore takes the examination's fuzziness into account by being suitably 'generous' to the candidate.

To give a specific example: consider an examination for which grade C is all marks from 55 to 59 inclusive, and grade B, 60 to 64. Suppose further that the fuzziness of this examination has been measured as $f = 4$ marks. Under the current grading policy, a script marked $m = 58$ would be awarded grade C; under a policy of determining grades based on $m + f = 58 + 2 = 62$, the script would be awarded grade B - as if the script had been marked by the 'most generous' examiner.

As is now familiar, in the absence of marking errors, the new policy for appeals is that the original grade is changed only if the re-mark $m*$ is less than $m - f$ or greater than $m + f$. The result of applying this new policy is illustrated in Figure 10, which shows the author's simulation of 2018 A level Biology, graded according to $m + f$, for $f = 6$, and then re-graded following a re-mark by a senior examiner.

Figure 17: Author's simulation of grades after appeal for 2018 A level Biology, graded according to $m + f$*, for* $f = 6$ *marks, under the new policy*

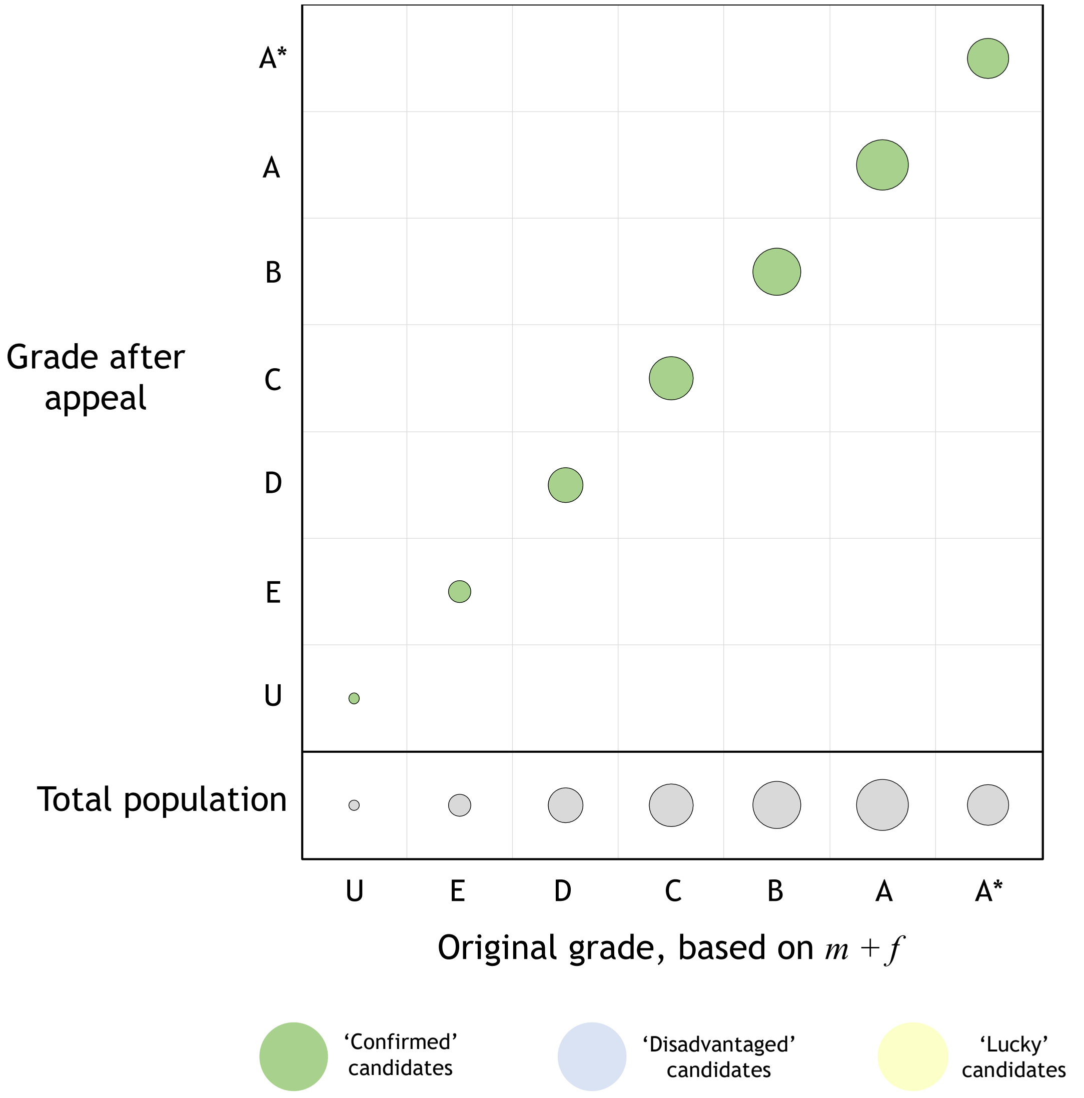
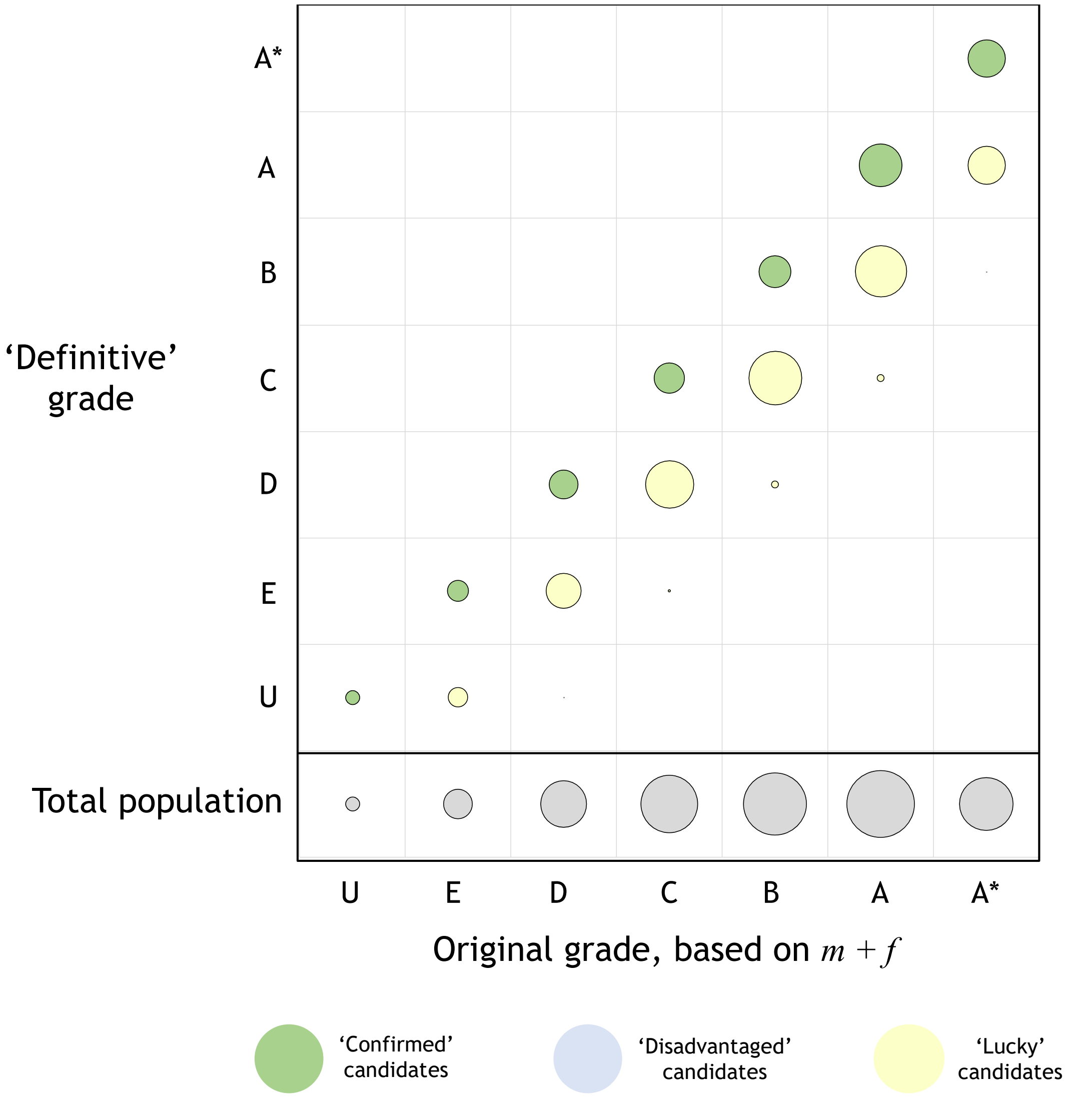


Figure 17 uses the same underlying marks, re-marks and grade boundaries as Figure 10: the only differences are the policies used for awarding the original grade, and for appeals. The effect of these policies is profound. Under the new policies, all the original grades are confirmed. This demonstrates that grading according to $m + f$ delivers reliable grades.

Figure 17, however, masks a subtlety. The vertical axis in Figure 17 represents the outcome of the new appeals policy under which the original grade (as determined by $m + f$) is confirmed if the senior examiner's re-mark m^* lies within the range $m - f$ to $m + f$, but changed otherwise. If, however, the vertical axis represents the 'definitive' grade as determined directly from the senior examiner's re-mark m^*, the result is as shown in Figure 18.

Figure 18: Author's simulation 2018 A level Biology, graded according to $m + f$, for $f = 6$ *marks, compared to the corresponding 'definitive' grades*



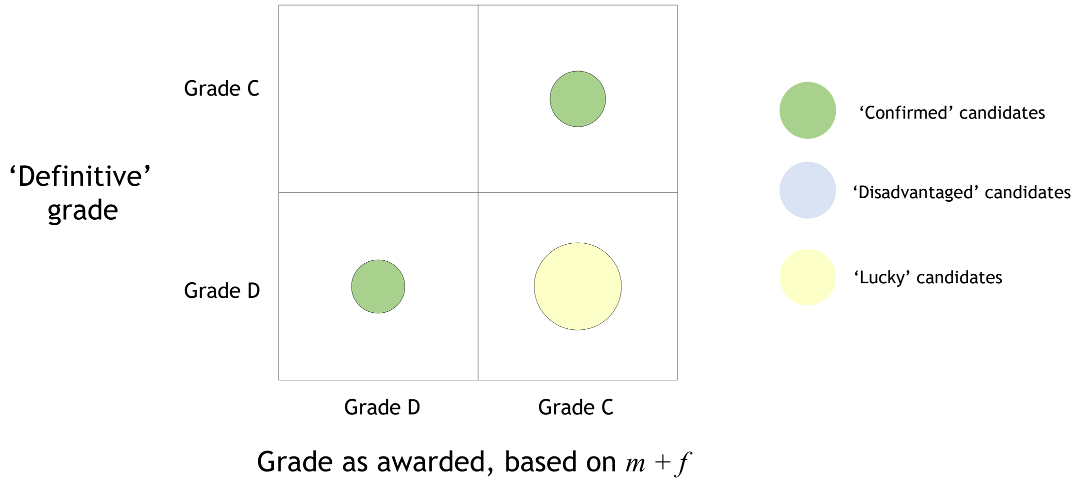
As can be seen from Figure 10, grading according to m results in approximately equal numbers of 'disadvantaged' and 'lucky' candidates; Figure 18, which uses exactly the same marks, re-marks and grade boundaries, shows that grading according to $m + f$ reduces the number of 'disadvantaged' candidates to zero, and increases the number of 'lucky' candidates, with a small number of candidates being 'doubly lucky'.

Importantly, what has *not* happened has been the shrinkage of both the 'disadvantaged' and 'lucky' populations simultaneously, accompanied by a corresponding increase in the populations of those candidates originally awarded the right grade, as represented by the green bubbles. Were this to happen, this would indicate an increase in grade accuracy. As discussed on pages 49 and 50, achieving accurate grades is impossible, but achieving reliability – as evidenced by Figure 17 – is a reality.

If grading – and assessment in general – were totally accurate, then charts such as those shown in Figures 10 and 18 would show only green bubbles along the upwards-sloping diagonal. As has been shown, the policy of grading according to $m + f$ does not deliver accuracy; rather it delivers reliability. But at a price. Generosity. More candidates are 'lucky', and awarded a grade higher than the grade that would be awarded by a senior examiner. But in reality, senior examiners do not mark many scripts, and no one knows what the 'right' grade actually is. What people do know – or could know if they were to appeal, and (given Ofqual's 2016 change in the rules for appeals) allowed to appeal – is the grade resulting from a fair re-mark. Under all the suggested policies, including grading according to $m + f$, the probability of a grade change as a result of a fair re-mark can be reduced almost to zero. Grades will be reliable.

Figure 19 shows the detail of the D/C grade boundary when graded according to $m + f$, this being the $m + f$ equivalent of Figure 11.

Figure 19: The D/C grade boundary: author's simulation of 2018 A level Biology, graded according to $m + f$, for f = 6 marks, compared to the corresponding 'definitive' grades



As can be seen, the population of 'disadvantaged' candidates has vanished: no candidate is denied an opportunity as a result of being an awarded a grade too low; at the same time, the population of 'lucky' candidates has increased.

No, this is not grade inflation

Comparison of Figures 10 and 18 shows that grading according to $m + f$ increases the (simulated) number of candidates awarded higher grades (grades A*, A, B: Figure 10, 30,115 candidates; Figure 19, 40,697) and decreases the number awarded lower grades (grades C, D, E, U: Figure 10, 33,704 candidates; Figure 18, 23,122). This appears to be 'grade inflation', but in fact, this is not the case.

Grade inflation is a steady, year-on-year, progressive increase in the number of candidates awarded higher grades. Comparison of Figures 10 and 18 does indeed show an increase in the number of candidates awarded higher grades and a corresponding decrease in the number awarded lower grades, and this is without doubt a result of awarding grades according to the 'upwards adjusted' mark $m + f$ rather than the original mark m. But this is not grade inflation. Rather, it is a recalibration of the baseline, which is a single event that takes place only at the time that the policy of basing grades on $m + f$ replaces the policy of awarding grades based on m. Once the new policy has been implemented, there is no further effect in subsequent years.

Furthermore, such an effect does not have to happen at all, even when the $m + f$ policy is first implemented. The policy that determines how an assessment is determined from the corresponding original mark - based on m, three grades, based on $m + f$, whatever - is totally independent of, and separate from, the policy determining where the grade boundaries are located. So, for example, it is possible to implement a policy of awarding grades based on $m + f$, and, simultaneously, for an examination for which $f = 8$, to move all the grade boundaries 8 marks upwards. As a result all the candidates are awarded exactly the same grades as they would have received, had grades been awarded on the basis of the original mark m, using the original grade boundaries. So grades are now awarded according to $m + f$, but there is no apparent grade inflation. But it also appears that absolutely nothing different has happened - so what is the point?

In fact, something new, and important, has happened - but it has happened 'behind the scenes'. As well as changing the policy for awarding grades from being based on m to being based on $m + f$, and as well as changing the grade boundaries to avoid the apparent grade inflation, the policy for appeals has changed too. Under the current policy, if the entire cohort were to appeal, about 1 grade in every 4 would be changed. The current policy delivers grades that are very unreliable. But under the $m + f$ policy - with or without changing the grade boundaries - if the entire cohort were to appeal, very few grades would be changed. No matter where the grade boundaries are, no matter what the grade widths are, a policy of awarding grades based on

$m + f$ delivers reliable grades. And because the grades are reliable, they can be trusted.

A plausible, but false, claim

A further point about grading according to $m + f$ is best explained by example. Suppose that, for a particular examination subject:

- grade C is all marks from 51 to 59 inclusive;
- grade B, 60 to 68;
- grade A, 69 to 77, and
- the value of f for this examination subject is 6 marks.

A candidate is given $m = 59$ marks, implying that $m + f = 65$, and under the policy of basing grades on $m + f$, the candidate is awarded grade B. The candidate appeals, and is given a fair re-mark $m^* = 64$ marks. Since the re-mark $m^* = 64$ is within the range from $m - f = 59 - 6 = 53$ to $m + f = 59 + 6 = 65$, the original grade B is confirmed.

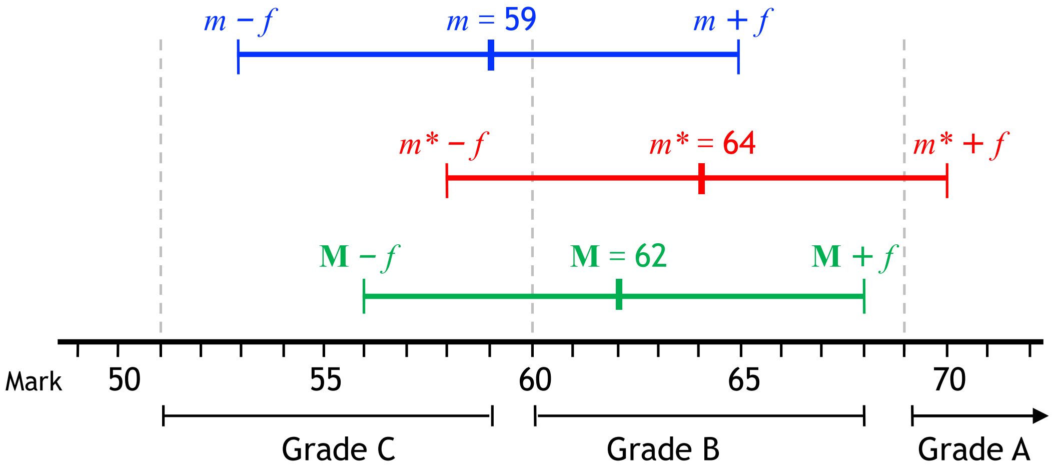
At which point, the candidate makes these claims:

- The original mark $m = 59$ was adjusted to $m + f = 59 + 6 = 65$, resulting in the award of grade B.
- The original mark could have been any mark in the range $m - f = 59 - 6 = 53$ to $m + f = 59 + 6 = 65$, as proven by the re-mark $m^* = 64$. A mark of 64 is therefore a valid mark.
- Had the original mark been the valid mark of 64, then the adjusted mark would have been $m + f = 64 + 6 = 70$, resulting in the award of grade A.
- Since an original mark of 64 is valid, and corresponds to grade A, the grade B as actually awarded on the basis of the actually awarded mark $m = 59$ is unfair, and this unfairness was not corrected by the appeal.

The candidate is correct in that there is indeed a possibility that the script might have originally been marked $m = 64$, resulting in an adjusted mark $m + f = 70$ and the award of grade A. This is true. But is the original award of grade B, and its confirmation, unfair?

The answer to this question is ‘no’. The candidate has been treated fairly, as can be appreciated from Figure 20.

Figure 20: When graded according to $m + f$, a candidate originally marked m = 59, and awarded grade B, is treated fairly, even if the re-mark m^* = 64



In Figure 20, the green line represents the distribution of marks as would be determined if, say, 100 different examiners were each to mark the candidate's script. This 'panel distribution' (see page 49, and also pages 97 and 98 in the Appendix) is important, for any single mark given by any one examiner must lie within this distribution, and any mark that lies within this distribution is correspondingly a valid, legitimate, mark for that script. Figure 20 shows that the median **M** of this distribution is **M** = 62 marks; also, the end-to-end range of this distribution is 68 – 56 = 12 marks = $2f$, implying that f = 6 marks. As discussed on pages 50, and also pages 97 to 100 of the Appendix, the median mark **M** may be taken as 'definitive' or 'right', and so the 'right' mark for this script is **M** = 62, corresponding to grade B. In reality, neither the 'right' mark **M** = 62, nor the 'right' grade B, are known. But if the discussion is about 'fairness', there needs to be a definition of what 'fair' actually is - for example, the 'right' grade as specified by the median **M** = 62. For this example, the 'fair' grade is therefore grade B.

In reality, the candidate's script is given the single mark m = 59 by a single examiner. This mark is within the 'panel distribution' as shown by the green line, and so is a valid mark; accordingly, under the current policy of determining the grade based on the given mark m = 59, the candidate would be awarded grade C. The candidate's 'right' grade, however, is grade B. The candidate is therefore 'disadvantaged'.

If the candidate were allowed to appeal, and the script re-marked m^* = 64, then, as shown by the green line in Figure 20, this mark lies within the same panel distribution as the original mark m = 59, as is required, and should, in principle, result in an up-grade to the 'right' grade B. Under the current policy, however, the appeal would be disallowed, and no re-mark would take place, on the grounds that the original mark m = 59 is not associated with any marking errors and is, to use Ofqual's word, '*sound*'. Any request for a re-mark is therefore a '*second bite of the cherry*', and therefore, to quote Ofqual a third time, '*unfair*'.

Under the policy in force at the time of writing, the candidate is awarded a certificate showing grade C, and denied the opportunity to appeal. Even though, in principle, the 'right' grade is grade B. To the author, this in an injustice. And even though the 'right' grade can never be known in practice, this does not negate the truth. If enough resources were available, 100 examiners could mark the script, and the median $\mathbf{M}$ of the resulting panel distribution determined, so resolving the matter. In reality, these resources are not available, and the truth is never discovered. But the truth is there nonetheless: the candidate was awarded grade C, was not allowed to appeal, and possibly denied life chances as a result. Yet that candidate's 'right' grade was grade B.

Suppose, however, that the same script is given a single mark $m = 59$, and awarded a grade based on $m + f = 59 + 6 = 65$. As shown in Figure 20, the 'upwards adjusted' mark $m + f = 65$ corresponds to grade B, and it is grade B - the 'right' grade - that appears on the candidate's certificate. Under the current policy, the candidate would be awarded grade C, be 'disadvantaged', and denied recourse to appeal; under a policy of grading according to $m + f$, the candidate is awarded the 'right' grade B at the outset, a grade that is confirmed by a re-mark $m* = 64$, which is within the range from $m - f = 53$ to $m + f = 65$.

Because the re-mark $m* = 64$ lies within the panel distribution as shown by the green line in Figure 20, then, as the candidate claims, it is possible that this might have been the originally-given mark. Had this happened, the candidate is correct in claiming that the grade awarded would have been based on $m + f = 64 + 6 = 70$ marks, corresponding to grade A. Given that the 'right' grade is grade B, the award of grade A is 'lucky'.

The distinction between 'disadvantaged', 'right' and 'lucky' is central to a discussion of 'fairness'. Under the current policy, the original mark $m = 59$ corresponds to grade C, and the candidate is both 'disadvantaged' and denied the opportunity to appeal. Under a policy of grading according to $m + f$, the candidate is awarded the 'right' grade B, as is confirmed on appeal.

Had the original mark been 64, which is quite possible, then under the current policy, the candidate would have been awarded the 'right' grade, grade B; under a policy of grading according to $m + f$, the candidate would be awarded grade A, and so would be 'lucky'.

In this example, the candidate has claimed that it is unfair to be awarded grade B when there is a possibility that the grade awarded might have been grade A. In the author's view, this claim is unfounded. As summarised in Figure 20, the award of grade B is 'right', and the award of grade A is 'lucky'. By being awarded grade B, the 'right' grade, it is true that the candidate has not been 'lucky'. But is 'not being lucky' the same as 'being treated unfairly'? The author thinks not. To the author, being awarded the 'right' grade is fair, especially when contrasted with the current, deeply unfair, policy in which the award of grade C implies that the candidate is 'disadvantaged', and denied the opportunity to appeal.

This is the essence of the comparison between Figures 10 and 18. Under the current policy, as shown in Figure 10, some candidates are awarded the 'right' grade, some are 'disadvantaged', and some 'lucky'. When grades are based on $m + f$, some candidates are awarded the 'right' grade, and rather more (compared to grading according to m) are 'lucky'. But – and very importantly – when grades are based on $m + f$, very few candidates, if any, are 'disadvantaged'. And as a result, very few, if any, are denied potentially life-changing opportunities.

The lottery-of-the-first-mark revisited

In this example, and as illustrated in Figure 20, both marks $m = 59$ and $m^* = 64$ are members of the same panel distribution, and so both are valid marks for the same script. Furthermore, either mark might have been the original mark, and as discussed, these two different marks result in different grades, even when the policy for grading is based on the 'upwards adjusted' mark $m + f$: for the mark $m = 59$, $m + f = 59 + 6 = 65$, grade B; for the mark $m = 64$, $m + f = 64 + 6 = 70$, grade A. The grade awarded therefore depends on which examiner marks the script first, even when grades are based on $m + f$. Is this another example of the lottery-of-the-first-mark, as discussed on page 11? This lottery was identified as a major injustice of the current policy of basing grades on the mark m, and so if basing grades on $m + f$ does not resolve this problem, then it appears that nothing has changed, and nothing gained. So why bother with all the fuss of grading according to $m + f$?

In fact, two important things have changed, but the fundamental point is true: grading according to $m + f$ does not eliminate the lottery-of-the-first-mark. As the example shows, different examiners marking the same script could result in different grades, with the grade actually awarded still being a lottery, even when grades are based on $m + f$.

But as shown in Figures 10 and 11, grading according to m results in three populations – 'confirmed' candidates, 'disadvantaged' candidates, and 'lucky' candidates. Figures 18 and 19, however, show that grading according to $m + f$ results in only two populations – 'confirmed' and 'lucky'. The 'disadvantaged' population has been eliminated, and, from the standpoint of social policy, no young person would be denied an opportunity.

Secondly, there is an important difference as regards the consequences of a fair re-mark resulting from appeal. Under the current policy, a fair re-mark m^*, if allowed, takes precedence over the original mark m, and if the re-mark m^* lies on the other side of a grade boundary as compared to the original mark m, the grade is changed accordingly. That is why current grades are unreliable.

If, however, the grade is based on $m + f$, and if f has been determined statistically correctly, then the likelihood that a fair re-mark m^* would be less than $m - f$ or greater than $m + f$ is very low. It is therefore very unlikely that the originally-awarded grade would be changed. This applies whether or not, in the example used, the originally-awarded graded was grade B or grade A: whichever of these two grades was originally awarded, that grade would be confirmed. That is why grading according to $m + f$ results in reliable grades.

The lottery-of-the-first-mark is not a result of the policy used for determining grades, and the lottery exists whether grades are determined by m, $m + f$, or any other algorithm. Rather, the lottery is a direct consequence of the unavoidable reality that different examiners can give different marks to the same script. Since that original mark is a lottery, everything that follows has this fundamental lottery as its basis.

This discussion also highlights the distinction between accuracy and reliability, as examined on pages 49 to 56. Accurate grades can be given only if every examiner gives the same mark to the same script without exception. If this is not the case, if different examiners can indeed give different marks to different scripts, if marks are fuzzy, then there must be a lottery-of-the-first-mark, however grades are awarded, and even if grades are not awarded at all. Given that, for essay-style examinations, fuzziness must exist, then, as stated on page 50, the search for accuracy is a search for a chimera, a snark. And an inevitable result of the impossibility of achieving accuracy is that the first-given mark is necessarily a lottery.

Importantly, however, the consequences of that lottery can be managed and controlled. As has been shown, grading according to $m + f$ results in reliable grades, and the elimination of ‘disadvantaged’ candidates; as will be shown in the next section, grading according to $m - f$ achieves an outcome which is similar, but different in one important respect.

One grade based on $m - f$

Under this policy, the assessment of a script originally marked m is a single grade based on the ‘downwards adjusted’ mark $m - f$. Like grading according to $m + f$, grading according to $m - f$ also results in reliable grades, as shown in Figure 21, which shows the grades after appeal for 2018 A level Biology for $f = 6$.

Figure 21: Author's simulation of grades after appeal for 2018 A level Biology, graded according to $m - f$, for $f = 6$ *marks*

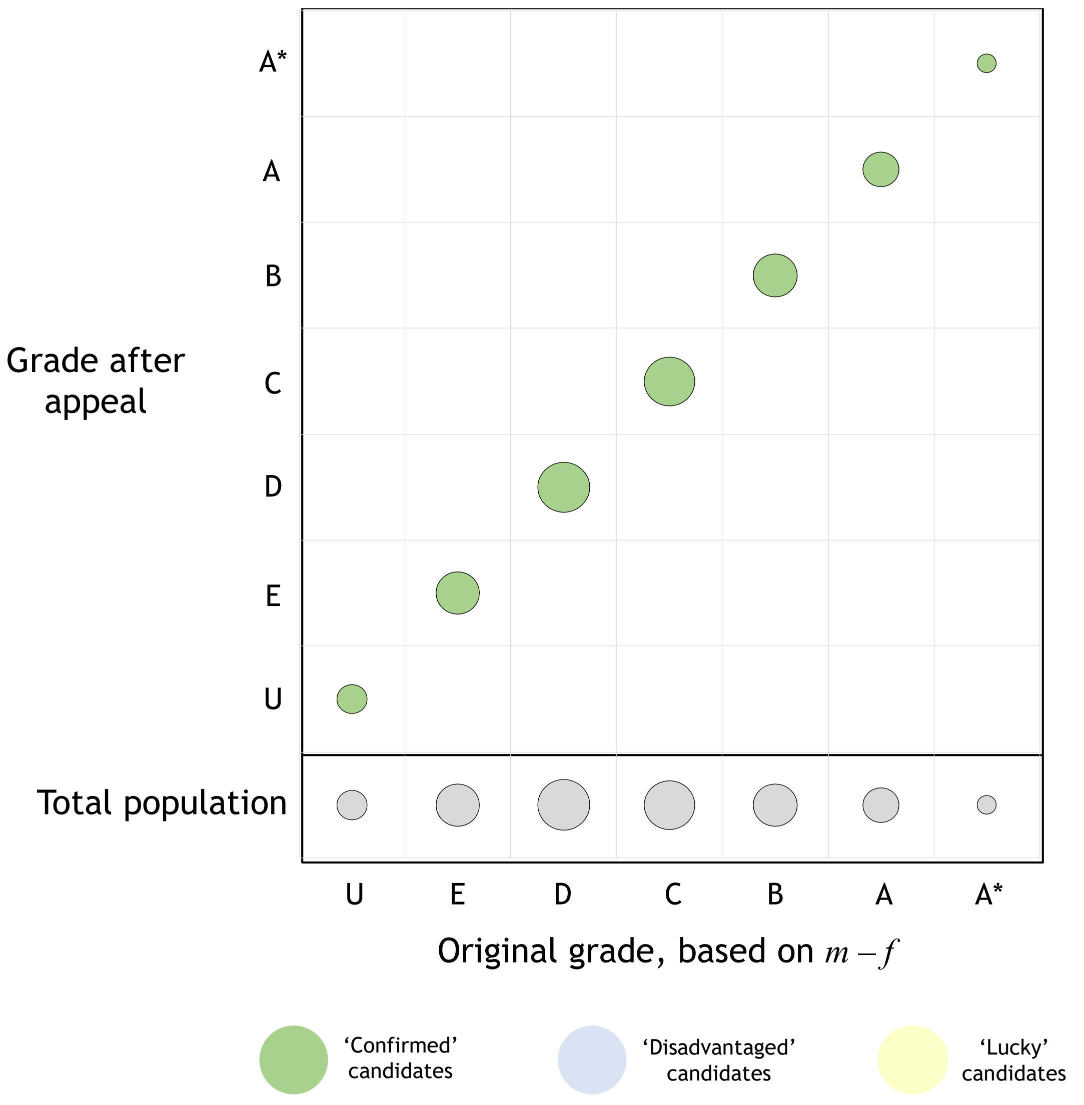


Figure 21, for grading according to $m - f$, may be compared to Figure 18, for grading according to $m + f$. At first sight, Figures 18 and 21 appear to be identical, for both show only green bubbles along the upwards-sloping diagonal. Closer inspection, however, shows that the sizes of both the green and the grey bubbles are different - in Figure 18, the bubbles for the higher grades are larger, and those for the lower grades smaller, than the bubbles for the corresponding grades in Figure 21.

This is because, as has been discussed, and as depicted in Figures 18 and 19, awarding grades based on $m + f$ is generous, minimising the population of 'disadvantaged' candidates and increasing the population of 'lucky' candidates; by contrast, grades according to $m - f$ is stringent, minimising the population of 'lucky' candidates and increasing the population of 'disadvantaged' candidates, as shown in Figures 22 and 23.

Figure 22: Author's simulation 2018 A level Biology, graded according to $m - f$, for $f = 6$ marks, compared to the corresponding 'definitive' grades

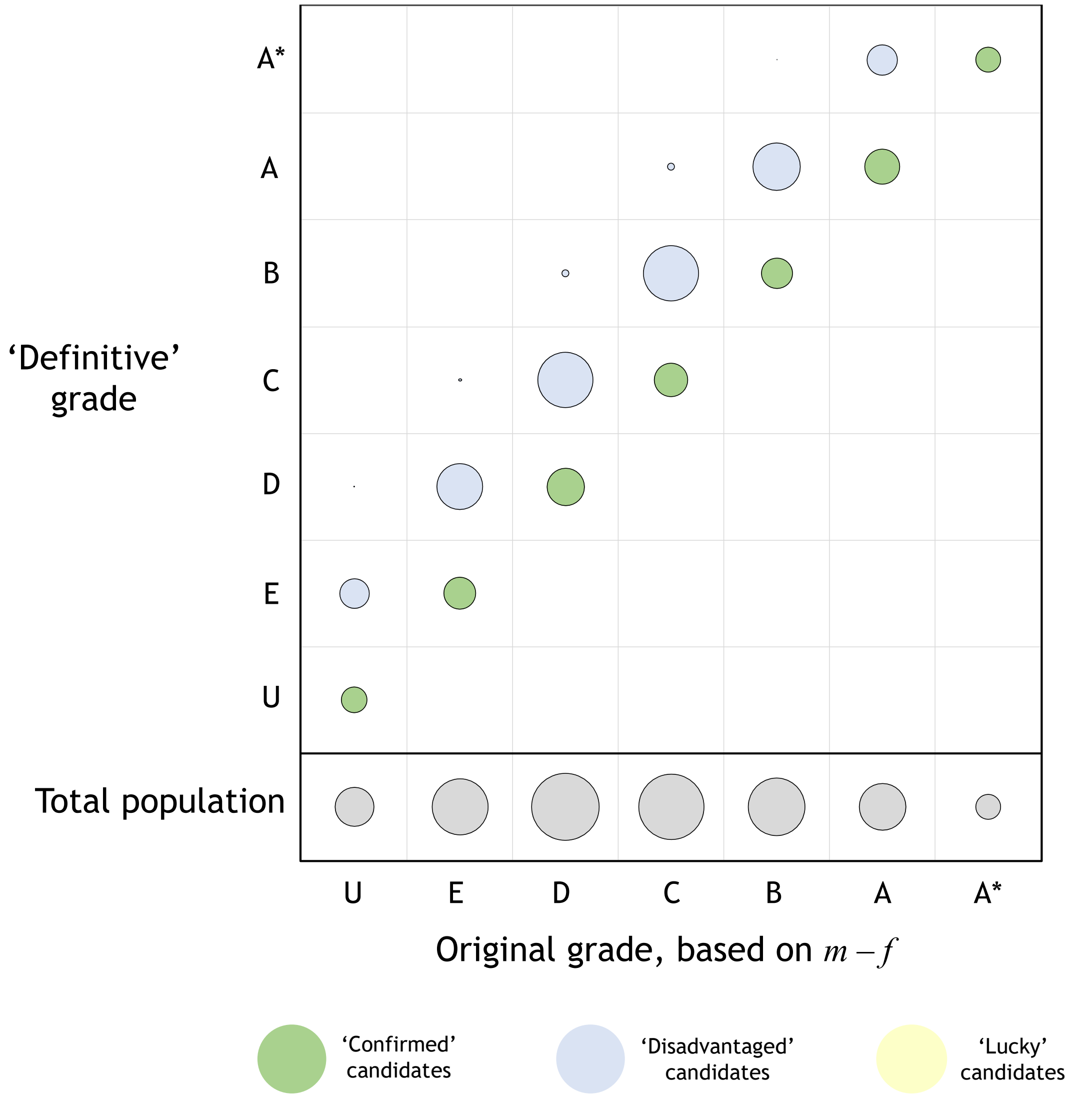
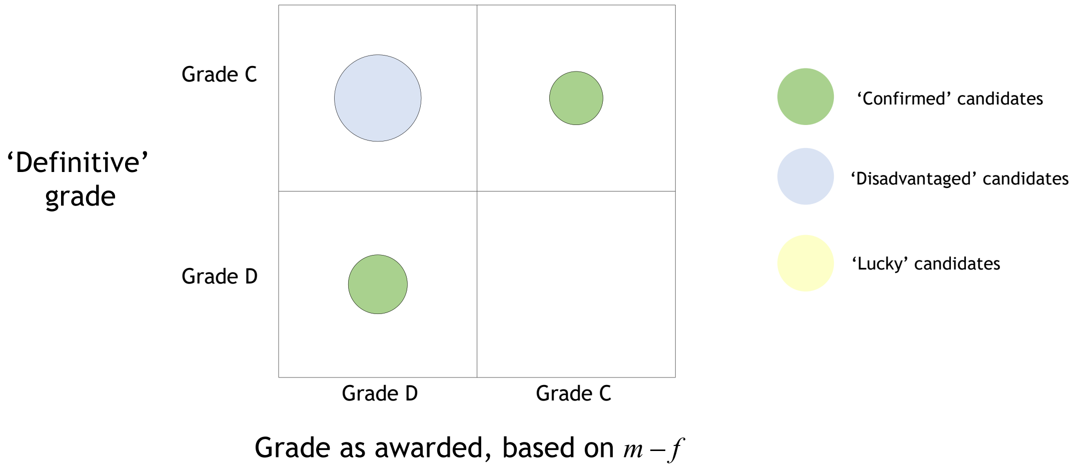


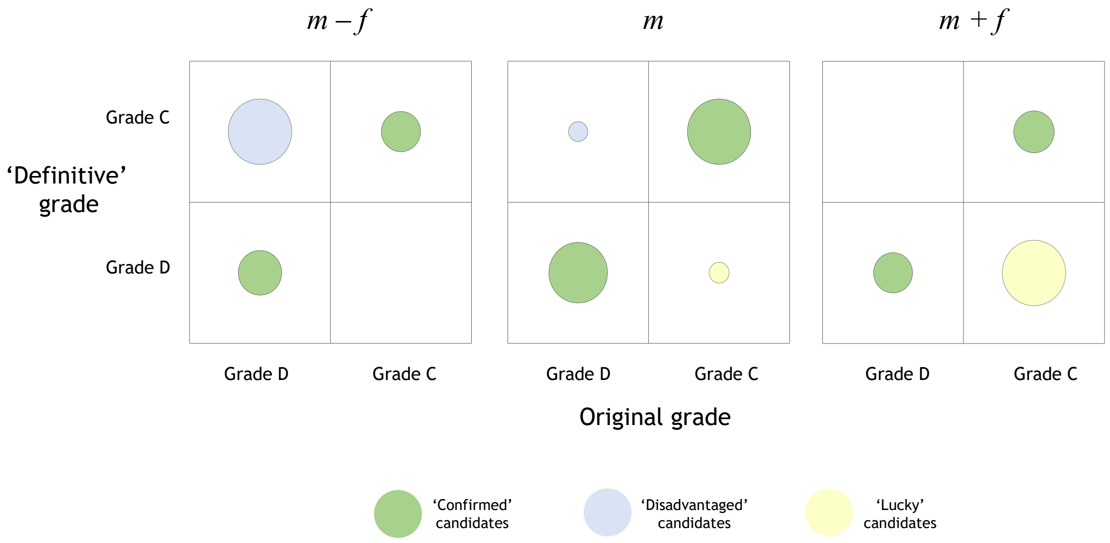
Figure 23: The D/C grade boundary: author's simulation 2018 A level Biology, graded according to $m - f$, for f = 6 marks, compared to the corresponding 'definitive' grades



Overall, comparison of Figure 22 with Figure 10 shows that the number of candidates awarded higher grades has reduced, and the number awarded lower grades increased. Just as grading according to $m + f$ appeared to be driving grade inflation, grading according to $m - f$ appears to be driving grade deflation. As discussed on pages 68 and 69, this is not the case; rather what is happening is a once-only grade re-calibration, taking place at the cut-over from awarding grades based on m to grade based on $m - f$.

Figure 24 draws Figures 11, 19 and 23 together, so as to highlight a key implication of the choice of policy for assigning grades from the original mark m for any examination subject associated with a given value of f.

Figure 24: A matter of policy



For all three cases illustrated in Figure 24, the mark m is the same, as are the locations of the grade boundaries. The difference is solely attributable to the policy adopted for the assessments shown as the grades appearing on candidates' certificates: according to $m - f$ (left), m (centre) or $m + f$ (right).

Grading according to m results in almost equal populations of 'disadvantaged' and 'lucky' candidates, the sizes of which are greater, the larger the value of f. Grading according to $m + f$ reduces the population of 'disadvantaged' candidates to very close to zero and simultaneously increases the number of 'lucky' ones; grading according to $m - f$ has the opposite effect, minimising the population of 'lucky' candidates and increasing the number of 'disadvantaged' ones.

From a rather different standpoint, the policy of grading according to $m - f$ ensures that it is most unlikely that anyone awarded a given grade might have done so 'undeservingly', for the absence of 'lucky' candidates ensures that no candidates actually awarded any grade might be down-graded should their scripts be re-marked by a senior examiner.

Similarly, a policy of grading according to $m + f$ results in a negligible population of 'disadvantaged' candidates, and ensures that very few candidates are awarded a grade lower than the grade that would be awarded by a senior examiner. Very few candidates are therefore denied an opportunity which they deserve, but some of the candidates awarded any grade – the 'lucky' ones – might be 'under-qualified'.

Given that the underlying marks are the same in all three cases, it as a matter of policy choice as to which is adopted – and different policies might be appropriate for different circumstances. Few people, for example, would feel happy with the possibility that their brain surgeon had been 'lucky' in his or her final examinations, so perhaps a policy of grading those examinations according to $m - f$ is wise. The same might apply to qualifications relating to, for example, gas fitting, electricians, the driving test – and perhaps some of the new T levels too. For subjects such as GCSE Geography, however, surely there is nothing to be lost, and much to be gained, by being generous, and giving candidates the 'benefit of the doubt': is it not better for young people who might be under-qualified to be offered opportunities than for candidates who are qualified to be denied them?

This is indeed a matter of policy that needs to be debated.

Figure 24 has one further, important, feature – a feature that is self-evident once attention is drawn to it, but a feature that is easily overlooked. The policy, currently in place, of grading all school examinations according to the original mark m is based on an assumption. As is evident from Figure 24, this assumption is that, for all subjects, and at each of GCSE, AS and A level, the value of f is zero.

Currently, there are no measures of f for any examinations, other than the inferences that can be drawn from charts such as those shown in Figures 5, 6 and 7, as discussed on page 24. But although no-one knows what any value of f actually is, everyone knows the one value that f is *not*. Zero. As Ofqual explicitly state *'There is often no single, correct mark for a question. In long, extended or essay-type questions it is possible for two examiners to give different but appropriate marks to the same answer. There is nothing wrong or unusual about that.'* All marking is fuzzy. Therefore all marking is associated with a value of f that must be greater than zero. Yet Ofqual's policy as regards not only grading, but also (and importantly) appeals, denies this: Ofqual's policy assumes that the value of f is the most unrealistic value imaginable, zero.

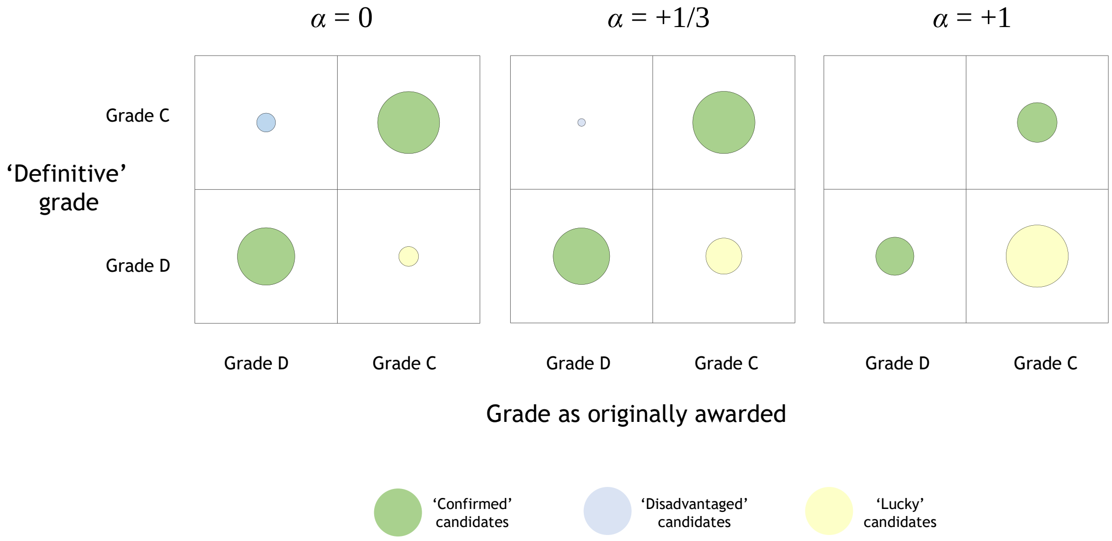
Not just that: as will be shown in the next section, the assumption that $f = 0$ results in the maximum possible unreliability for any examination subject - any other value gives more reliable grades.

Basing grades on $m + \alpha f$

The policy of basing grades on $m + \alpha f$ is a generalisation: α is a parameter that can take any value from -1 to $+1$ as determined for any particular examination subject in accordance with an agreed policy, and αf can be used instead of f in each of the three grade, two grade and one grade possibilities just discussed. So, for example, for the 'one grade' solution, the choice $\alpha = +1$ results in awarding grades based on $m + f$; similarly, the choice $\alpha = -1$ results in awarding grades based on $m - f$; the choice $\alpha = 0$ results in awarding grades based on m as is the current policy - although describing the current policy as a 'choice', with the implication that other possibilities were identified, wisely evaluated, and then explicitly rejected, is something on which the reader may wish to form his or her own opinion.

The effect of selecting a value for α other than $+1$, -1 or 0 can be inferred from Figure 24. A positive value for α, less than $+1$, will result in a population of 'disadvantaged' candidates smaller than that for the policy of grading according to m (corresponding to $\alpha = 0$, as shown in the central diagram), but greater than that for the policy of grading according to $m + f$ (corresponding to $\alpha = +1$, as shown in the right-hand diagram). Similarly, the population of 'lucky' candidates will be greater than as shown for m, but fewer than as shown for $m + f$. In short, the distribution will be intermediate between those shown in the centre and the right, such that the closer the value of α to zero, the more like the centre, the closer the value of α to $+1$, the more like the right. Figure 25 shows an example based on the author's simulation of 2018 A level Biology, for which $f = 6$, using $\alpha = 1/3$ (implying that $\alpha f = 2$ marks, corresponding to grading according to $m + 2$).

Figure 25: The D/C boundary, graded according to $m + \alpha f$ *for* $\alpha = 0, +1/3$ *and* $+1$*, author's simulation of 2018 A level Biology, for which* $f = 6$



By the same token, a negative value for α, but not more negative than -1, will result in a distribution intermediate between the diagram at the centre if Figure 24, and that on the left: the closer the value of α to zero, the more like the centre, the closer the value of α to -1, the more like the left, as exemplified by Figure 26.

Figure 26: The D/C boundary, graded according to $m + \alpha f$ *for* $\alpha = -1, -1/3$ *and* 0*, author's simulation of 2018 A level Biology, for which* $f = 6$

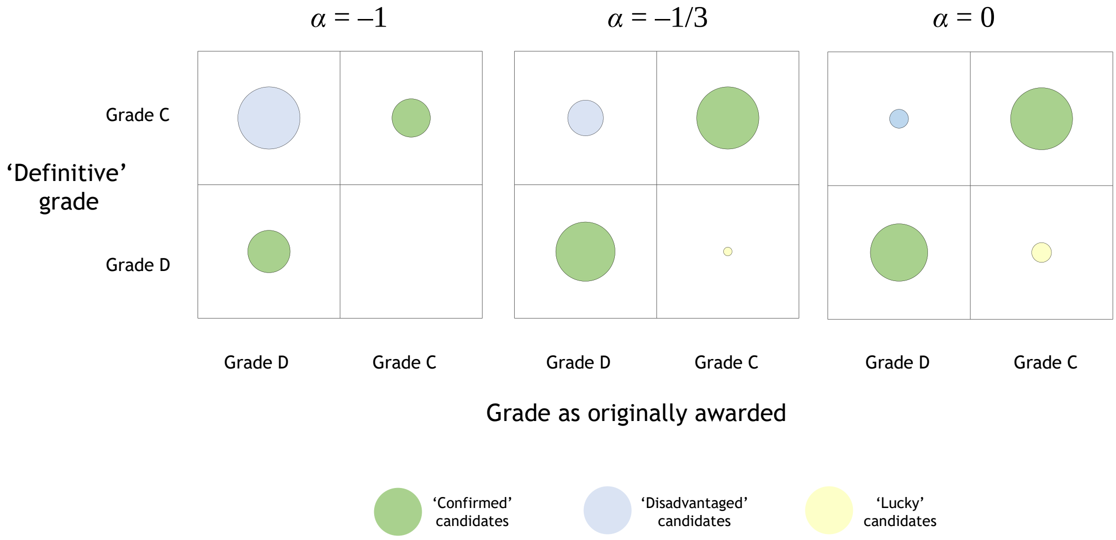
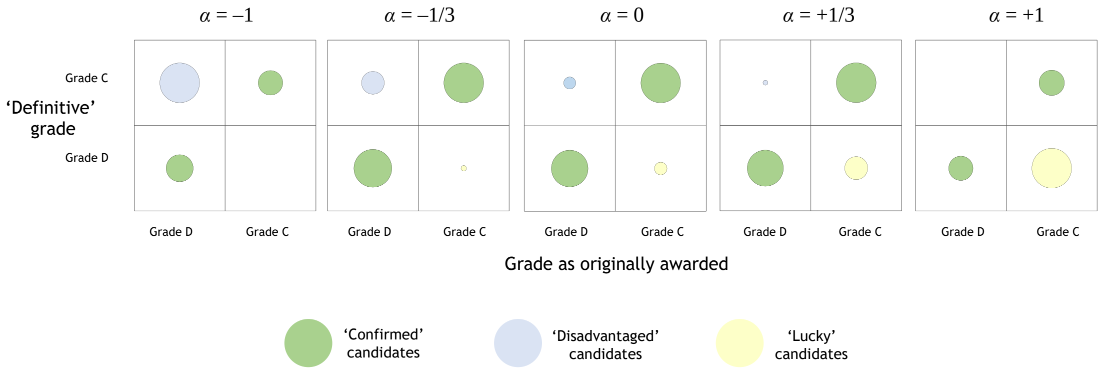


Figure 27 combines Figures 25 and 26, and represents the 'full spectrum' from $\alpha = -1$ to $\alpha = +1$.

Figure 27: The D/C boundary, graded according to $m + \alpha f$ *for* $\alpha = -1, -1/3, 0, +1/3$ *and* $+1$*, author's simulation of 2018 A level Biology, for which* $f = 6$



Fundamentally, the policy choice of the value of α enables control over the relative populations of 'disadvantaged' and 'lucky' candidates, and hence over grade reliability. When $\alpha = 0$, corresponding to the assumption that $f = 0$, as is the case for the current policy, the reliability of a given examination subject is determined by that subject's intrinsic fuzziness, as shown in Figure 1 for the 14 subjects measured by Ofqual. Each subject has its own 'signature' as visualised by, for example, a full chart of the type shown in Figure 10, or across any grade boundary, as exemplified by Figure 11 and the central chart in Figure 27. For more reliable subjects, the green bubbles along the diagonal will be relatively large, and closer to the sizes of the corresponding grey bubbles in the top row of Figure 10; the sizes of the off-diagonal bubbles, both blue (for 'disadvantaged' candidates) and yellow (for 'lucky' candidates) will be relatively small, and these will be present only one grade on each side of the diagonal. As the grade unreliability increases, the sizes of the green bubbles decrease; the sizes of the blue and yellow bubbles increase; and small bubbles begin to appear two, and eventually three, grades away from the diagonal, as exemplified by the simulation for 2018 GCSE English Language shown in Figure 12.

When $\alpha = +1$, if the value of f is the full end-to-end range of either of the two re-mark distributions (as compared to a senior examiner, or another ordinary examiner), then the grades are 100% reliable (relative to re-marks by a senior examiner or an ordinary examiner, as determined by which re-mark distribution is being used). Exactly the same applies if $\alpha = -1$, and the grades determined according to $m + \alpha f = m - f$.

For intermediate values of α, awarding grades based on $m + \alpha f$ results in grade reliabilities less than 100%, but, for any given examination subject, greater than the reliability for that subject when $\alpha = 0$, with grades awarded on the basis of the original mark m. This is the justification of the statement made at the end of the last section that the current policy, under which grades are awarded on the basis of the original mark m, maximises grade unreliability.

For any given examination subject, and hence a corresponding intrinsic value of f, it is possible to compute the resulting grade reliability for any value of α in the range from -1 to $+1$. As an example, Ofqual's research, as summarised in Figure 1, shows that the grades for GSCE English Language are about 61% reliable, when measured by reference to re-marks by a senior examiner. According to the author's simulation, this suggests an estimate for f of about 12 marks, and Table 6 shows the results of further simulations for different values of α, each chosen such that the product αf is a whole number.

Table 6: Author's simulation of average grade reliabilities for 2018 GCSE English Language, when graded according to $m + \alpha f$, where the values of α can be positive or negative, and $f = 12$

α	αf	Average grade reliability
0.0000	0	61%
0.0833	1	64%
0.1667	2	69%
0.2500	3	76%
0.3333	4	83%
0.4167	5	89%
0.5000	6	94%
0.5833	7	97%
0.6667	8	98%
0.7500	9	99%
0.8333	10	99.7%
0.9167	11	99.9%
1.0000	12	100.0%

As can be seen, the average grade reliability has its lowest value for $\alpha = 0$, and a maximum when $\alpha = +1$; also, the reliabilities have the same values when α is both positive or negative, implying that the reliability is 76% for $\alpha = +0.250$ (corresponding to grading according to $m + 3$), and also $\alpha = -0.250$ (corresponding to grading according to $m - 3$). As can be seen, the grade reliability rises quite quickly from 61% ($\alpha = 0$) to 94% ($\alpha = 0.5$ or $= -0.5$), and then approaches 100% rather more slowly, this being a consequence of 'flat tails' towards either extreme of the re-mark distribution, as exemplified by Figure 8.

This leads to in important policy question: what level of grade reliability do we seek? For GCSE English Language, is a reliability of 61% acceptable? If it is, then better to leave the current processes alone. But should the reliability be 98%? 99%? 99.9%? In which case, the current grading policy needs to be replaced by a different one. And what about all the other subjects? This is a debate that needs to be held.

But beware attractive-looking numbers. An average grade reliability of 98% (α = 0.6667 with grading based on m + 10) appears to be very good, and is the kind of number that would be claimed to be 'world class'. For 2018 GCSE English language graded 9, 8, 7..., the cohort was 683,838, and so a reliability of 98% implies that about 13,600 students would be awarded the wrong grade, about 6,800 being awarded too high a grade, and about 6,800 too low a grade. Compared to the total cohort of 683,838, 6,800 students awarded too low a grade might be regarded as a 'small number'. But to each one of those 6,800 students, the award of too low a grade could deny a life-changing opportunity. Especially since those students could be awarded a reliable grade simply by adopting the policy α = + 1, and basing grades on m + 12.

Dispensing with grades altogether

Fundamentally, the problem of grade reliability is attributable to the attempt to map a necessarily fuzzy mark (such any mark in the range 57, 58, 59, 60, 61) onto a necessarily 'cliff-edged' grade boundary (grade C is all marks from 50 to 59 inclusive; grade B, 60 to 69). The various solutions discussed so far have had the common thread that each is a different way of identifying a single mark that can be mapped onto the cliff, with the objective of avoiding the possibility that other marks might be dangling over the edge.

This solution is different. Instead of manipulating the marks, get rid of the cliff. If grades are the problem and no longer fit-for-purpose – as indeed they are – what would happen if grades were no longer used? The answer is that the assessment as awarded to candidates, and as recorded on their certificates, would no longer be in the form of grades, but in a different form - a form with the considerable benefit of being reliable.

So instead of awarding unreliable grades of the form

- History, grade C
- Mathematics, grade B

a candidate's certificate might show, for example

- English Language, 59 marks (+/– 12 marks)
- Mathematics, 62 marks (+/– 2 marks)

these being the marks m as given by the examiners who marked the scripts, (scaled, as required, to a standardised range such as from 0 to 100), and also the corresponding measures of the fuzziness for each examination subject.

It really is as simple as that.

Declaring measures of fuzziness might be regarded as alarming. Indeed, page 70 of a report published in 2005 by the examination board AQA, includes these words:

> *However, to not routinely report the levels of unreliability associated with examinations leaves awarding bodies open to suspicion and criticism. For example, Satterly (1994) suggests that the dependability of scores and grades in many external forms of assessment will continue to be unknown to users and candidates because reporting low reliabilities and large margins of error attached to marks or grades would be a source of embarrassment to awarding bodies. Indeed it is unlikely that an awarding body would unilaterally begin reporting reliability estimates or that any individual awarding body would be willing to accept the burden of educating test users in the meanings of those reliability estimates.*

Measures of fuzziness may indeed be '*a source of embarrassment*'. But fuzziness is real. It exists. It has significant consequences. So pretending that it does not exist and covering it up just to avoid '*embarrassing*' the examination boards could well be regarded as a very high price to pay - and, in the author's opinion, an unacceptable one.

Some implications

No solution is perfect

If it were possible for every script to be marked by a single, expert, individual, for that individual to apply consistent standards throughout, and for that individual to give the same script the same mark on a fair re-mark, then this would assure that every candidate is awarded a grade that is both accurate and reliable. This might be possible in practice for subjects with few candidates, such as some of the modern foreign languages. But for subjects with tens or hundreds of thousands of candidates, such an approach is totally impracticable.

As has been discussed many times, for examinations structured around essay-style questions, different examiners can give different marks to the same script. There is no single, unique, right mark. As a consequence, the lottery-of-the-first-mark is unavoidable and inevitable, which in turn implies that accuracy - giving every script the right first mark – can never be achieved. But reliability - confirming the originally-awarded grade as the result of a fair re-mark - can.

As has been shown, the current policy of awarding grades based on the original mark m is deeply flawed. The various other policy opportunities suggested all have the very important benefit of delivering reliability, but all have implications, and none is perfect.

Given that no solution is perfect, it is important that each of these possible solutions is examined thoroughly and in a balanced and fair way, and compared against a continuation of current policies. This document does not attempt to do this, but might be useful in setting the scene; that said, this final section opens the discussion on some of the key issues.

Measuring fuzziness

An important assertion throughout this paper has been that the fuzziness, as measured by the parameter f, is a property of an examination subject, and not a property of the mark or the candidate. As discussed on page 35, the significance of this is that it implies that the fuzziness of any subject needs (in principle) to be measured only once, so enabling that single value to be used for all scripts, for example, by awarding all grades based on $m + f$.

The author is not aware of any published data of actual measurements of f, nor of measurements of f for different marks within the same examination. There is therefore no current evidence that f does indeed have sensibly the same value for all marks – where 'sensibly' means that the value of f does not need to be identical 'to six decimal places' for all marks, but that an average value of f can be used in practice.

The author's assertion that the value of f depends only on the examination subject is based on:

- firstly, the regularity of the shapes of the 'arches' shown in Figures 5, 6 and 7, and in the (small number) of similar charts as published in Ofqual's November 2016 and November 2018 reports; and
- secondly, the results of the author's (many) simulations in which a single value of f has been used, and the general agreement of the results of those simulations (such as the charts illustrated in Figure 9 and 13) with Ofqual's published findings, as well as the validity of the insights obtained from the various bubble diagrams such as those shown in Figures 10, 12, 17, 18, 21 and 22.

This is inadequate. As a matter of urgency, it is imperative that a rigorous statistical study be made of actual data, so that the assertion that f depends only on the examination subject is either confirmed or refuted.

Once fuzziness has been measured

On the assumption that the assertion that f depends only on the examination subject will be confirmed, the measurement of f – perhaps as suggested on pages 56 and 57 – can become a routine process for all subject examinations.

It would therefore be wise, at the outset, to anticipate the possibility that the measured values of f for a given subject – say, GCSE Geography – in any year might be meaningfully different for different examination boards, where 'meaningfully different' means that the observed differences cannot be explained in terms of the statistical errors associated with any measurement. Should this happen, a consequence is likely to be the raising of questions such as:

- Why is the value of f for the same subject different for different examination boards?
- What impact, if any, does the value of f have on the choices made by schools as to which examination board to use?

These questions are difficult. And their difficulty is a possible incentive not to open this particular Pandora's Box.

Pandora's Box, however, is already open. Ofqual's November 2016 report loosened the lid; the November 2018 report made the contents fully visible.

Another possible Pandoran consequence

In 2016, a former student sued Oxford University claiming that, 16 years previously, 'inadequate teaching' had prevented him from being awarded a first-class degree, so denying him a successful career. In fact, he had been awarded an upper second, and in a judgement made in 2018, he lost the case.

In response to the allegation of 'inadequate teaching', one might imagine that the University might counter-allege 'indolent learning' on the part of the student. Both the allegation and the counter-allegation are hard to prove.

So imagine the possibility that, at some time in the future, an individual brings a case that the grade B awarded for an A level taken in, say, 2016, had resulted in failure to win a place at a particular university, so denying the individual a successful career. The basis of the claim might be statistical, referencing, for example, the chart shown in Figure 1; alternatively, the original script might still exist and so be available for a fair re-mark, which might turn out to be grade A.

Who, though, is the defendant? If it is an examination board, they might (validly) state that the original grade B was based on a mark that was '*appropriate*', and that no marking errors had been made. The original award of grade B was therefore fully compliant with the grading policy in force at the time – a policy that the examination board has a duty to follow, but a policy determined elsewhere. If a consequence of the policy is that some candidates are 'disadvantaged', then culpability must lie at the door not of those who compliantly execute that policy, but of those who set it.

Such an eventuality is indeed truly Pandoran, let alone the precedent that is set for others to dip into the box too.

It could well be worthwhile to think this through at the outset.

Educating the stakeholder community

Rather less dramatic, but nonetheless of considerable importance, is the impact that any change in the way in which candidates' assessments are presented on certificates might have on those who use those assessments – stakeholder communities such as schools, colleges, universities and employers, and of course parents and the students themselves. As the experience of the recent change in the grading structure for GCSE examinations from A*, A, B... to 9, 8, 7... bears witness, any change that affects so many very different people must be preceded, and accompanied, by an extensive, and well-formulated communication campaign.

Grades are familiar, and (more or less) well understood; they are also easy-to-use in that, for example, an organisation that has a policy that to be eligible for a particular programme, the candidate must have a minimum of grades BBB can use this as a filter, and so reject any candidate with grades BBC. The candidate can use this as a filter too, and so any candidate with grades BBC might not apply in the first place. Since grades are so unreliable, such filters are untrustworthy from both perspectives, but while stakeholders have the belief that grades are reliable, they will continue to be used as if they are indeed reliable, with no one the wiser.

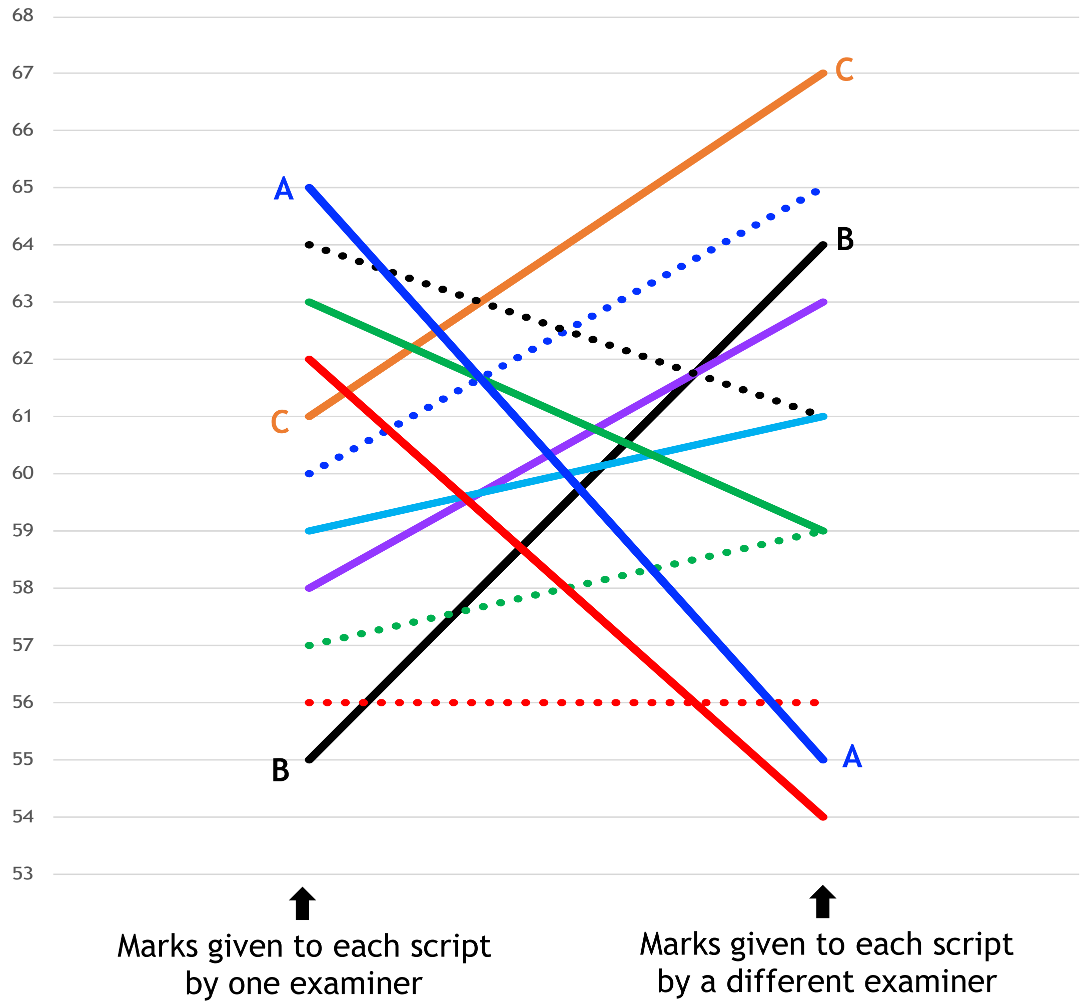
Familiarity and ease-of-use present two significant barriers to change, and will undoubtedly be cited as insuperable by those who will argue to maintain the *status quo*. Because grades are familiar and have been used for such a long time without question, the evidence that grades are in fact unreliable is quite likely to be dismissed. The question "If grades are unreliable, why has this only now been discovered?" is a good one, but perhaps asked not in a spirit of enquiry but rather as a mask for "this must be fake news – if it were true, we would have heard about it years ago".

Ease-of-use is important, for interpreting a candidate's assessment should be neither a burden, nor a form of ordeal-by-statistics. A present-day certificate stating 'Geography Grade B' is easy to understand, and offers a (simplistic, and therefore easy-to-use) blend of bunching (all candidates awarded grade B are indistinguishable) and stratification (all candidates awarded grade B are more able than those awarded grade C, but less able than those awarded grade A).

A certificate stating (for example) 'Geography, 59 marks +/– 9 marks' (see Figure 8) is more obscure, requires arithmetic ("just what is 59 plus 9?"), and raises questions such as "Did the candidate get 59 marks? Or 68? or 50? And if the candidate did get 68, why have all the clutter?" Furthermore, if two candidates are competing with one another for the same position, one with a certificate showing 59 +/– 9, the other, 61 +/– 9, which candidate is 'better'? In the 'good old days', the first candidate was awarded grade B, and the second grade A, so the choice was easy. And although 59 +/– 9 and 61 +/– 9 overlap substantially, 61 is still greater than 59, and, for those who can be bothered to do the arithmetic, 70 is greater than 68 too. Which 'proves' that candidate 2 is 'better' than candidate 1 – so, once again, why the clutter?

In the author's opinion, the 'clutter' is useful in that it informs any user of the grades that the two candidates are in essence indistinguishable based on the examination results alone, and that other evidence – perhaps an interview – would provide valuable further insight. The clutter also highlights an important truth: that the concept of a 'true rank order', such that Chris is 'better' than Alex, is misleading – it is an artefact, attributable to the lottery-of-the-first-mark. Essay-based examinations necessarily have fuzzy marks. And an unavoidable consequence is that the ranking of results is correspondingly blurred, as illustrated in Figure 26, based on the author's simulation of 2018 A level English Literature.

Figure 26: Same scripts, different examiners, different rank orders – 2018 A level English Literature



On the left are the marks given by an ordinary examiner to each of 11 scripts. This defines a rank order from a highest mark of 65 (awarded to candidate A) to a lowest mark of 55 (candidate B). On the right are the marks given to each of the same scripts by a different (ordinary) examiner. As can be seen, the resulting rank order is very different: Candidate A, ranked 1st on the left, is ranked 10th on the right; candidate B, originally ranked 11th, is ranked 3rd on the right; candidate C, originally ranked 5th, is ranked 1st on the right, with a mark higher than the top mark originally given on the left.

The rank orders shown are just two specific cases out of a huge number of possibilities, all for the same scripts, but as marked by different examiners. In practice, each script is marked just once, and that single mark determines the candidate's grade. Those first-given marks might correspond to those on the left of Figure 26, or to those on the right – or indeed to any other of the many possible rank orders not shown. This is the lottery-of-the-first-mark made real. Rank orders are indeed blurred.

But for stakeholders who have become accustomed to a rank order that 'works' – even though it is wrong, misleading, and does great injustice to very many young people – getting used to more sophisticated, more nuanced, information is difficult. But this undoubted difficulty is not of itself overwhelming – especially when considered in the light of the current policy, which, though familiar and easy-to-use, has a devastating consequence. More than 1.6 million wrong grades. Every year. Perhaps the 'simplicity' of the current rank order is too high a price to pay.

This example illustrates a more fundamental, general, and important point. Any solution to the grade reliability problem will require a carefully planned, suitably resourced, and well-executed programme to inform the stakeholder community. This will take effort, people's time and energy, and money. These needs create a barrier, and provide any number of pretexts whereby any proposed change can be rejected, and the *status quo* maintained.

It could be the case that the *status quo* should indeed be maintained. But that would be valid only on the grounds that, after a detailed, thorough and balanced evaluation of all the possibilities, there is widespread agreement that maintaining the *status quo* is the best option – and not just because organising a communications programme is too much like hard work, or some vague claim that 'stakeholders won't like it'.

That is why, alongside the statistical analysis of fuzziness, a vital next step is the wise evaluation of the various solutions outlined in this paper – and indeed all others that can be identified.

Appendix

The statistics of marking and re-marking

Why a statistical analysis is needed

This Appendix presents the relevant mathematics and statistics of examination marking and re-marking. If the marking of GCSE, AS and A level scripts were precise, such that the same mark would be given to the same script by all examiners (as is the case for examinations based on unambiguous multiple-choice questions), then no statistical analysis would be needed: any originally-given mark m would be confirmed by the re-mark $m*$ given by any other examiner. For examinations largely structured around more open-ended questions, and especially for those that require essays as answers, then a re-mark $m*$ by another examiner, as equally qualified and as equally conscientious as the first, might result in the same mark m as the original mark, but might not: the re-mark $m*$ might be a number of marks higher than the original mark m, or it might be a number of marks lower.

Given that, for any original mark m, there are a number of different possible values that the re-mark $m*$ might take, any questions concerning any relationships between the original mark m and the re-mark $m*$ have to be expressed in probabilistic terms, as exemplified by questions such as:

- For a given value of the original mark m, what is the probability that the re-mark $m*$ will be the same as the original mark m?
- For a given value of the original mark m, what is the probability that the re-mark $m*$ will be h marks different from the original mark m, such that $m* = m + h$? So, for example, for an original mark $m = 59$, what is the probability that the re-mark $m*$ is 61, two marks higher (implying that $h = 2$ so that $m* = m + h = 59 + 2 = 61$)?

Since these questions enquire about probabilities rather than certainties, any answers to these questions must be based on a statistical analysis of marking and re-marking, as presented here. Much of the analysis is therefore mathematical, and so the discussion presented assumes some familiarity with mathematics, and mathematical symbols and representations. Sometimes a symbol will be used to represent a quantity, or variable, in general: so, for example, the symbol m represents any mark that might be given to any script by any examiner. There are occasions, however, when it is helpful to represent a specific instance of that quantity, in which case the variable symbol will be associated with the † symbol: accordingly, the composite symbol m† represents a specific mark (say, 59) given to a particular script.

Probability distributions

Measuring probabilities - the probability distribution $t(m)$

Suppose that a single script is marked once by each of 75 different equally-qualified and equally-conscientious examiners. Suppose further that 6 examiners give a mark $m = 56$, 10 give $m = 57$, and 9 give $m = 61$. The overall outcome for all 75 examiners is shown Table A1.

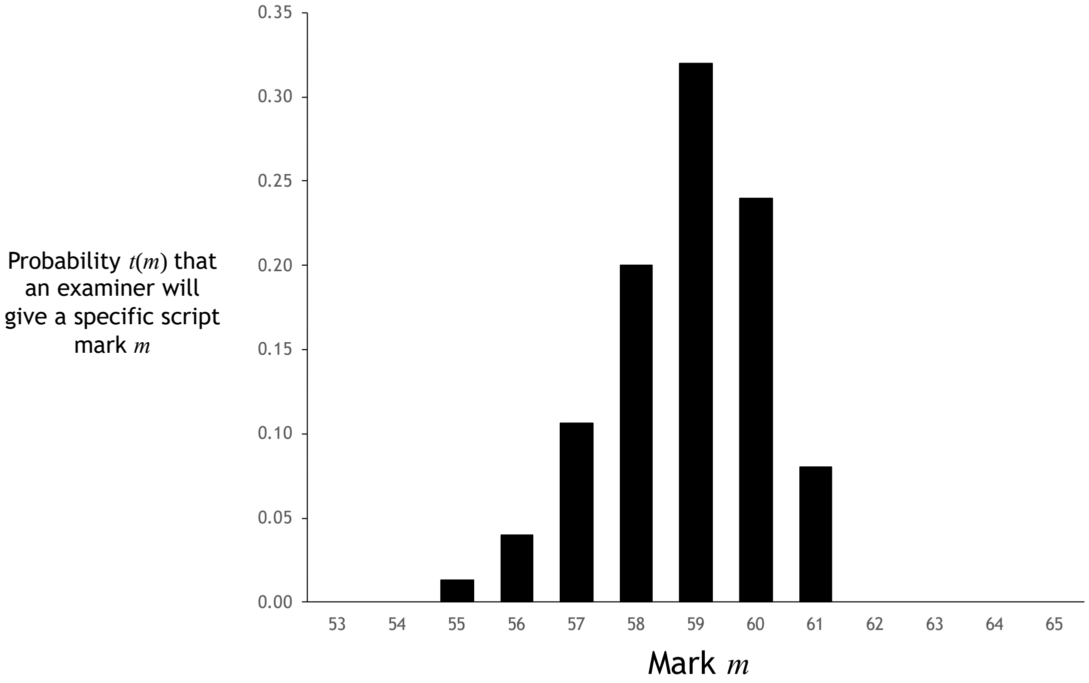
Table A1: Marks given by 75 different examiners to the same script

Mark m	Number of examiners giving mark m		Percentage of examiners giving mark m	Probability $t(m)$ that an examiner will give mark m
	Actual	Cumulative		
≤ 53	0	0	0.00%	0.0000
54	0	0	0.00%	0.0000
55	1	1	1.33%	0.0133
56	3	4	4.00%	0.0400
57	8	12	10.67%	0.1067
58	15	27	20.00%	0.2000
59	24	51	32.00%	0.3200
60	18	69	24.00%	0.2400
61	6	75	8.00%	0.0800
62	0	75	0.00%	0.0000
≥ 63	0	75	0.00%	0.0000
Total	75	75	100.00%	1.0000

In this table, the percentages are calculated based on the total of 75 = 100%, and the probabilities are defined by reference to the corresponding percentages, but expressed as a number between 0 and 1.

If 100 further examiners were to mark that script, what marks would be given? This question cannot be answered with certainty, but if the new examiners are as well-qualified and as conscientious as each of the previous 75, then the data in Table A1 suggests that it is extremely unlikely (but none the less still possible) that any new mark would be 54 or lower, likewise 62 or higher; a reasonable inference is that about 20 would give 58, and about 32 would give 59, in accordance with the probabilities as shown. The set of probability figures defines a ‘probability distribution’, as represented graphically by the histogram shown in Figure A1.

Figure A1: The probability distribution $t(m)$ for the data shown in Table A1



Formally, this distribution is described by a 'mathematical function' $t(m)$, where the value of $t(m)$ for any specific mark m is as shown in Table A1, and as represented by the height of the corresponding column in Figure A1. Distributions of different shapes will be associated with different functions, all of which have different shapes, but all generically written as $t(m)$.

A feature of a distribution of probabilities is that the sum of all the column heights is 1.00, or 100% – expressed mathematically as

$$\sum_m t(m) = 1$$

In this expression, the symbol $\sum_m$ indicates a summation over all possible values of m. In principle, this range of marks extends from 0 to 100; since in this particular example the values of $t(m)$ are all zero for values of $m \leq 54$ and $m \geq 62$, the effective range of the summation is from $m_{min} = 55$ to $m_{max} = 61$.

Since a probability of 1 = 100% is a certainty, the 'real world' interpretation of this is that there is in essence an absolute certainty that a given mark m is within the range from $m_{min} = 55$ to $m_{max} = 61$, and that the probability that a mark m might be outside this range is less than, say, 0.0001 = 0.01%.

Three different measures of a distribution's centre

For any distribution, it is helpful to identify a measure of a 'representative' member of that distribution, and so statisticians define

- the **mode** M;
- the **mean** $\langle \mathrm{M} \rangle$;
- the **median** $\mathbf{M}$.

Each of these specify a single number towards the centre of the corresponding distribution, and with reference to the data shown in Table A1, and as illustrated in Figure A1:

- The **mode** M corresponds to the mark m given by more examiners than any other mark, as identified by the peak of the corresponding distribution. Accordingly, for the example shown, $\mathrm{M} = 59$.
- The **mean** $\langle \mathrm{M} \rangle$ is the arithmetical average, defined mathematically as

$$\langle \mathrm{M} \rangle = \frac{\sum_m m\ t(m)}{\sum_m t(m)}$$

  in which the product $m\ t(m)$ weights each mark m by the probability $t(m)$ of that mark's occurrence. For the example shown, the mean $\langle \mathrm{M} \rangle$ computes to $\langle \mathrm{M} \rangle = 58.13$.

- The **median** $\mathbf{M}$ is the 'half-way' mark, defined such that this mark is equal to, or greater than, that given by one-half of the examiners; by the same token, it is also the mark equal to, or less than, that given by the other half of the markers. Operationally, the median can be determined by listing all the individual examiners, and the corresponding mark given, in ascending order of the mark, and then identifying the mark given by the examiner in the middle of resulting list. In the example shown in Figure A1, there were 75 examiners: the 'middle' examiner is therefore the 38th, and, as can be seen from the 'cumulative' column in Table A1, among the 24 examiners who gave the script 59 marks. The median of the distribution shown in Figure A1 is therefore $\mathbf{M} = 59$.

In this example, the median $\mathbf{M} = 59$ happens to be have the same value as the mode $\mathrm{M} = 59$, but a value different from the mean $\langle \mathrm{M} \rangle = 58.13$. For some distributions, all three measures have the same value, in which case a graphical representation of the distribution is left-right symmetrical. For some distributions, the median $\mathbf{M}$, mode M and mean $\langle \mathrm{M} \rangle$ have different values, in which case a graphical representation of the distribution is skewed, to a greater or lesser extent, with the mode M either towards the right (as is the case for the distribution shown in Figure A1), or towards the left.

Two different measures of a distribution's width

The median $\mathbf{M}$, mode M and mean $\langle \mathrm{M} \rangle$ are three different measures of the centre of a distribution, but any one of these measures, though informative, gives no indication of the distribution's shape - and in particular, whether the distribution is narrow or broad. This is important, for the median $\mathbf{M}$ is much more informative when associated with a measure of the corresponding distribution's width than as a number by itself. As an example, if a distribution has a median $\mathbf{M} = 59$, and a minimum mark of 55 and a maximum mark of 61, then all the marks are closely clustered around the median $\mathbf{M} = 59$; in contrast, a different distribution, also of median $\mathbf{M} = 59$, but with a minimum of 45 and a maximum of 71, is much broader. If the only knowledge is that the median $\mathbf{M} = 59$, then the range of marks might be from 58 to 60 - or from 8 to 100.

Accordingly, two measures of the width of a distribution are

- the **standard deviation**, σ; and
- the **end-to-end range** N.

The **standard deviation** σ is defined mathematically as

$$\sigma^2 = \frac{\sum_m (m - \langle \mathrm{M} \rangle)^2 \ t(m)}{\sum_m t(m)}$$

In this expression, the difference $(m - \langle \mathrm{M} \rangle)$ represents the distance between any mark m and the mean $\langle \mathrm{M} \rangle$, and so is a larger number for a mark m further from the mean than for a mark m closer in. Since the difference $(m - \langle \mathrm{M} \rangle)$ can be both positive (for marks m greater than the mean $\langle \mathrm{M} \rangle$) and negative (for marks m smaller than the mean $\langle \mathrm{M} \rangle$), the square $(m - \langle \mathrm{M} \rangle)^2$ is always positive. The standard deviation σ therefore represents a measure of the average actual distance of a mark m from the mean $\langle \mathrm{M} \rangle$, this being a measure of the width of the corresponding distribution. For the example shown in Figure A1, σ computes to 1.313.

The end-to-end range N is simpler to identify and compute: any distribution of marks will extend from a minimum mark m_{min} to a maximum mark m_{max}, and the end-to-end range N is defined as

$$N = m_{max} - m_{min}$$

In the example shown in Figure A1, $m_{min} = 55$ marks and $m_{max} = 61$ marks, from which $N = 61 - 55 = 6$ marks.

The **end-to-end range** N is a measure of the distance between m_{min} and m_{max}, the total range of marks over which the distribution extends. Note that a count of the number of individual marks included in the distribution is always $N + 1$, one mark greater than the end-to-end range N - in this example, the end-to-end range $N = 6$ marks, but there are $N + 1 = 7$ marks included in the distribution itself (55, 56, 57, 58, 59, 60 and 61).

Other representations of the distribution $t(m)$

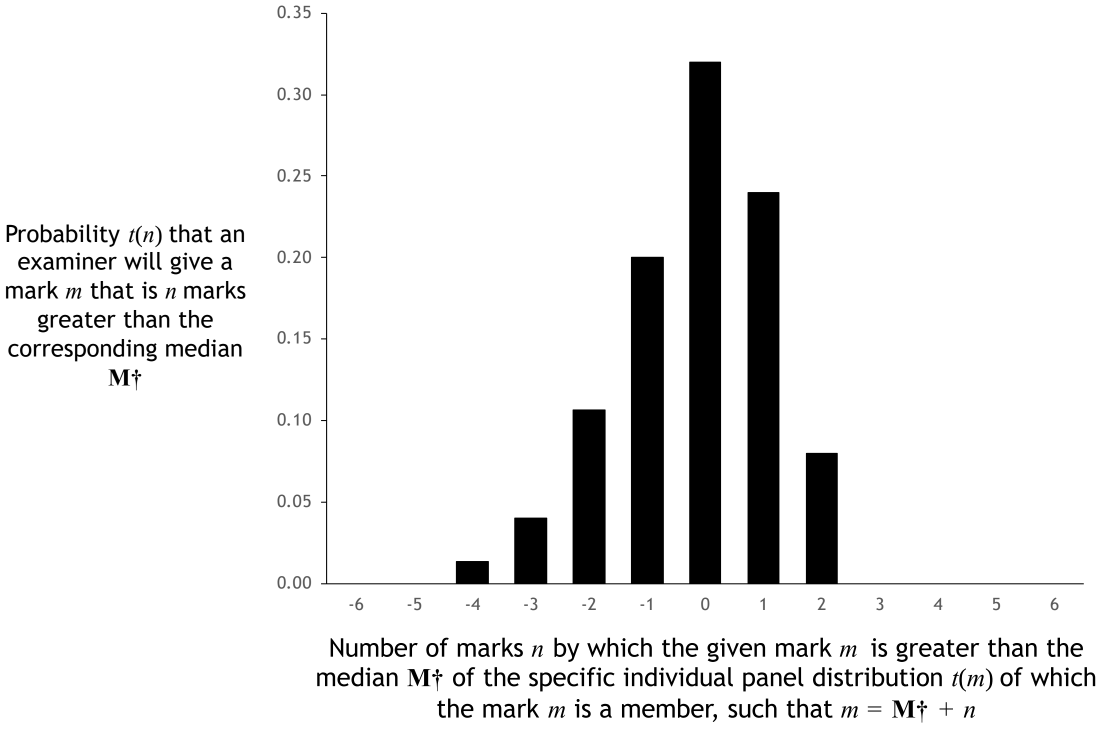
As shown in Figure A1, the distribution $t(m)$ extends from $m_{min} = 55$ marks to $m_{max} = 61$ marks, with median $\mathbf{M}† = 59$ marks, where the composite symbol $\mathbf{M}†$ indicates that this median is specific, being the median of that particular distribution $t(m)$ of which the mark m is a member.

If a new variable n is defined such that

$$m = \mathbf{M}† + n$$

then n represents the number of marks by which any mark m is greater than the median $\mathbf{M}†$ of the distribution $t(m)$ of which the mark m is a member. So, for example, $m_{max} = 61$ corresponds to $n = 2$ $(61 = 59 + 2)$, and $m_{min} = 55$ corresponds to $n = -4$ $(55 = 59 - 4)$. For the distribution of Figure A1, for every value of m from $m_{min} = 55$ to $m_{max} = 61$, a total end-to-end range $N = 61 - 55 = 6$ marks, there is a corresponding value of n from $n_{min} = -4$ to $n_{max} = 2$, this being a total range of $2 - (-4) = 6$ marks $= N$ also. The distribution represented as $t(n)$, expressed in terms of the variable n rather than the variable m, therefore has the same shape as the distribution $t(m)$, but extends from $n_{min} = -4$ to $n_{max} = 2$, with a median $\mathbf{M} = 0$, as shown in Figure A2.

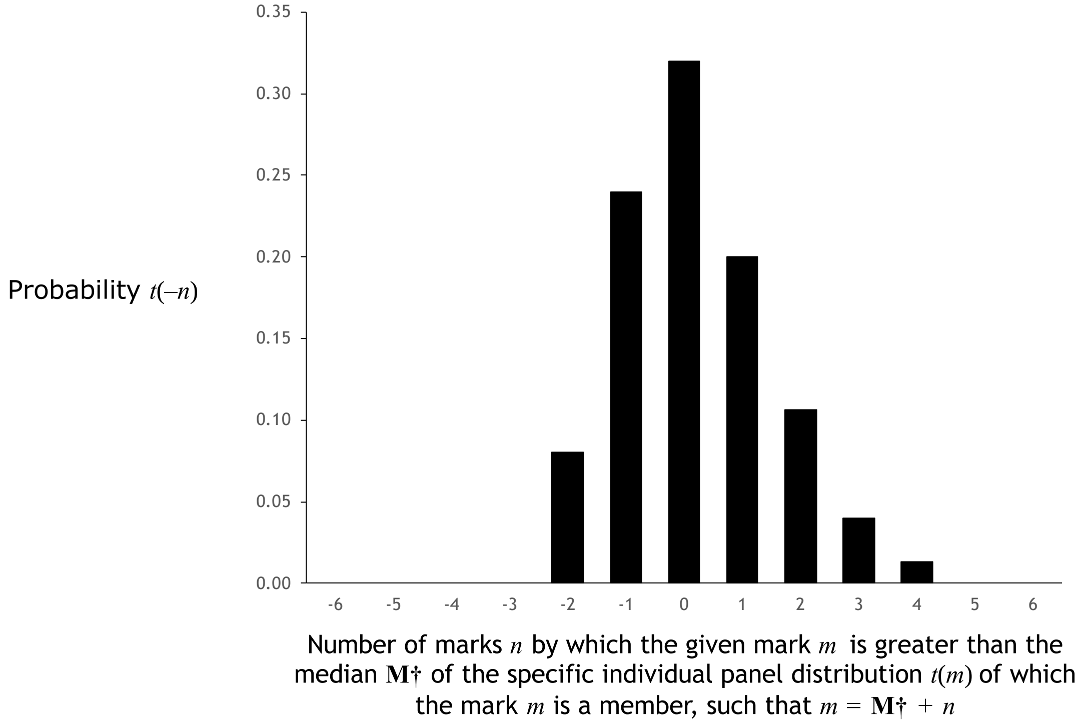
Figure A2: The distribution $t(n)$



To verify this, consider the specific value $n = -3$, for which, according to Figure A2, $t(n) = t(-3) = 0.04$. Since $m = \mathbf{M}\dagger + n$, then, for $\mathbf{M}\dagger = 59$, $m = 59 - 3 = 56$. According to Figure A1, $t(m) = t(56) = 0.04$, so demonstrating that $t(n) = t(m)$. This is also true for all other values of n, and the corresponding values of m, so proving that the distributions $t(m)$ and $t(n)$ have identical shapes, but with $t(m)$ straddling the median $\mathbf{M}\dagger = 59$, as shown in Figure A1, and $t(n)$ straddling the median $\mathbf{M} = 0$, as shown in Figure A2.

One further distribution is of interest, that represented by the function $t(-n)$. To determine the shape of $t(-n)$, consider the specific value $n = +1$. When $n = +1$, the value of $t(-n)$ is given by the corresponding value of $t(-1)$ as shown in Figure A2 (and Table A1), namely 0.24. The same applies to all other values of n, and so the shape of the distribution $t(-n)$ is as shown in Figure A3, which, as can be seen, is the left-right mirror image of the shape of the distribution $t(n)$ as shown in Figure A2.

Figure A3: The distribution $t(-n)$



If the distribution $t(n)$ is left-right symmetrical (as, in practice, it often is), then the shapes of the two distributions $t(n)$ and $t(-n)$ are indistinguishable, for the symmetry of $t(n)$ implies that it is its own left-right mirror image; if, however, $t(n)$ is not symmetrical (as in Figure A2), then $t(n)$ and $t(-n)$ can be distinguished, as shown by comparing Figures A2, for $t(n)$, and A3, for $t(-n)$.

As will be shown, the distributions $t(n)$ and $t(-n)$ play an important role in the statistics of marking and re-marking, and provide the mathematical foundations of the measurement of grade reliability.

Three important probability distributions

Three statistical probability distributions play an especially important role in the analysis of marking and re-marking. These are briefly introduced here; each will be discussed in more detail later:

- The **generic panel distribution**, represented mathematically as $T(n)$. This distribution defines the distribution of marks given to the same script by each examiner drawn from a panel of equally-qualified, equally-conscientious, examiners. This distribution answers the question "If a number of different examiners were each to mark the same script, what is the probability that the mark m given by any one examiner is n marks greater than the median $\mathbf{M}\dagger$ of the distribution of all marks given to that script, such that $m = \mathbf{M}\dagger + n$?". In this question, the parameter n may take both positive and negative values, as well as a value of zero, so that the mark m can be greater than, less than, or equal to the median mark $\mathbf{M}\dagger$. This median $\mathbf{M}\dagger$ is important in that, as will be discussed on pages 99 and 100 , it can be used to define the 'right' mark for any given script.

- As will be shown, an important feature of the statistics of marking is that a script given a specific mark $m\dagger$ by a single examiner can be a member of any one of a number of different generic panel distributions, each with its own median $\mathbf{M}_p$. In practice, this implies that knowledge of an originally-given mark m does not give sufficient information to determine unambiguously the median $\mathbf{M}\dagger$ of the specific generic panel distribution of which the mark m is a member. The **special re-mark distribution**, represented mathematically as $Q(p)$, answers the question "What is the probability that the specific single mark $m\dagger$ is a member of the generic panel distribution of median $\mathbf{M}_p$ such that $\mathbf{M}_p = m\dagger + p$?". The significance of this distribution is that it defines the probability that a mark $m\dagger$ is associated with a particular median $\mathbf{M}_p$. If the median $\mathbf{M}_p$ is the 'right' mark, this in turn defines the probability that the 'right' mark corresponding to an original mark m is $\mathbf{M}_p = m\dagger + p$. Furthermore, the distribution $Q(p)$ defines the probability that a script, originally given the specific mark $m\dagger$, would be re-marked $\boldsymbol{m*} = m\dagger + p$ by a senior examiner – hence the description of $Q(p)$ as the special re-mark distribution.

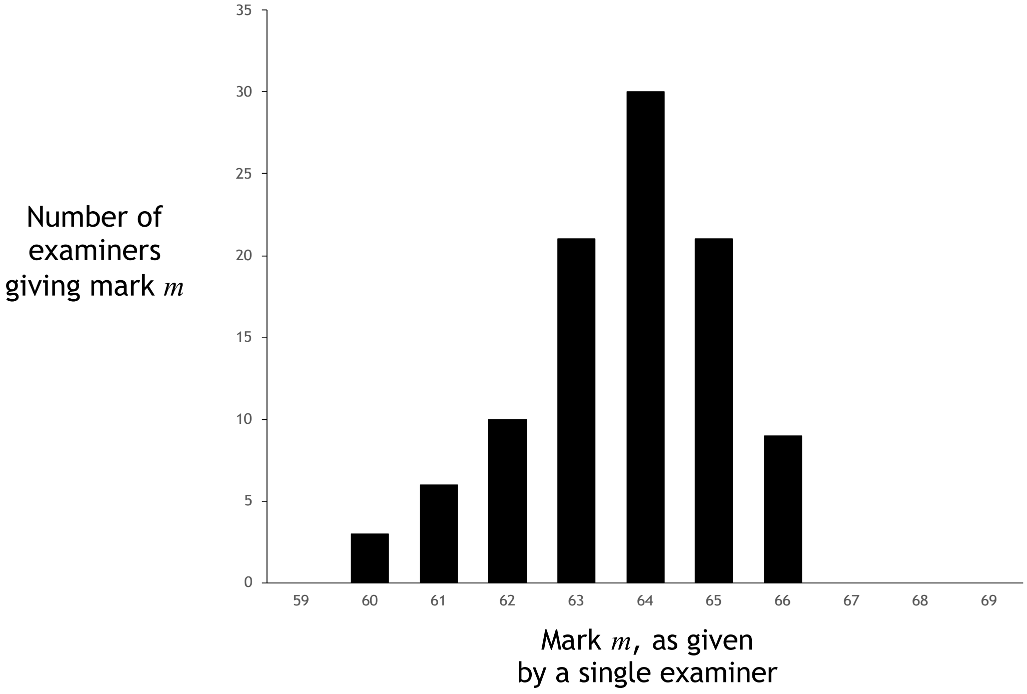
- The **ordinary re-mark distribution**, represented mathematically as $r(h)$, answers the question "If a script originally given the specific mark $m\dagger$ is re-marked $m*$ by any examiner (and so not only by a senior examiner), what is the probability that the re-mark $m*$ will be h marks different from the original mark m, such that $m* = m\dagger + h$?". As will be shown, the ordinary re-mark distribution $r(h)$, which is defined by reference to a re-mark $m*$ by any examiner, is (importantly) different from, and broader than, the special re-mark distribution $Q(p)$ resulting from re-marking that same script by a senior examiner. It is the special re-mark distribution $Q(p)$ that explains Ofqual's research, all of which was based on a comparison to the 'definitive' mark given by a senior examiner; it is the ordinary re-mark distribution $r(h)$, however, that provides a realistic method of measuring grade reliability in practice.

The analysis starts, however, with a statistical discussion of the difficulties of determining the 'right' mark.

Which mark is 'right'?

As has been mentioned several times, for all examinations, other than those structured as right/wrong multiple-choice questions, it is possible that different, equally qualified, examiners might award different marks m to the same script. If, for example, 100 examiners each mark the same script once, the marks given will form a distribution such as that shown in Figure A4 (which is superficially similar in shape to the distribution shown in Figures A1 and A2, but is in fact different).

Figure A4: A representative individual panel distribution $t(m)$



As can be seen, 64 is the most 'popular' mark, given by 30 examiners; 9 examiners give the highest mark, 66; the lowest mark, 60, is given by 3 examiners.

Figure A4 shows the distribution of marks m given by each of 100 different examiners to the same script; strictly speaking, however, Figure A4 does not show a *probability* distribution, for the vertical axis represents the number of examiners who actually gave the script the mark m; furthermore, the sum of all the columns is 100, the total number of examiners. By contrast, the vertical axis of Figure A2 shows a mathematical probability, and the sum of all the columns is 1. Corresponding actual and probability distributions have the same shape, and the one may be derived from the other by adjusting the vertical axis according to the total population: given the actual distribution, the probability distribution is obtained by dividing by the total population; given the probability distribution, the actual distribution is obtained by multiplying by the total population.

For a specific script, the distribution obtained (whether the distribution of actual marks, or the corresponding probability distribution) will be referred to as the ***individual panel distribution*** $t(m)$ – 'individual' because this distribution relates to one, specific, individual script; this is in contrast to the ***generic panel distribution***, which, as will be seen in the next section, relates to any script for the given examination subject.

The individual panel distribution illustrated in Figure A4 happens not to be left-right symmetrical, but in practice it often is. Whatever the shape might be, as discussed on pages 92 to 94, the distribution is always associated with a number of statistical characteristics, for example, for the distribution shown in Figure A4:

- The **mode** M, as shown by the peak in the distribution, 64 marks.
- The **mean** $\langle \mathrm{M} \rangle$, in this case 63.68 marks.
- The **median** $\mathbf{M}$, which, in this example, is 64, the same as the mode.
- The **standard deviation** σ, which, for the distribution shown in Figure A4 computes as 1.45 marks.
- The **end-to-end range** N, the range of marks from the lowest mark m_{min} = 60 to the highest mark m_{max} = 66, given by the difference $m_{max} – m_{min}$ = 66 – 60 = 6 marks.

This distribution shown in Figure A4 represents the marks given by 100 examiners, each marking the same script once. Following the same line of reasoning as on page 90, if another examiner were to mark that script, it is highly unlikely that the mark will be lower than m_{min} = 60 or higher than m_{max} = 66; furthermore, since 30 examiners of the original 100 gave 64 marks, there is 30% probability that this additional examiner will also gave 64 marks; likewise a 10% chance of 62 marks.

The individual panel distribution $t(m)$ shown in Figure A4 can therefore be used to determine the probability that any suitably qualified examiner will give the script any particular mark. No mark is favoured, or 'special' – it really is a lottery as to which mark is actually given, with some marks (such as 64) being more likely than others (such as 61).

This range of marks creates a problem if a single mark has to be chosen as a measure of the candidate's assessment, this being the mark that determines the grade that appears on the candidate's certificate and therefore widely accepted as the 'right' mark by all who might take that grade into consideration when making a decision, such as the offer of an apprenticeship, a job or a place at a college or university.

Is the 'right' mark the mark that happens to be given by one examiner who, by chance, happens to mark the script – which could be any mark from 60 to 66 – as is the current policy for awarding grades?

Is the 'right' mark that given by a 'special' examiner, such as a senior examiner? If it is, and if the senior examiner's mark is, say, 61, then an inference from Figure A4 is that there is only about a 6% chance that an ordinary examiner would give this mark. Perhaps it would be fairer to the candidates if marking were done only by senior examiners – but even then there must be assurance that all senior examiners always agree, and that the distribution of marks is always a 'spike' at a single mark, rather than a distribution, albeit probably narrower than the distribution shown in Figure A4.

Or is the 'right' mark one of the characteristics of the distribution, such as the mode M, the mean $\langle \mathrm{M} \rangle$, the median $\mathbf{M}$, the highest mark m_{max}, or the lowest m_{min}? If it is one of these, then it appears that the distribution needs to be determined first, but for a public examination, marking every individual script in the cohort multiple times is a huge amount of work, and so totally impracticable.

Perhaps, though, it might be possible to use statistics to help. Suppose, for example, that a script is given a single mark, say, 63. If it were possible to estimate that there is, say, about a 20% probability that a mark of 64 is the median of the individual panel distribution $t(m)$ of which this mark is a member, then that might be quite informative.

Deciding which single mark is 'right' is problematic, but supposing for the moment that defining a particular single mark as 'right' might be useful, perhaps it does not matter which single number is chosen from the individual panel distribution $t(m)$, provided that three conditions are simultaneously fulfilled:

- The number chosen must be uniquely representative of the individual panel distribution $t(m)$ with which it is associated.
- That number must be reproducible, in that, for any specific script, the same number must be obtained from all possible individual panel distributions $t(m)$, as generated by using different panels of suitably qualified examiners.
- The principle that defines the chosen number must be used consistently for all candidates.

According to the first of these conditions, the individual panel distribution $t(m)$ could, in principle, be represented by, for example, the mean $\langle \mathrm{M} \rangle$, the median $\mathbf{M}$, the highest mark m_{max}, or the lowest mark m_{min}. The mode M, however, must be excluded since, if the distribution is somewhat flat, or if there are two or more equally high 'humps', there is more than one mode, and so the mode is not uniquely defined as a single mark.

The second condition, reproducibility, is fulfilled by the definition of the individual panel distribution $t(m)$ as being a distribution that is independent of the examiners. In practice, however, there is the possibility that different sets of examiners might result in slightly different distributions, especially as regards the low-end and high-end 'outliers', so implying that m_{max} and m_{min} are unsuitable. According to various academic studies[4], the median is more stable with respect to outliers than the mean, and so it is the median $\mathbf{M}\dagger$ that this paper will use as representative of the corresponding individual panel distribution, where the composite symbol $\mathbf{M}\dagger$ emphasises that this is the specific median of the single individual panel distribution of which the given mark m is a member. The third condition is then easily fulfilled – if the median of the every candidate's individual panel distribution $t(m)$ is chosen as the basis of grading, then all candidates are being treated fairly.

For any script, and the corresponding individual panel distribution $t(m)$, the selection of the median $\mathbf{M}\dagger$ as the mark that determines the candidate's grade does not imply that the median is the 'right' mark. What is, or is not, the 'right' mark is of no consequence: the important point is that there is a mark which acts as a representative of the corresponding individual panel distribution $t(m)$, and that this mark is used consistently for all scripts.

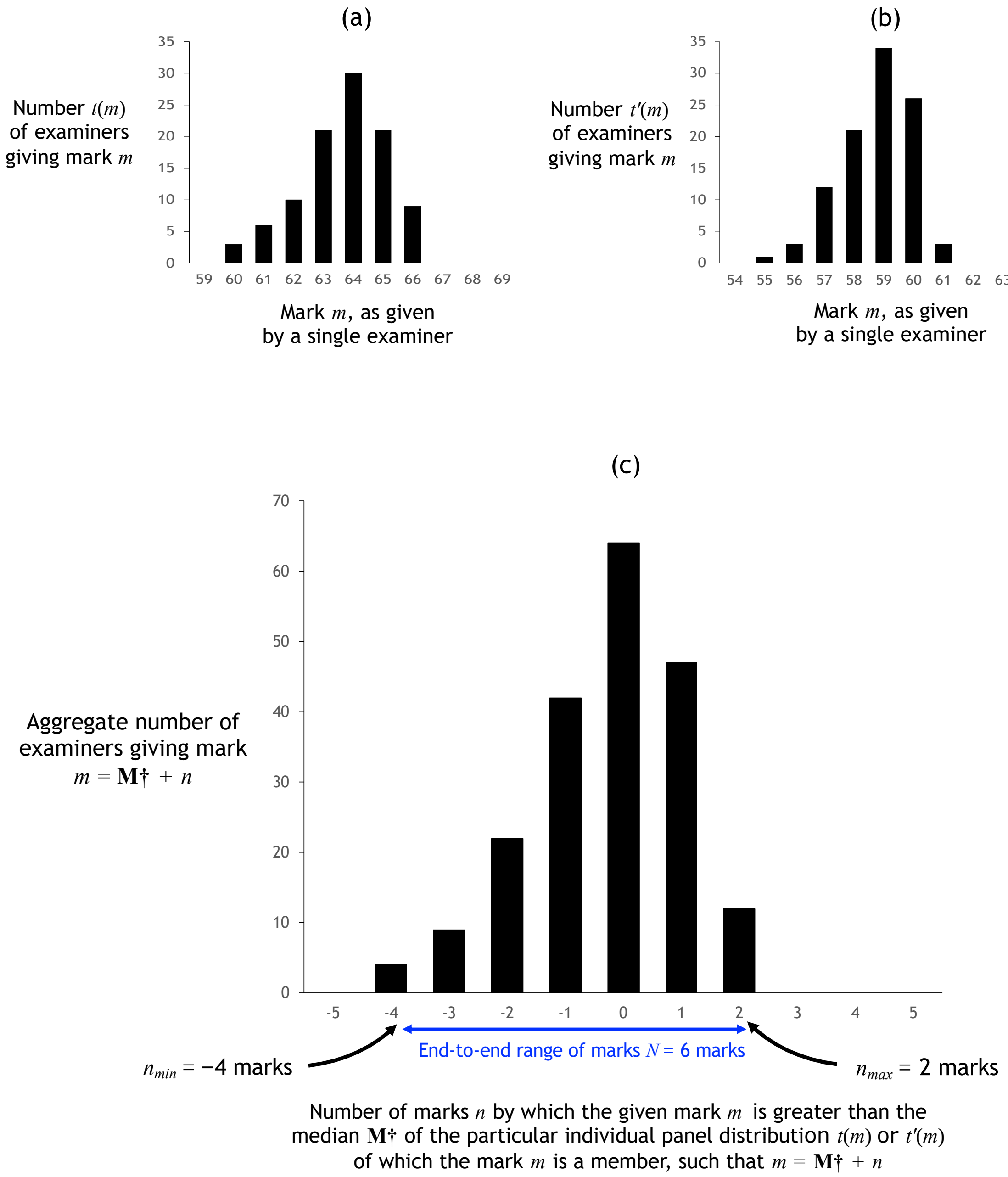
A central theme of Ofqual's November 2016 and November 2018 reports, however, is the use of a senior examiner's mark as a reference point, defining the 'definitive' mark and the corresponding 'definitive' grade – and Figures 12 and 13 of the November 2016 report even refer to the 'true grade'. In the absence of any other information, this paper will assume that the senior examiner's mark corresponds to the median of the corresponding individual panel distribution $t(m)$.

The generic panel distribution $T(n)$

The individual panel distribution $t(m)$ shown in Figure A4 refers to a single, specific, script. Suppose that a second script is randomly chosen, and also marked by a panel of 100 examiners, so generating a second, different, individual panel distribution $t'(m)$, shown as (b) on the upper right-hand side of Figure A5.

[4] See, for example, RS Pindyck and DL Rubinfeld, *Econometric Models and Economic Forecasts* (4th edition, 1998), Irwin/McGraw Hill, p 47.

Figure A5: Aggregating two individual panel distributions $t(m)$ and $t'(m)$



The individual panel distribution $t(m)$ shown in Figure A5(a) is the same as that shown in Figure A4, with a median $\mathbf{M}†$ = 64; A5(b) is the individual panel distribution $t'(m)$ for a second script, with median $\mathbf{M}†$ = 59, and although different in detail, the two distributions are quite similar in shape. If these two distributions are shifted along the horizontal axis so that they both have a median $\mathbf{M}$ = 0, the two distributions will overlap, and can be added, resulting in the distribution shown in Figure A5(c).

In Figure A5(c), the definition of the horizontal axis has changed from 'Mark *m*, as awarded to a single examiner' to 'Number of marks *n* by which the given mark m is greater than the median $\mathbf{M}\dagger$ of the particular individual panel distribution $t(m)$ or $t'(m)$ of which the mark m is a member, such that $m = \mathbf{M}\dagger + n$'. This is a consequence of the shift of each individual panel distribution to a common median $\mathbf{M} = 0$, and the parameter n defines the number of marks by which a mark m is greater than the median $\mathbf{M}\dagger$ of the particular individual panel distribution of which that specific mark m is a member, where n can be positive (implying that the mark m is greater than the corresponding median $\mathbf{M}\dagger$), negative (m is less than $\mathbf{M}\dagger$), or zero (m is equal to $\mathbf{M}\dagger$).

Suppose that this process is carried out for 10 randomly selected scripts, so giving a total of 10 individual panel distributions of the type shown in Figures A5(a) and A5(b). Each of these 10 distributions has its own median, and its own shape, but it is likely that the shapes will be similar. If each of these 10 distributions is shifted to a common median of 0, they can then be added, resulting in an aggregate distribution like that shown in Figure A5(c), but representing 10 contributing individual panel distributions, rather than just two.

The total number of scripts marked is 1,000, corresponding to 100 examiners for each of 10 scripts, and the resulting histogram, the equivalent of Figure A5(c), would show a number of columns (say, seven, as in Figure A5), and the height of each column would show the numbers of scripts given the median mark (corresponding to $n = 0$); one mark greater than the corresponding median ($n = 1$); one mark lower from the corresponding median ($n = -1$); and so on for each integral value of n from $n_{min} = = -4$ to $n_{max} = 2$, such that the total of the heights of all the columns is equal to the total number of scripts marked, 1,000. If the height of each column is divided by 1,000, the total of the heights of all the columns is then 1, and each column has a height represented by a number less than 1. The overall result is represented as the probability distribution shown in Figure A6, with the corresponding numerical values in Table A2.

Figure A6: The generic panel distribution $T(n)$

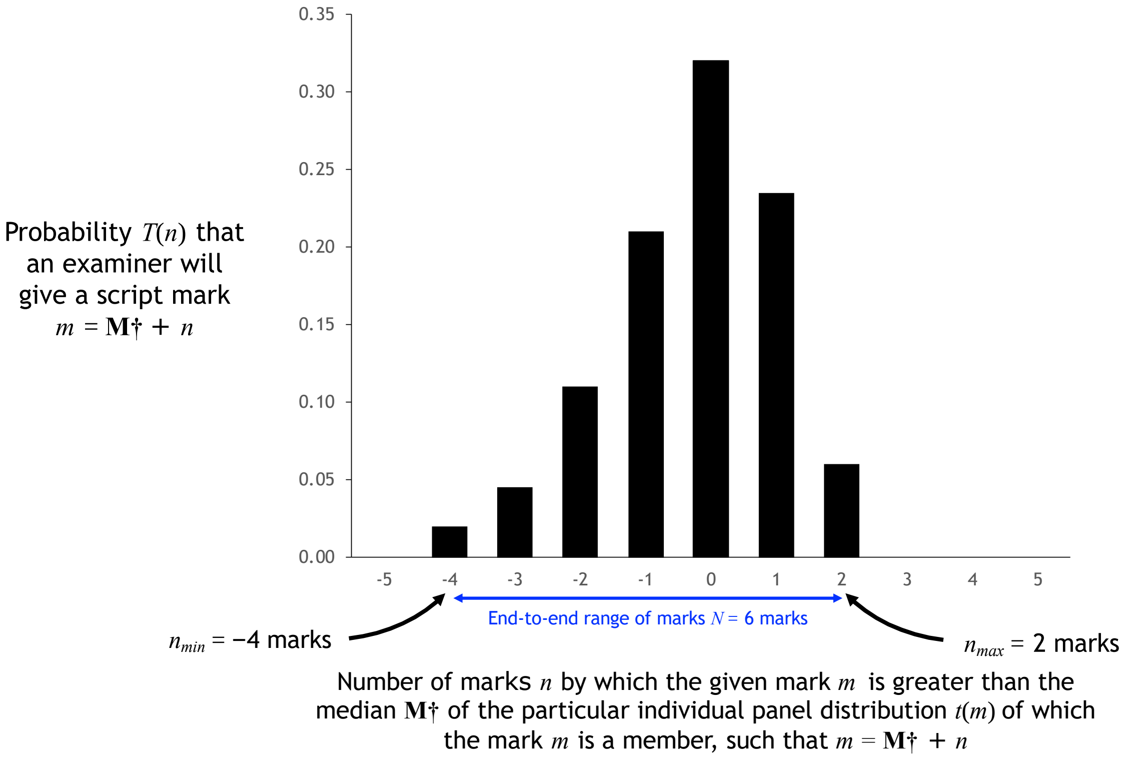


Table A2: Numerical values corresponding to the generic panel distribution $T(n)$ shown in Figure A6

n	Probability $T(n)$	
	Percentage	Numeric
≤ -5	< 0.1%	< 0.001
-4	2.0%	0.020
-3	4.5%	0.045
-2	11.0%	0.110
-1	21.0%	0.210
0	32.0%	0.320
1	23.5%	0.235
2	6.0%	0.060
≥ 3	< 0.1%	< 0.001
Total	100.0%	1.0000

As can be seen from the total in Table A2, the summation of all the probabilities $T(n)$ is 1.000 = 100%; in real terms, this means that it is virtually certain that any mark m will be within – 4 and + 2 marks of the median of the generic panel distribution associated with that mark. Mathematically, the distribution $T(n)$ is said to be 'normalised', as represented as

$$\sum_{n} T(n) = 1$$

where the symbol $\sum_{n}$ means 'add successive values of $T(n)$ for all values of n'. In principle, for an examination given standardised marks, n extends from – 100 and + 100; in practice, the probability $T(n)$ that a script will be marked tens of marks away from the associated median is in essence zero, and non-zero values will be within a relatively narrow range of values of n such as from – 4 to + 2 as in the current example.

In compiling Figure A6, the assumption has been made that each of the contributing individual panel distributions are 'different but similar' - and, in this particular case - each has a total end-to-end range of N = 6 marks, extending from n_{min} = – 4 to n_{max} = + 2.

This assumption is important, for it implies that the shape defined by Figure A6:

- is sensibly representative of the examination as a whole;
- is independent of the examiners; and
- can be applied to all scripts.

In fact, there are two circumstances in which the first of these conditions breaks down: for very low marks, and for very high marks. On a standardised mark scale, no script can be given a mark less than zero, and so the individual panel distribution for a script given a mark of say, 1, 2 or 3, by any one examiner is likely to be truncated on the left. Similarly, no script can be given a mark greater than 100, and so the individual panel distribution for a script given a mark in the high 90s by any one examiner is likely to be truncated on the right. These extreme individual panel distributions are therefore likely to be narrower than any others, and more skewed. Very few scripts, however, are given such low or high marks, and so, for the purposes of this paper, these distributions will be regarded as 'outliers', and ignored.

Accordingly, this paper will continue to assume that the three conditions mentioned above hold for the vast majority of scripts. As noted on page 84, however, it is important that this assertion is verified by a detailed statistical analysis; but if the three conditions can be accepted as valid, then, as is about to be shown, it unlocks the statistics of marking.

The distribution illustrated in Figure A5 will be referred to as the ***generic panel distribution***, for it refers to the examination as a whole, so distinguishing this distribution from any one script's individual panel distribution. Mathematically, this distribution may be represented as a function $T(n)$ of the generalised parameter n. As well as having a defined shape, an important characteristic of any generic panel distribution is its end-to-end range, represented as N marks, such that $N = n_{max} - n_{min}$, which in this example is $N = n_{max} - n_{min} = 2 - (- 4) = 6$ marks.

Each subject examination has its own generic panel distribution $T(n)$, implying that if, for any particular examination, its shape can be determined - for example, by using statistically valid samples - then that same shape can be used as a surrogate for the individual panel distribution for any individual script given any specific mark. Furthermore, the end-to-end range N of any examination's generic panel distribution $T(n)$ correlates with that examination subject's fuzziness: the value of N for a more fuzzy subject such as History will be considerably greater than the value of N for a less fuzzy subject such as Chemistry.

As an example of how knowledge of the generic panel distribution for a particular examination subject can be used, Figure A6 and Table A2 imply that:

- The probability that a mark m given to any script is the median mark $\mathbf{M}\dagger$ is 32%, corresponding to $n = 0$.
- If the mark m given to any script is known (say, 54), then there is an 11% probability that this mark is 2 marks lower than the median mark $m = \mathbf{M}\dagger + n$, corresponding to $n = -2$, and implying that $54 = \mathbf{M}\dagger - 2$, from which $\mathbf{M}\dagger = 56$...
- ...and, conversely, if the median mark $\mathbf{M}\dagger$ is known (say, 56), then there is an 11% probability that the script will be given a mark m that is 2 marks lower: $n = -2$ and so $m = \mathbf{M}\dagger + n = 56 - 2 = 54$.

If the definition of the 'right' mark is the median $\mathbf{M}\dagger$, then these inferences are important: they state, for this example, that there is a probability of 32% (about 1 chance in 3) that any script will be given the 'right' mark when marked by any examiner, drawn at random from the team of examiners, as happens under the grading policy in force at the time of writing. Even more important is what this does not say, at least explicitly: if there is about 1 chance in 3 that a script's mark is 'right', then there are about 2 chances in 3 that it is wrong.

The special re-mark distribution $Q(p)$

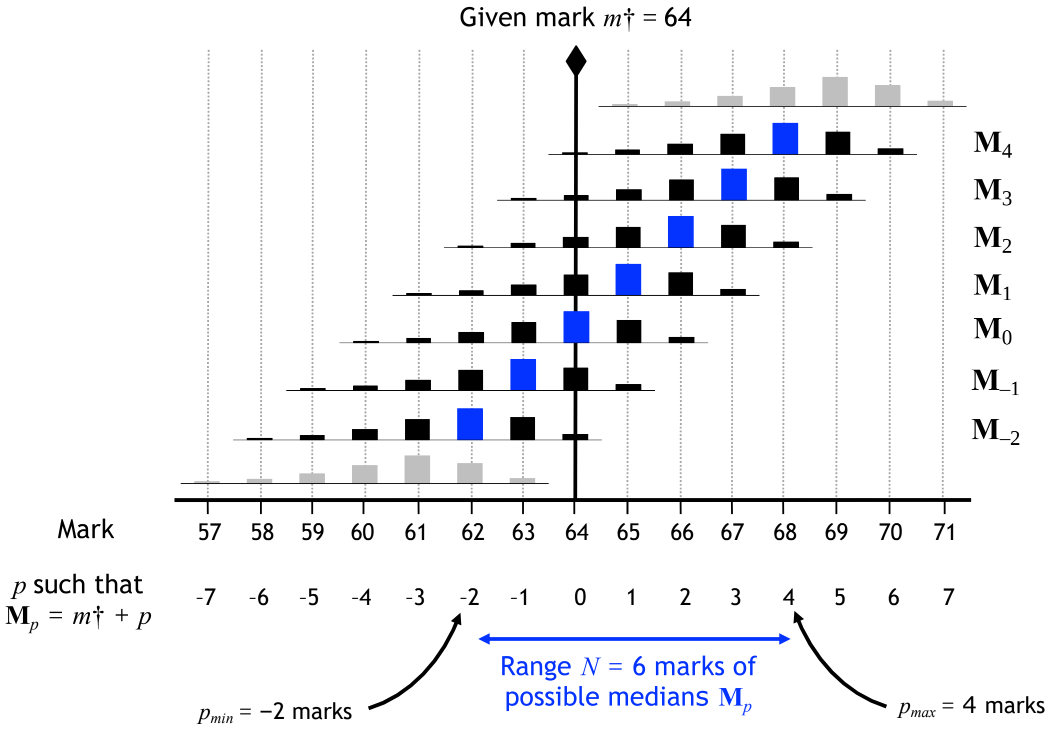
The medians $\mathbf{M}_p$

In practice, a single script is given a single valid mark m by a single examiner. Since, in principle, it is desirable to award the candidate the 'right' grade, and if it is agreed that the 'right' grade corresponds to the median $\mathbf{M}\dagger$ of the individual panel distribution of which the given mark m is a member, then it is clearly useful if that median $\mathbf{M}\dagger$ can be determined.

One way to determine the median $\mathbf{M}\dagger$ is for the script to be marked by a panel, and to compile the script's individual panel distribution – but that is expensive and impracticable. So might some statistics help?

At first sight, that appears to be impossible: if only the mark m is known, then the median $\mathbf{M}\dagger$ might be equal to the given mark m, but it might be higher, or it might be lower. It therefore seems that the median $\mathbf{M}\dagger$ might be any number, and that the problem is insoluble. But if the generic panel distribution $T(n)$ can be estimated by a sampling process, and if it is valid to assume that the generic panel distribution is a valid surrogate for any specific individual panel distribution, then the shape of $T(n)$ can be applied to any script, so limiting the possible values of $\mathbf{M}\dagger$, as represented in Figure A7.

Figure A7: The uncertainty of the medians $\mathbf{M}_p$ for the generic panel distributions, of the form shown in Figure A6, associated with the given mark $m\dagger$ = 64



Suppose that a script is given a specific mark $m\dagger$ = 64, where the composite symbol $m\dagger$ indicates reference to a specific mark given to a specific script by a single examiner. Suppose further that the generic panel distribution $T(n)$ for the subject examination takes the form shown in Figure A6. Because the generic panel distribution $T(n)$ can act as a surrogate for the individual panel distribution for this script, then the mark $m\dagger$ = 64 must be a member of that distribution. But since the generic panel distribution $T(n)$ shown in Figure A6 has an end-to-end width N of only 6 marks, that constrains the number of possible generic panel distributions that:

- have a shape defined by $T(n)$; and also
- contain the given mark $m† = 64$.

This is illustrated in Figure A7, which shows the given mark $m† = 64$, and also (rather vertically compressed) representations of the all the generic panel distributions of the shape shown in Figure A6, and with medians $\mathbf{M}$ from $\mathbf{M} = 61$ to $\mathbf{M} = 69$.

Since the given mark $m† = 64$ must be a member of its own generic panel distribution, it is extremely unlikely that this is the case for any generic panel distribution $T(n)$ for which the median $\mathbf{M} \leq 61$; likewise, for $\mathbf{M} \geq 69$. It is therefore almost certain that the median $\mathbf{M}†$ of the specific generic panel distribution of which the given mark $m† = 64$ is a member lies in the range $62 \leq \mathbf{M}† \leq 68$ This range is $68 – 62 = 6$ marks, the same as the end-to-end range N of the associated generic panel distribution $T(n)$.

Figure A7 identifies all these possibilities. The distribution $T(n)$ associated with the median $\mathbf{M} = 61$, as shown in grey at the bottom, is ruled out, for its end-to-end range does not include $m† = 64$; likewise, the distribution $T(n)$ associated with the median $\mathbf{M} = 69$, at the top. By contrast, The distribution $T(n)$ associated with the median $\mathbf{M} = 63$ does include $m† = 64$, and so it is possible that a script marked $m† = 64$ might be a member of this distribution, in which case the 'right' mark for that script is $\mathbf{M} = 63$. As Figure A7 vividly shows, however, this is not the only possibility: the distribution $T(n)$ associated with the median $\mathbf{M} = 67$ also includes $m† = 64$, and so the script's 'right' mark might also be $\mathbf{M} = 67$. As can be seen, a total of $7 = N + 1$ different distributions $T(n)$ include $m = 64$, and so the 'right' mark is constrained to one of the seven values from 62 to 68 inclusive.

For a mark $m† = 64$, as actually given to the script, any of the $7 = N + 1$ allowed values of the median $\mathbf{M}$ can be written as $\mathbf{M}_p$, where the parameter p is such that $\mathbf{M}_p = m† + p$. Accordingly, when $p = 2$, $m† + p = 54 + 2 = 56$, corresponding to $\mathbf{M}_2$, as shown in Figure A7. Furthermore, the parameter p can take any of $N + 1$ values, ranging from $p_{min} = – 2$ to $p_{max} = 4$, including $p = 0$. Reference, to Figure A6, which shows the generic panel distribution $T(n)$ on which Figure A4 is based, will show that $T(n)$ also includes a total of $N + 1$ marks extending from $n_{min} = – 4 = – p_{max}$ to $n_{max} = 2 = – p_{min}$.

These are particular cases of the general principles that:

- Any generic panel distribution $T(n)$ extending from n_{min} to n_{max}, corresponding to a total end-to-end range $N = n_{max} – n_{min}$ marks, and including $N + 1$ individual marks ...
- ... will be associated with $N + 1$ values of possible medians $\mathbf{M}_p$ ...
- ... corresponding to a range $p_{min} = – n_{max}$ to $p_{max} = – n_{min}$.

Figure A7 demonstrates that if an examination subject's generic panel distribution $T(n)$ is known, and has an end-to-end width of N marks from n_{min} to n_{max}, then the range of possible 'right' marks for a script given any mark m is limited to $N + 1$ possibilities $\mathbf{M}_p$, such that $\mathbf{M}_p = m\dagger + p$.

This immediately links to an intuitive understanding of fuzziness and grade reliability. A less fuzzy subject, such as Physics, will be associated with a more narrow generic panel distribution $T(n)$, and the corresponding value of N will be small - perhaps, say, 2 marks. Any Physics script marked m is therefore associated with $N + 1 = 3$ possible values of $\mathbf{M}_p$; by contrast, the generic panel distribution $T(n)$ for Religious Studies is likely to be broader - say, $N = 8$ marks - implying that any mark m is associated with $N + 1 = 9$ possible values of $\mathbf{M}_p$. If the grade widths are similar for both examination subjects, the likelihood that a Religious Studies mark will straddle a grade boundary is therefore greater than for a Physics mark; accordingly, the grades awarded for Religious Studies are less reliable than those awarded for Physics.

For any subject examination, the generic panel distribution $T(n)$ can be determined, and this will have an end-to-end range of N marks. Accordingly, any script given $m\dagger$ marks can be associated with $N + 1$ possible values of $\mathbf{M}_p = m\dagger + p$, any one of which is that script's 'right' mark. Limiting the range of possible 'right' marks in this way is helpful, but even better would be to have some information as regards their respective probabilities. So, for example, taking the case illustrated in Figure A7, for a script marked $m\dagger = 64$, the 'right' mark is any one of the seven possible values of $\mathbf{M}_p$ from 62 to 68 inclusive. Are each of these equally probable, with a 1 in 7 chance (a probability of about 0.14, or 14%)? Or are some values of $\mathbf{M}_p$ more likely than others? Or, more generally, what is the probability distribution of the medians $\mathbf{M}_p$, a distribution represented mathematically as $Q(p)$ such that, for any given mark $m\dagger$, the value of $Q(p)$ for any value of p defines the probability that the specific mark $m\dagger$ is associated with the median $\mathbf{M}_p = m\dagger + p$?

The distribution $Q(p)$

To determine $Q(p)$, consider an example of an examination subject for which the Figures A6 and A7 apply, and the particular case of a script marked $m\dagger = 64$ which is in fact a member of the generic panel distribution for which the median $\mathbf{M}\dagger = \mathbf{M}_2 = 66 = 64 + 2$, implying that $p = 2$. What is the corresponding probability $Q(2)$?

Reference to Figure A7 will verify that, of the seven possible generic panel distributions that include the mark $m\dagger = 64$, the one for which the median is 66 is that identified as $\mathbf{M}_2$. As shown in Figure A6, generic panel distributions $T(n)$ are defined in terms of a variable n defined such that a given mark m is related to the median $\mathbf{M}\dagger$ of its generic panel distribution as $m = \mathbf{M}\dagger + n$. In this particular case, $m = 64$ and $\mathbf{M}\dagger = 66$ implying that $n = -2$, as indeed is verified by Figure A7 which shows that the mark $m = 64$ lies two marks to the left of the median $\mathbf{M}_2$.

According to Figure A6, however, the probability that a given mark m is 2 marks less than the associated median $\mathbf{M}\dagger$ is 0.11 = 11%. Conversely, the probability that a median $\mathbf{M}\dagger$ is 2 marks more than a given mark m is also 11%. This is the case of interest, and so the probability $Q(2)$ in this particular instance is 11%, the value of $T(-2)$.

By exactly the same reasoning, comparing Figures A6 and A7, for any value of p, the value of $Q(p)$ is given by the corresponding value of $T(-p)$. A depiction of the probability distribution $Q(p)$ is shown in Figure A8.

Figure A8: The distribution $Q(p)$ of the medians $\mathbf{M}_p$ shown in Figure A7

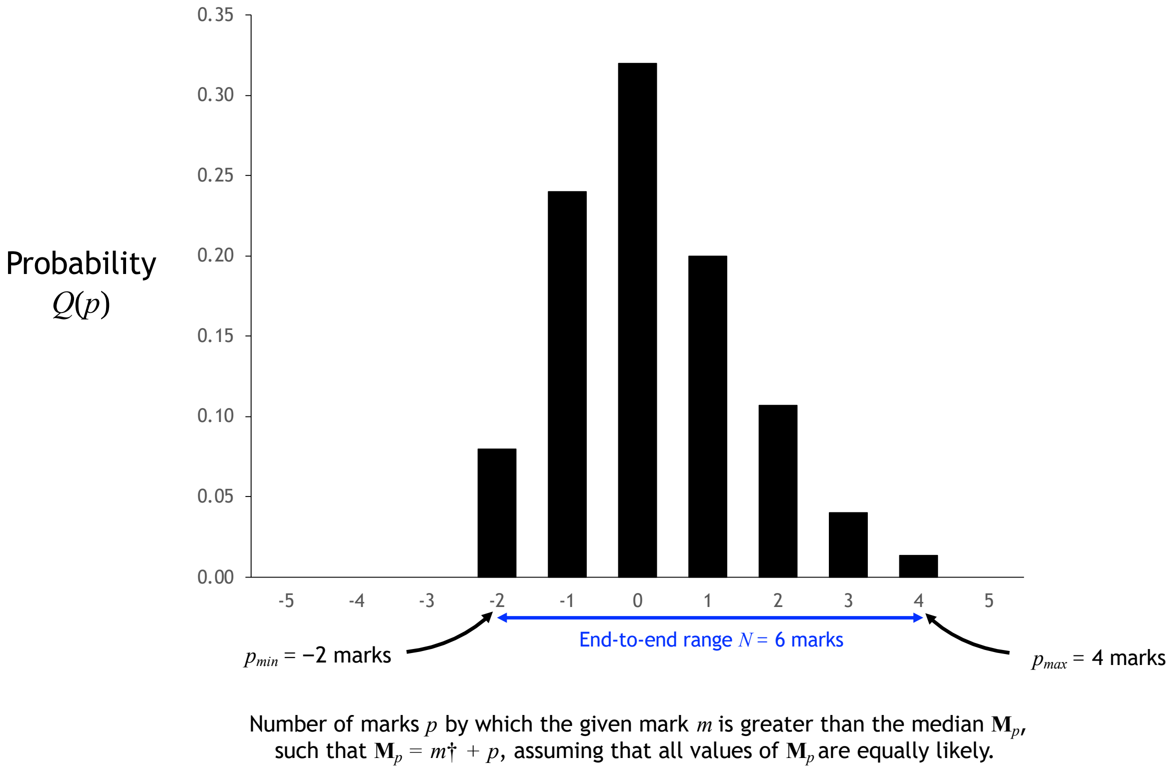


Figure A8 is consistent with Figure A7, with each of the columns in Figure A8 corresponding to the equivalent median $\mathbf{M}_p$ as shown in blue in Figure A7. Figure A8, however, not only identifies the range of possible values of $\mathbf{M}_p$ (as does Figure A7), but also identifies their probabilities: for a script marked $m\dagger$ = 64, the most likely median $\mathbf{M}_p$ with which that mark is associated corresponds to p = 0, implying that the probability that $\mathbf{M}_0 = m\dagger + 0 = 64$ is 0.32 = 32%; the probability that $\mathbf{M}_4 = m\dagger + 4 = 68$ is 0.02 = 2%.

Furthermore, as can be seen by comparing Figures A8 and A6, the distribution $Q(p)$ of the medians $\mathbf{M}_p = m\dagger + p$ is the left-right mirror image of the corresponding generic panel distribution $T(n)$; this verifies that, as discussed on page 95, if the generic panel distribution is defined mathematically as $T(n)$, then the distribution of medians is defined mathematically as

$Q(p) = T(-p)$. Accordingly, the end-to-end range of the distribution $Q(p)$ is the same as the end-to-end range of the associated distribution $T(n)$, namely, N marks.

The Bayesian assumption

The result $Q(p) = T(-p)$, however, masks an unstated assumption - that all the distributions associated with all the allowed values of $\mathbf{M}_p = m\dagger + p$ are equally probable. If this is not the case, and if there is prior information as to the distribution $H(\mathbf{M}_p)$ of the medians $\mathbf{M}_p = m\dagger + p$ over the entire mark range, then Bayesian probability theory states that

$$Q(p) = \frac{H(\mathbf{M}_p)\ T(-p)}{\sum\limits_p H(\mathbf{M}_p)\ T(-p)}$$

This reduces to $Q(p) = T(-p)$ if the values of $H(\mathbf{M}_p)$ for all values of p are equal, or approximately so – as they often are, especially when the total end-to-end width of marks is relatively small.

In fact, $H(\mathbf{M}_p)$ can never be known - it is impossible for a panel to mark all scripts, and so to determine all the medians $\mathbf{M}_p$. What is known, however, is the distribution $H'(m)$ of actual marks m over all scripts, encompassing all possible values of m over the entire range of marks from 0 to 100. The distribution $H'(m)$ for an examination in a given subject is known year-on-year, and if the annual examinations are regarded as equivalent, then the aggregate of each annual $H'(m)$ might be taken as a reasonable approximation of $H(\mathbf{M}_p)$.

As was discussed on page 59, all the solutions to the grade reliability problem require the measurement of an examination subject's fuzziness, as represented by the parameter *f*. This in turn is determined not by the shape of the distribution $Q(p)$ but by its end-to-end range N. If all the associated values of $H(\mathbf{M}_p)$ are non-zero, then the end-to-end range N of $Q(p)$ can be derived directly from $T(-p)$, regardless of the behaviour of $H(\mathbf{M}_p)$. If only the value of N is to be estimated, then the details of $Q(p)$ do not need to be determined, nor does knowledge - or lack of knowledge - of $H(\mathbf{M}_p)$ matter.

If, however, there are values of $\mathbf{M}_p$ for which $H(\mathbf{M}_p) = 0$, then $H(\mathbf{M}_p)$ does matter, for this reduces the end-to-end range N of $Q(p)$, making it less than the end-to-end range of $T(n)$. However, if $H(\mathbf{M}_p)$ is zero, then the corresponding mark $\mathbf{M}_p$ can never be given. This can only be the case for all marks outside the marking range (say, below 0 or greater than 100), or perhaps for marks at the very extreme ends (say, below 3 or greater than 98) - neither of which are important as regards the grade reliability.

In many practical contexts, the values of $H(\mathbf{M}_p)$ for any particular script may be assumed to be equal, or nearly so, implying that, to a good approximation

$$Q(p) = T(-p)$$

and that the distribution $Q(p)$ can be represented, as exemplified in Figure A8.

In Figure A8, the height of the column for any value of p represents the probability that a script, given a single mark $m†$ by a single examiner, is a member of the generic panel distribution characterised by the median $\mathbf{M}_p = m† + p$. If that median mark $\mathbf{M}_p$ has a special significance - for example, if it is the conventionally-agreed definition of the 'right' mark, or if it happens to correspond to the 'definitive' mark as given by a senior examiner - then it is this median mark $\mathbf{M}_p$ that determines the script's grade. And it is the distribution $Q(p)$ that answers the question 'If a script is a single mark $m†$ by a single examiner, what is the probability that the 'right' mark for this script is $\mathbf{M}_p = m† + p$?'. But not just that. Since a senior examiner, by definition, gives the 'right' mark, which must be one of the median marks $\mathbf{M}_p$, the distribution $Q(p)$ also answers the question 'If a script is a single mark $m†$ by a single examiner and then given a fair re-mark $\boldsymbol{m*}$ by a senior examiner, what is the probability that re-mark $\boldsymbol{m*}$ is such that $\boldsymbol{m*} = m† + p$?' - where the bold symbol $\boldsymbol{m*}$ indicates that the re-mark is done by a senior, and not by an ordinary, examiner.

Quantifying grade reliability

The distribution $Q(p) = T(-p)$, as exemplified by Figure A8, is therefore very important as regards quantifying grade reliability. Assuming for the moment that the median $\mathbf{M}†$ of a specific script's generic panel distribution $T(n)$, as illustrated in Figure A6, has the 'special' significance of being the 'right' mark, then the distribution $Q(p) = T(-p)$ has these characteristics:

- The shape - and in particular the end-to-end range N - of $Q(p)$ depends on the examination subject: the fuzzier the subject, the broader the distribution.
- For a script given any mark m, there is only one actual 'right' mark, but this mark can be determined only if a panel, or a senior examiner, were to mark that script. If the only information available is the script's mark m, then the 'right' mark can be any mark. But if the generic panel distribution $T(n)$ can be determined for the examination subject (as is quite practicable), then the distribution $Q(p) = T(-p)$ can also be determined. This then limits the possibilities as regards what that script's 'right' mark might be: there is a very high probability that 'right' mark is one of the $N + 1$ marks defined by the distribution $Q(p)$, as exemplified in Figure A8.
- For a script given any mark $m†$, the probability that the 'right' mark is $\mathbf{M}_p$, such that $\mathbf{M}_p = m† + p$, is given by the corresponding value of $Q(p)$, as exemplified by the height of the corresponding column in Figure A8.

This last point unlocks the quantification of grade reliability as measured by reference to a 'special' mark, such as the the mark given by a senior examiner, this being assumed to be the median **M** of the examination's generic panel distribution $T(n)$. If a script is given a mark $m\dagger$, say, 64, and if the examination subject's generic panel distribution $T(n)$ is as illustrated in Figure A6, then Table A3 shows the probabilities that the 'right' mark is one of the seven possibilities from 62 to 68.

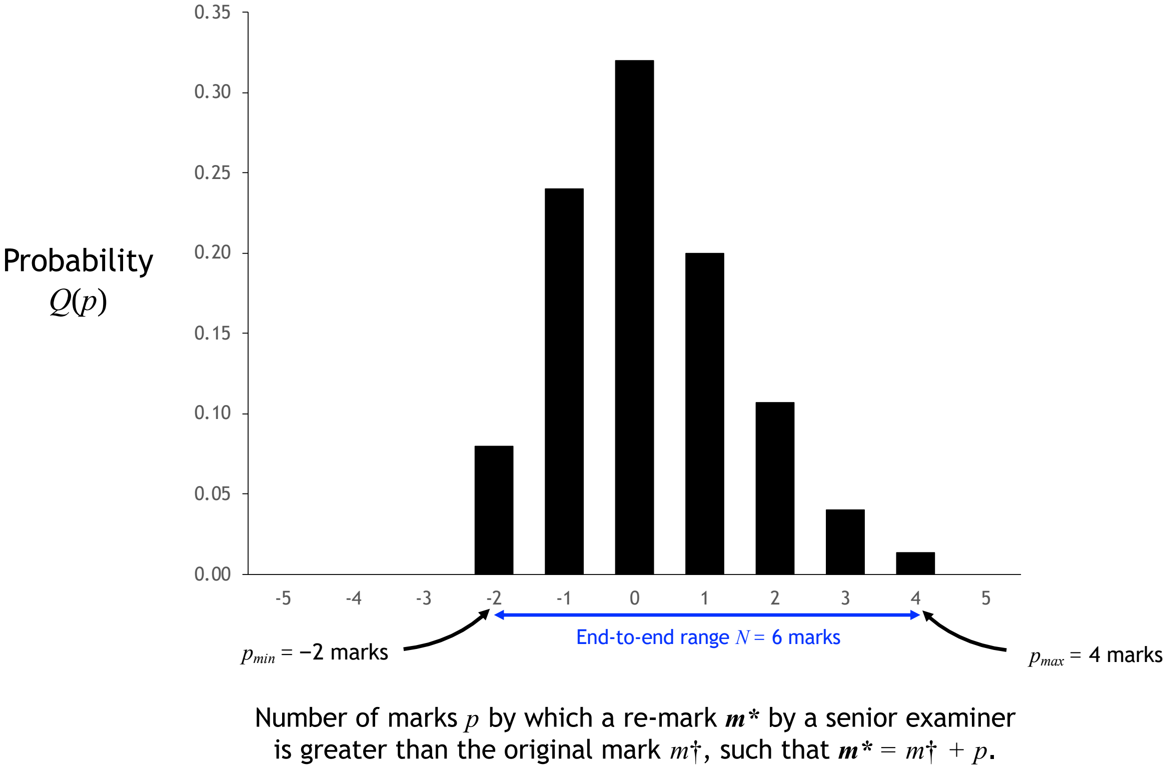
Table A3: The probability $Q(p)$ that a script originally marked $m\dagger = 64$ is associated with a particular 'special' mark $= m\dagger + p$, these being the medians of successive individual panel distributions as illustrated in Figure A7

'Special' mark	p	Probability $Q(p)$	
		Percentage	Numeric
≤ 61	≤ -3	< 0.1%	< 0.010
62	– 2	6.0%	0.060
63	– 1	23.5%	0.235
$m\dagger = 64$	0	32.0%	0.320
65	1	21.0%	0.210
66	2	11.0%	0.110
67	3	4.5%	0.045
68	4	2.0%	0.020
≥ 69	≥ 5	< 0.1%	< 0.001
Total		100.0%	1.000

Suppose that, for this examination, grade B is defined as all marks from 61 to 65 inclusive, and grade A marks from 66 to 70 inclusive. A script is marked 64 is awarded grade B, but according to the data shown in Table A3, there is a probability of 21.0% that the corresponding 'special' = 'right' mark is 65; 11.0%, 66; 4.5%, 67; and 2.0%, 68. This implies that there is a probability of 21.0 + 11.0 + 4.5 + 2.0 = 38.5% that the 'special' = 'right' grade is grade B. The reliability of the originally-awarded grade is therefore 61.5%.

The distribution $Q(p) = T(-p)$ therefore defines the probability that a script given any original mark $m\dagger$ would be given a different, 'special', mark $\boldsymbol{m^*} = m\dagger + p$ as the result of a fair re-mark. The distribution $Q(p) = T(-p)$ is therefore known as the **special re-mark distribution**, as illustrated in Figure A9 - noting that the histogram in Figure A9 is identical to that shown in Figure A8, but the caption is different.

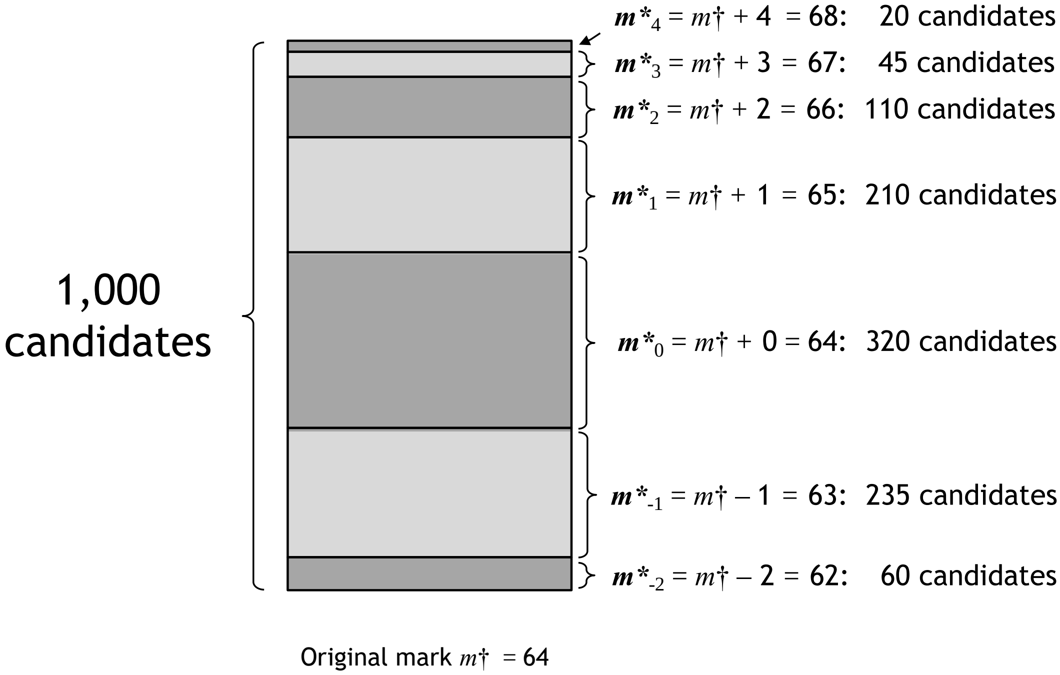
Figure A9: The special re-mark distribution, $Q(p)$ defining the probability that a script originally marked $m†$ will be re-marked $\boldsymbol{m*} = m† + p$ by a senior examiner



$Q(p)$ and grade reliability

Suppose that the special re-mark distribution $Q(p)$ as illustrated in Figure A9, and the associated data as shown in Table A3, are known, and valid for a particular examination subject. Suppose further that 1,000 candidates are marked $m†$ = 64 marks. There is therefore a probability of 0.32 = 32% that a senior examiner would re-mark any of those 1,000 scripts $\boldsymbol{m*}$ = 64, corresponding to p = 0. The number of candidates re-marked $\boldsymbol{m*}$ = 64 by a senior examiner may therefore be estimated as 0.32 x 1,000 = 320 candidates. Similarly, the number of candidates re-marked $\boldsymbol{m*}$ = 63 by a senior examiner, corresponding to p = – 1, is 0.235 x 1,000 = 235, and likewise for all re-marks from $\boldsymbol{m*}$ = 62 (p = – 2) to $\boldsymbol{m*}$ = 68 (p = 4), with the number of candidates being re-marked $\boldsymbol{m*}$ = 61 or lower, or $\boldsymbol{m*}$ = 69 or higher, estimated as zero. These inferences can be represented as shown in Figure A10.

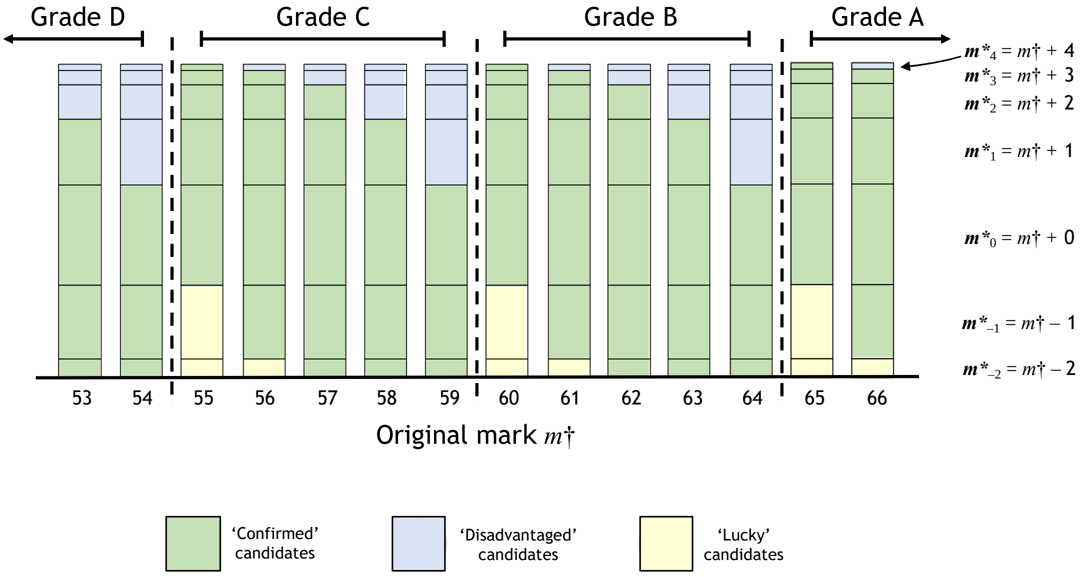
Figure A10: Re-marks $\boldsymbol{m^*}$ *by a senior examiner for a cohort of 1,000 candidates, all given an original mark* $m\dagger$ *= 64, for an examination for which the special re-mark distribution* $Q(p)$ *as shown in Figure A9 is valid*


$m^*_4 = m\dagger + 4 = 68$: 20 candidates
$m^*_3 = m\dagger + 3 = 67$: 45 candidates
$m^*_2 = m\dagger + 2 = 66$: 110 candidates
$m^*_1 = m\dagger + 1 = 65$: 210 candidates
1,000 candidates
$m^*_0 = m\dagger + 0 = 64$: 320 candidates
$m^*_{-1} = m\dagger - 1 = 63$: 235 candidates
$m^*_{-2} = m\dagger - 2 = 62$: 60 candidates
Original mark $m\dagger$ = 64

In Figure A10, there are a total of $N + 1 = 7$ 'layers', corresponding to each of the allowed values of p from $p_{min} = -2$ to $p_{max} = 4$, as shown in Figure A7; furthermore, the 'thickness' of each layer is proportional to the corresponding value of the special re-mark distribution $Q(p)$. In essence, the distribution $Q(p)$ is being displayed vertically, from p_{min} at the bottom to p_{max} at the top.

Figure A11 brings together a series of representations of the type illustrated in Figure A10 for a sequence of marks from 53 to 66, for an examination with the grade boundaries as shown.

Figure A11: A visualisation of grade reliability



For clarity, the figure is based on the (unrealistic) assumption that the same numbers of candidates are given each of the original marks $m†$, as shown by the equal heights of all the columns. As a consequence, the (much more valid) assumption that the same special re-mark distribution $Q(p)$ applies to all original marks $m†$ implies that any given layer has the same thickness across the diagram.

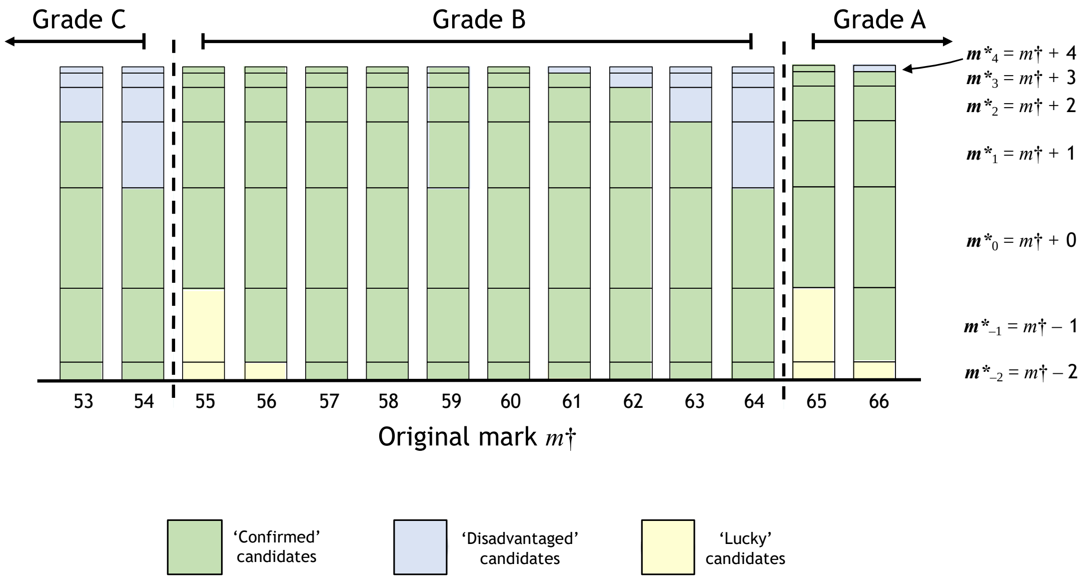
Taking as an example those candidates all originally given $m†$ = 64, and all awarded grade B, the bottom three layers represent the numbers of candidates whose scripts, if re-marked by a senior examiner, would be given 62 (60 candidates), 63 (235 candidates) or 64 (320 candidates), all of whom have their original grades confirmed. The top four layers represent the numbers of candidates whose scripts would be re-marked 65 (210 candidates), 66 (110 candidates), 67 (45 candidates) or 68 (20 candidates), all of whom would be up-graded to grade A; using the terminology introduced on page 39, these 385 candidates are all 'disadvantaged'. Of the 1,000 candidates originally marked $m†$ = 64, a total of 615 would have their grades confirmed by a re-mark by a senior examiner, and 385 would have their grades changed; the reliability of the 1,000 grades originally marked $m†$ = 64 is therefore 61.5%.

For the 1,000 candidates originally given $m†$ = 60, the bottom two layers represent candidates whose scripts would be re-marked 58 (60 candidates) or 59 (235 candidates), resulting in a down-grade to grade C; these 295 candidates are therefore 'lucky'. The remaining 705 candidates would be re-marked 60 (320 candidates), 61 (210), 62 (110), 63 (45), or 64 (20), all of which are confirmed as the original grade C. The reliability of the 1,000 grades originally marked $m†$ = 60 is therefore 70.5%.

For the 5,000 candidates originally awarded grade B, Figure A11 indicates that 645 candidates are disadvantaged, and 345 lucky; 4,010 candidates would have their grade confirmed. The average reliability of grade B is therefore 4,010/5000 x 100 = 80.2%.

Figure 12 shows exactly the same data as that shown in Figure 11, but with grade B now being wider, encompassing all original marks $m†$ from 55 to 64 inclusive, corresponding to a total of 10,000 candidates.

Figure A12: The effect on grade reliability of grade width



It is immediately evident visually that, compared to Figure A11, the green area associated with the wider grade B is now much larger, both in absolute terms and also in relation to the associated pale blue and yellow areas, suggesting that the reliability of grade B has increased. This can be verified numerically: in Figure A12, of the 10,000 candidates originally awarded grade B, the number of candidates whose grades are changed as the result of a re-mark by a senior examiner is the same in both figures at 990, but the number of candidates whose grades are confirmed is now 9,010. The average reliability for the wider grade B shown in Figure A12 is therefore 90.1%, compared to an average reliability of 80.2% for the narrower grade B shown in Figure A11.

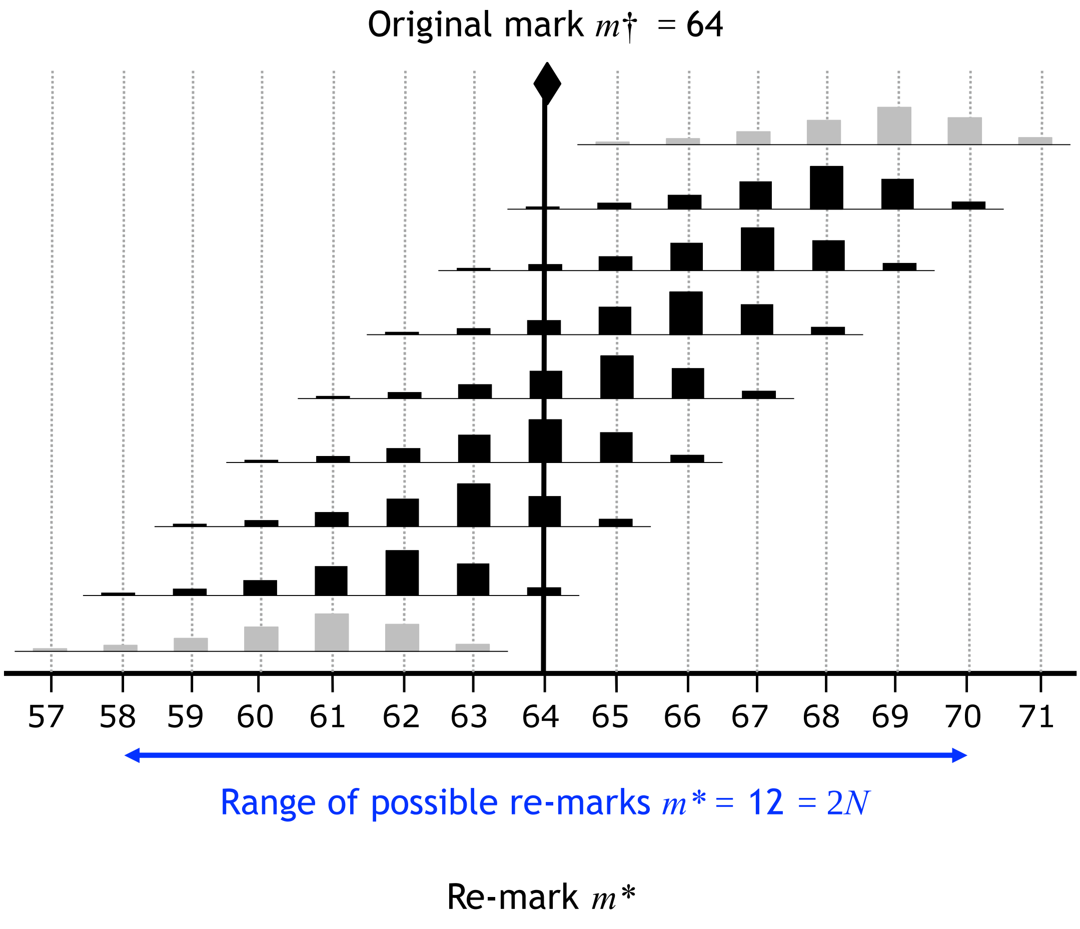
The ordinary re-mark distribution $r(h)$

The distribution $r(h)$

The special re-mark distribution $Q(p)$, as exemplified by Figure A9, is exactly that – ‘special’ – for it defines the probability that, on being fairly re-marked, a script originally mark m will be given a ‘special’ mark, such as the mark corresponding to the median $\mathbf{M}_p$ of an overlapping generic panel distribution, or the re-mark $\boldsymbol{m^*} = m\dagger + p$ given by a senior examiner.

Suppose, however, that a script originally marked $m\dagger = 64$ is re-marked m^* by an ordinary examiner, drawn at random from the entire team of examiners (where a re-mark by an ordinary examiner is symbolised by m^*, in contrast to the bold symbol $\boldsymbol{m^*}$ for a re-mark by a senior examiner). Both the original mark $m\dagger$ and the re-mark m^* must be members of the same generic panel distribution, but if only the original mark $m\dagger$ is known at the outset, there is no knowledge as to which particular generic panel distribution this might be, as exemplified in Figure A13.

Figure A13: Re-marking by an ordinary examiner



For an examination for which $T(n)$ is as shown in Figure A6, a script originally marked $m†$ = 64 can be a member of any of the seven distributions represented by the 'vertically squashed' representations of the distribution, as shown in Figure A13. If this script is fairly re-marked by an ordinary examiner, the re-mark $m*$ can be any mark from 58 to 70, 6 marks either side of the original mark 64 and spanning a total range of 70 – 58 = 12 marks – twice the range of the distributions of medians $\mathbf{M}_p$ shown in Figure A7.

As shown in Figure A13, for an examination characterised by the generic panel distribution $T(n)$ as illustrated in Figure A6, a script originally marked $m†$ = 64 by a first ordinary examiner might be given a re-mark $m*$ by another ordinary examiner such that $m*$ can be any number between 58 and 70, but these are not equally probable.

Accordingly, we may define a distribution $r(h)$ specifying the probability that a script originally marked $m†$ is re-marked $m* = m† + h$ by an ordinary examiner. The distribution $r(h)$, known as the **ordinary re-mark distribution**, can be obtained by weighting all possible distributions $T(n)$ by the probability of their occurrence as defined by $Q(p)$, implying (as will be proven on pages 125 to 133) that $r(h)$ is known mathematically as the 'convolution' of $Q(p)$ and $T(p)$, represented by the symbol ⁎ as

$$r(h) = Q(p) \ast T(p)$$

If, as discussed on pages 110 and 111, we may assume to an acceptable approximation that $Q(p) = T(-p)$, then

$$r(h) = T(-p) \ast T(p)$$

this being known mathematically as the 'auto-correlation' of the underlying distribution $T(n)$, as shown in Figure A14, with the corresponding numerical values in Table A3.

Figure A14: The ordinary re-mark distribution $r(h)$

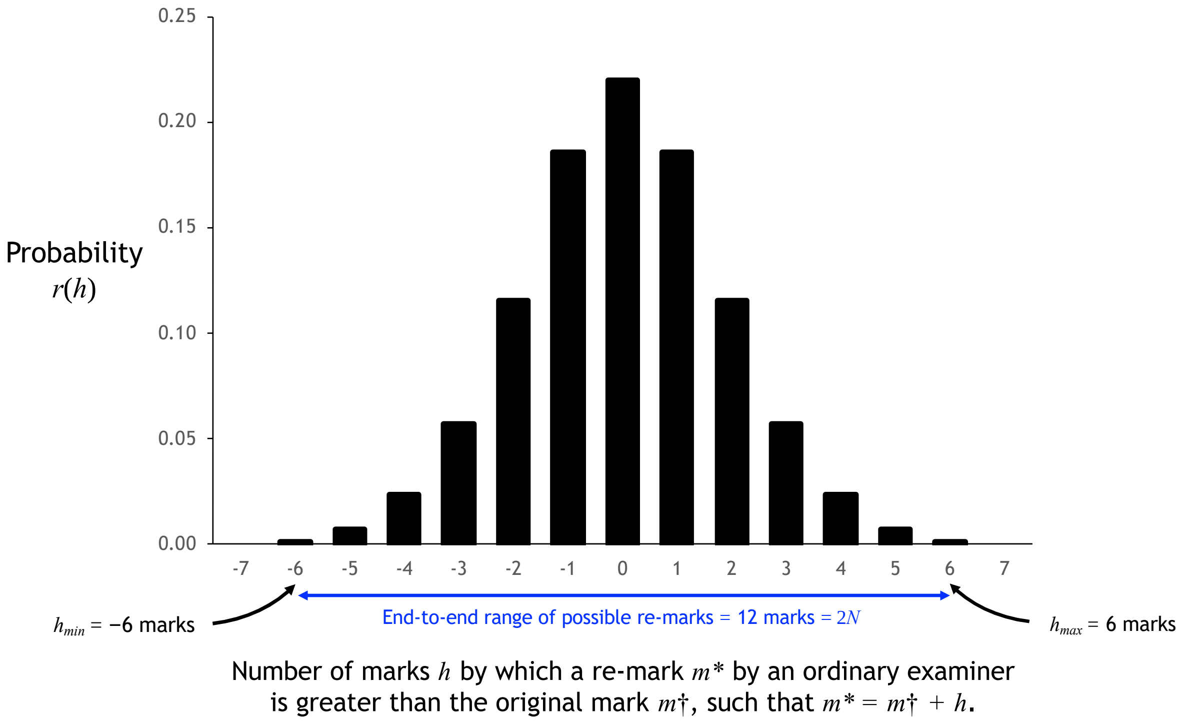


Table A4: The probability that a script originally marked $m†$ = 64 *will be re-marked* $m^* = m† + h$*, as shown in Figure A14.*

Re-mark m^*	h	Probability $r(h)$	
		%	Numeric
≤ 57	≤ − 7	< 0.01%	< 0.0001
58	− 6	0.1%	0.001
59	− 5	0.7%	0.007
60	− 4	2.4%	0.024
61	− 3	5.7%	0.057
62	− 2	11.5%	0.115
63	− 1	18.6%	0.186
$m†$ = 64	0	22.0%	0.220
65	1	18.6%	0.186
66	2	11.5%	0.115
67	3	5.7%	0.057
68	4	2.4%	0.024
69	5	0.7%	0.007

70	6	0.1%	0.001
≥ 71	≥ 7	< 0.01%	< 0.0001
Total		100.0%	1.000

For an examination characterised by a generic panel distribution $T(n)$, as illustrated in Figure A6, the distribution $r(h)$ shown in Figure A14, and the corresponding values in Table A4, define the probability that a script originally marked $m\dagger$ will be re-marked $m* = m\dagger + h$ by a second, ordinary examiner, drawn at random from the team of examiners. Figure A14 is a more informative representation of the same data as shown in Figure A13: Figure A13 shows separately the seven different individual panel distributions of which the original mark $m\dagger = 64$ is a member; in Figure A14, these seven individual distributions have each been weighted according to the appropriate probability of occurrence, and then aggregated.

Why the distribution $r(h)$ is important

The significance of the distribution shown in Figure A14 is that it defines the probability that a script given a mark $m\dagger$ by any one ordinary examiner will be given a mark $m* = m\dagger + h$ by another examiner – regardless of the order in which those marks are given. This distribution therefore quantifies the lottery-of-the-first-mark (see pages 11 and 12).

The fundamental measurement defined by this distribution is a comparison between two marks, $m\dagger$ and $m*$, each given by ordinary examiners. There is no assumption as to whether any one mark is 'right', or 'special'; what is important is that the two marks $m\dagger$ and $m*$ are different, and should they lie on different sides of a grade boundary, the corresponding grades will be different. Since the ordinary re-mark distribution $r(h)$ describes the statistics of ordinary marking, it is more realistic and practical than the special re-mark distribution $Q(p)$, as exemplified by Figure A9.

As already noted, however, any examination subject has a characteristic generic panel distribution $T(n)$. To what is usually a very good approximation (see pages 110 and 111), the distribution $T(n)$ can be used to determine the special re-mark distribution $Q(p)$ as

$$Q(p) = T(-p)$$

Furthermore, as already noted and as will be proven on pages 125 to 133, the ordinary re-mark distribution $r(h)$ is related to the two distributions $T(p)$ and $Q(p)$ by the mathematical process known as convolution as

$$r(h) = Q(p) \ast T(p) = T(-p) \ast T(p)$$

Accordingly, if the distribution $r(h)$ can be determined by statistical sampling as described on pages 56 and 57, then the mathematical process known as ‘deconvolution’ can be used to derive $T(p)$ and hence $Q(p) = T(-p)$.

The special re-mark distribution $Q(p) = T(-p)$ and the ordinary re-mark distribution $r(h) = Q(p) \ast T(p) = T(-p) \ast T(p)$ are therefore not independent: knowledge of the one implies knowledge of the other. The ordinary re-mark distribution $r(h)$, however, is the more pragmatic. Since it is based on the marks given by ordinary examiners, it can be measured by statistical sampling across the whole examiner community as described an pages 56 and 57; furthermore, unlike the special re-mark distribution $Q(p)$, the ordinary re-mark distribution $r(h)$, does not require, or rely on, a (conceptually problematic) definition of ‘right’ or ‘definitive’.

Comparison of Figure A9, which shows a representative special re-mark distribution $Q(p)$, and Figure A14, which shows the corresponding ordinary re-mark distribution $r(h)$, highlights three differences between these two distributions:

- The ordinary re-mark distribution $r(h)$ is necessarily – and therefore always – symmetrical about the mid-point, so explaining the symmetry of the ‘whiskers’ in Figures 3 and 4 of the main text (see pages 20 and 24). The special re-mark distribution $Q(p)$ will be symmetrical if the underlying generic panel distribution $T(n)$ is itself symmetrical, as it often is, but not always (as, for example, illustrated in Figures A6 and A9). The total end-to-end range of $r(h)$ can therefore be expressed as $2f$ marks, such that $r(h)$ extends from $h_{min} = m\dagger - f$ to $h_{max} = m\dagger + f$. It is this parameter f that features in the various solutions to the grade reliability problem, as discussed on pages 60 to 83.
- The ordinary re-mark distribution $r(h)$ is always both flatter ...
- ... and broader than the corresponding special re-mark distribution $Q(p)$.

This third point is especially important as regards measures of grade reliability. As has been mentioned many times, the fuzzier the subject, the more unreliable the corresponding grades. ‘Fuzziness’ is a vague, if descriptive, term; fuzziness, however, can be quantified in terms of measurements of the width of either the special re-mark distribution $Q(p)$ (if the re-mark is a ‘special’ mark, such as the ‘definitive’ mark given by a senior examiner) or the ordinary re-mark distribution $r(h)$ (if the re-mark is a mark given by a second ordinary examiner). There are a number of different possible measures of the widths of these distributions, the first being the standard deviation (which can be computed, but is not immediately obvious from depictions such as those in Figures A9 and A14), and the second the end-to-end range (which is statistically less rigorous, but easier to measure). But whichever measure is chosen, and ensuring that the same measure is used for corresponding special and ordinary re-mark distributions, there is a fundamental truth: the measure of grade reliability derived from the narrower special re-mark distribution will *always be a larger number* than the measure derived from the broader ordinary re-mark distribution. If the measure is the end-to-end range, then the width of the corresponding ordinary distribution – 12 marks for the example shown in Figure A13 – is double that of the width of the corresponding special distribution (6 marks in the example shown in Figure A8); if the measure is the standard deviation, and if the underlying generic distribution $T(n)$ is a Gaussian distribution of standard deviation σ, then the standard deviation of the special re-mark distribution $Q(p)$ is also σ, and that of the ordinary re-mark distribution $r(h)$, $\sigma\sqrt{2}$ (see page 5 here).

Accordingly, if measures of grade reliability are made using the special re-mark distribution by reference to a senior examiner, then grades will appear to be more reliable than as measured relative to another ordinary examiner. That explains the qualitative difference between the numbers shown in Table 5 on page 54 in the main paper; the quantitative difference will be explained on pages 125 to 133.

The double marking fallacy

A further feature of the ordinary re-mark distribution, as exemplified by Figure A14, concerns the widely-held belief that marking a script twice – 'double marking' – yields a more reliable mark. So, for example, if the original mark is $m†$, and the re-mark $m*$, then perhaps $m*$ is the 'right' mark – as it will be if the second examiner is a senior examiner whose mark is by definition 'definitive'. If, however, the second examiner is an ordinary examiner, then the re-mark $m*$ will be 'definitive' only if that ordinary examiner happens to give the same mark as that given by a senior examiner, which would be a statistical accident; but surely it is 'common sense' that, under all circumstances, the average $(m† + m*)/2$ is a 'better' mark than either $m†$ or $m*$. Is this true?

To explore this question, suppose that the generic panel distribution of the subject examination is as shown in Figure A6, implying that all the subsequent figures are valid. Suppose further that a 'secret study' has determined that a particular script is known to be a member of the individual panel distribution associated with a median $\mathbf{M}† = 66$, and that this median is the 'definitive' mark given by a senior examiner.

None of this is known to the ordinary examiners, one of whom marks the script $m† = 64$. The script is then fairly re-marked $m* = 62$ by a second ordinary examiner, so identifying the two-heads-are-better-than-one average mark $(m† + m*)/2 = 63$. Which of the three marks 64 (the original), 62 (the re-mark) and 63 (the average) is right?

Figure A15: Double marking. A script is given an original mark $m†$ = 64. The 'definitive' mark for that script is $\mathbf{M}†$ = 66. How useful is a re-mark $m*$ = 62?

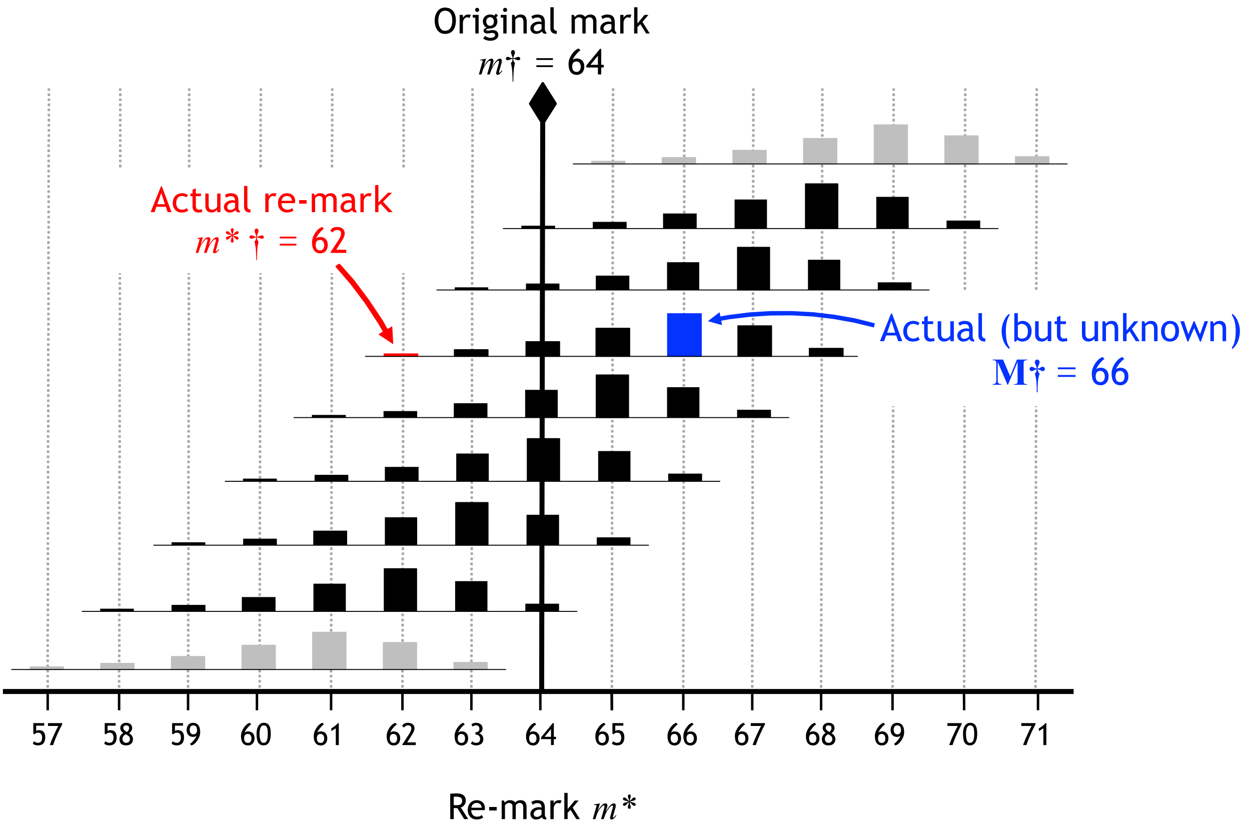


Figure A15, which contains much more information than is available to either of the two ordinary examiners, provides the context in which this example of double marking is taking place. The original mark $m†$ = 64 is a member of any one of seven possible generic panel distributions, each with its own 'definitive' mark, the median $\mathbf{M}_p$; in fact, the 'secret knowledge' is that the actual generic panel distribution for this particular script is that associated with the median $\mathbf{M}_2$ = 66 = $\mathbf{M}†$.

The second ordinary examiner gives the script a re-mark $m*$ = 62; as can be seen, this mark is also a member of the panel distribution associated with the median $\mathbf{M}_2$ = 66, and so is a valid re-mark. According to the 'secret knowledge', however, the 'definitive' mark for this script is $\mathbf{M}_2$ = 66, implying that both the re-mark $m*$ = 62 and the average mark ($m†$ + $m*$)/2 = 63 are even further from the 'definitive' mark than the original mark $m†$ = 64. In this instance, double marking has made matters worse, not better.

Reference to Figure A14, and the data in Table A4, shows that there is a probability of about 12% that a script originally marked $m†$ = 64 will be re-marked two marks lower, $m*$ = 62, for which the parameter h = – 2. Figure

A14 also shows that the most likely re-mark, with a probability of about 22%, corresponds to a value of h = 0, implying that the re-mark $m*$ is equal to the original mark m† = 64. A re-mark $m*$ = 64 might be interpreted as confirmation that 64 is indeed the 'right' mark – but reference to Figure A15 shows that the re-mark has in fact added no further useful information: the mark, and the re-mark, are both members of all of the original seven possible panel distributions.

If, however, the re-mark is $m*$ = 68, corresponding to a value of h = 4, then the average mark (m† + $m*$)/2 = 66, which is equal to the 'definitive' mark **M**† = 66. The probability that this will happen, according to Figure A14, is about 2% (compared to probabilities of about 22% for a re-mark $m*$ = 64, and about 12% that $m*$ = 62), and a re-mark $m*$ = 68 is the only value for which the average is 'right'. Furthermore, since both the original mark m† = 64 and the re-mark $m*$ = 68 must both be members of the same generic panel distribution, a re-mark $m*$ = 68 eliminates the possibility that this panel distribution is associated with a median of 62, 63, 64 or 65, but still leaves open the possibilities of medians 66, 67 or 68.

As noted earlier, reference to Figure A14 shows that the probability that a script originally marked m† = 64 and then fairly re-marked $m*$ = 62, for which the parameter $h = -2$, is about 12%. It might be argued, however, that to use Figure A14 to determine this probability is wrong. Since it is known that both the original mark m† = 64 and the re-mark $m*$ = 62 must be members of the same generic panel distribution, the correct distribution to use is that representing the generic panel distribution as shown in Figure A6 and Table A2: if a re-mark $m*$ is 2 marks lower than the original mark m†, the probability is therefore about 11%. This number happens to be rather close to the probability of 12% as inferred from Figure A14, but this is a numerical co-incidence rather than an indication of a deeper truth; the fundamental question remains – which of Figures A6 and A14 is the correct one to use?

The argument in favour of using Figure A6, the generic panel distribution, is apparently compelling, for it as indeed true that both the original mark m† = 64 and the re-mark $m*$ = 62 must indeed be members of the same generic panel distribution – and it is Figure A6, not Figure A14, that shows the probabilities that the same script is given different marks.

However compelling, this argument is false. It is, however, true that both the original mark m† and the re-mark $m*$ must be members of the same generic panel distribution. But when the only information available is the original mark m†, there is no knowledge as to which specific distribution this is; furthermore, the additional information provided by the re-mark $m*$ reduces this uncertainty to only a limited extent, if at all (for example, when m† = $m*$ = 64, and the full uncertainty remains), and the fuzzier the examination subject, the less helpful the second mark. Certainly, if the script is re-marked not just once, but progressively, then each successive re-mark provides further information: ultimately, 100 re-marks will reproduce the 'correct' individual panel distribution, the median of which is indeed the 'definitive' mark.

Overall, for any original mark $m\dagger$, any single re-mark $m*$ is as likely to be higher than the original mark $m\dagger$ as it is to be lower, and the resulting average $(m\dagger + m*)/2$ will also be lower or higher accordingly. Both marks $m\dagger$ and $m*$ are just random samples from the range ultimately determined by the examination subject's fuzziness; none of the marks $m\dagger$, $m*$ and $(m\dagger + m*)/2$ have any particular significance. Double marking adds little useful information, and, as discussed on pages 49 and 50, the search for the 'right' mark is illusory.

The mathematics of $Q(p)$ and $r(h)$

This section explores the mathematics of the special re-mark distribution $Q(p)$, as illustrated in Figure A8, and the ordinary re-mark distribution $r(h)$, as illustrated in Figure A11.

A script is given a single valid original mark $m\dagger$ by a single examiner, and a single valid re-mark $m*$ by another examiner. Both $m\dagger$ and $m*$ must be members of the same individual panel distribution. Operationally, however, there is no knowledge as to which particular individual panel distribution this is, and so its shape is approximated as that of the generic panel distribution $T(n)$. If the total end-to-end range of the distribution is N marks, then, as shown in Figure A7, the median can take any one of $N + 1$ values $\mathbf{M}_p = m\dagger + p$.

For any original mark $m\dagger$, the distribution $Q(p)$ defines the probability that the median $\mathbf{M}_p = m\dagger + p$ is the median of the actual generic panel distribution of which the original mark $m\dagger$ is a member. If it is this median mark that would be given if the script were fairly re-marked by a senior examiner, then the distribution $Q(p)$ is, as has been discussed, known as the special re-mark distribution. The distribution $Q(p)$ corresponding to the generic panel distribution $T(n)$ shown in Figure A6 is illustrated in Figure A8. As can be seen, the total end-to-end width of each of these distributions is the same, N marks, and mathematically $Q(p)$ is, to a good approximation as discussed on pages 110 and 111, the left-right mirror image of $T(n)$ such that

$$Q(p) = T(-p)$$

As was shown by the comparison between Figures A7 and A13, if a script originally marked $m\dagger$ is fairly re-marked $m*$ by a second ordinary examiner, the end-to-end range of possible re-marks is $2N$ marks from a lowest possible mark $m*_{min}$ to a highest possible mark $m*_{max}$. In general, $m* = m\dagger + h$, where the parameter h can take any one of $2N + 1$ values, including zero. For any original mark $m\dagger$, the probability that the re-mark $m* = m\dagger + h$ is given by the value of the ordinary re-mark distribution $r(h)$ for the corresponding value of h.

Figure A16: The variables p, n and h, showing that $h = p + n$

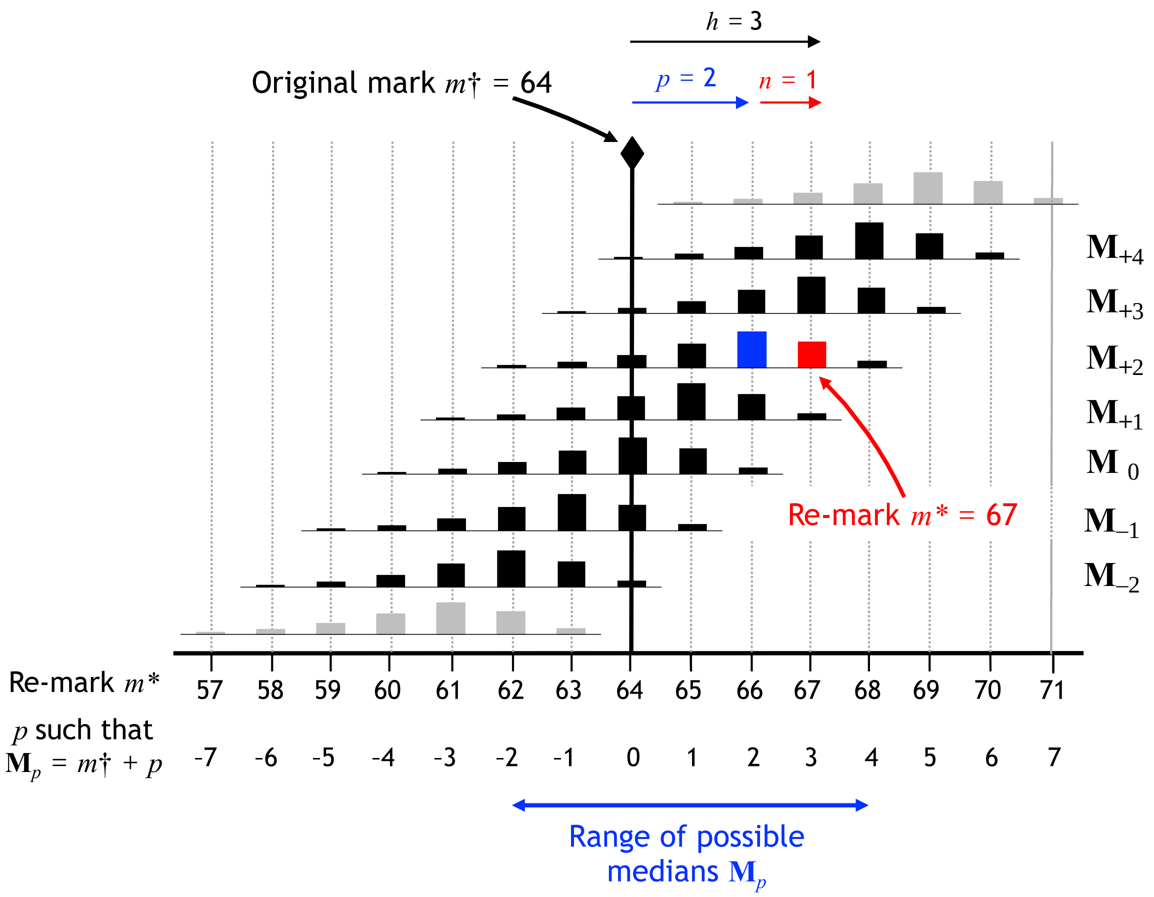


Figure A16 shows an example of a script originally marked $m^\dagger$ = 64, and subsequently re-marked m^* = 67. Since the parameter h is defined such that $m^* = m^\dagger + h$, then $h = m^* - m^\dagger$, which in this case implies that $h = m^* - m^\dagger$ = 67 - 64 = 3.

As illustrated in Figure A16, an original mark $m^\dagger$ = 64 and a re-mark m^* = 67 imply that the original mark $m^\dagger$ = 64 can be a member of the four generic panel distributions corresponding to values of p = 1, 2, 3 or 4, with medians $\mathbf{M}_1$, $\mathbf{M}_2$, $\mathbf{M}_3$ or $\mathbf{M}_4$; suppose for the moment that the actual generic panel distribution is that for p = 2, with median $\mathbf{M}_2$ as shown in blue. Within this particular generic panel distribution, the re-mark m^* = 67 corresponds to a value of n = 1, from which it is evident that

$$h = p + n$$

and therefore that

$$n = h - p$$

Although these relationships between the variables n, p and h have been demonstrated for a particular case, they are general.

Within any particular generic panel distribution, the probability that a script is given a mark $m = \mathbf{M}† + n$, n marks greater than that distribution's median $\mathbf{M}†$, is given by the corresponding value of $T(n)$. But since $n = h - p$, this probability may be written as $T(h - p)$, representing the probability that a script originally marked m is re-marked m^* such that $m^* = m† + h$, under the assumption that both the original mark $m†$ and the re-mark m^* are members of the specific generic panel distribution of median $\mathbf{M}_p = m† + p$.

The specific value of p, however, is unknown, but the probability of any of the $N + 1$ allowed values of p is defined by the distribution $Q(p)$, which is known once the underlying generic panel distribution $T(n)$ has been determined.

The probability $r(h)$ that a script given an original mark m will be given a re-mark $m^* = m† + h$ is therefore determined by weighting any particular $T(h - p)$ by the probability that the script is indeed a member of that specific distribution of median $\mathbf{M}_p = m† + p$, this being the distribution $Q(p)$, and then summing over all allowed values of p:

$$r(h) = \sum_p Q(p)\, T(h - p)$$

This summation is the mathematical definition of the convolution $Q(p) * T(p)$.

In general, as we saw on page 110,

$$Q(p) = \frac{H(\mathbf{M}_p)\, T(-p)}{\sum_p H(\mathbf{M}_p)\, T(-p)}$$

in which the $H(\mathbf{M}_p)$ terms are associated with Bayesian theory, and represent the overall probability distribution of the medians $\mathbf{M}_p$ over the entire mark range. As also described on pages 110 and 111, in many practical circumstances, the Bayesian modification is not required, and so to a good approximation

$$Q(p) = T(-p)$$

implying that

$$r(h) = \sum_p T(-p)\, T(h - p) = \sum_p T(p)\, T(h + p)$$

This expression is known as the ‘auto-correlation’ of $T(p)$. Furthermore, if, as is often the case, $T(p)$ is left-right symmetrical, $T(p) = T(-p)$ and so

$$r(h) = \sum_{p} T(p)\, T(h-p)$$

this expression being known as the ‘self-convolution’ of $T(p)$.

A mathematical expression such as

$$r(h) = \sum_{p} Q(p)\, T(h-p) = \sum_{p} T(-p)\, T(h-p)$$

can be intimidating, as can technical terms such as ‘convolution’ and ‘auto-correlation’. To interpret the expression, notice firstly that the symbol Σ indicates a summation, and the subscript $_p$ that this summation is over all allowed values of the parameter p. This parameter was introduced in Figure A7, and represents the number of generic panel distributions that include the original mark $m†$, and can take $N + 1$ values, where N is the end-to-end range of the examination subject’s generic panel distribution $T(n)$, the fundamental statistical description of that examination subject’s fuzziness. For the example used in this paper, the generic panel distribution is shown in Figure A6, and has an end-to-end range of $N = 6$ marks, implying that $N + 1 = 7$. There are therefore 7 terms in the summation.

Each of these terms is a distribution represented as $T(h-p)$. The distribution $T(n)$, the generic panel distribution, is illustrated in Figure A6 in terms of a variable n, but the shape is exactly the same if the variable used is h, such that the distribution is written as $T(h)$.

For any value of p, the distribution $T(h-p)$ has the same shape as $T(h)$ (and hence $T(n)$) but is shifted by p marks to the right (if p is positive), or to the left (if p is negative). Since, in this example, the variable p can take $N + 1 =$ 7 values from $p_{min} = -2$ to $p_{max} = +4$, including $p = 0$, the summation

$$\sum_{p} T(h-p)$$

therefore represents the summation of seven distributions, each of the same shape (as shown in Figure A6), but ‘spread’ from left to right, as illustrated in Figure A17 (in which, for clarity, each individual distribution is shown by a continuous line rather than a sequence of columns).

Figure A17: The summation $\sum_{p} T\,(h-p)$

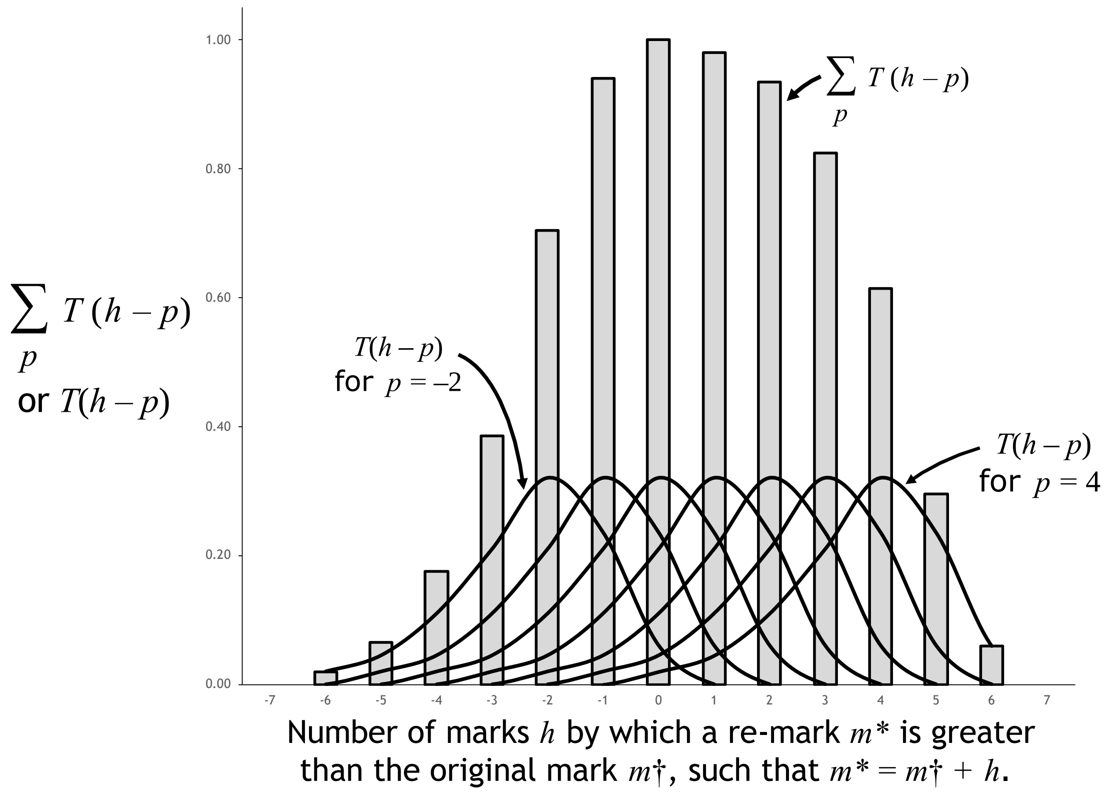


Figure A17 shows $N + 1 = 7$ generic panel distributions $T(h-p)$ of the general shape of $T(n)$ as shown in Figure A6, corresponding to each of the seven distributions shown in Figure A7, and then aggregated. When the values of each of these for any value of h are added, the result is as depicted by the histogram; the corresponding numeric values are shown an Table A5.

Table A5: Values of $T(h-p)$ *- the data corresponding to Figure A17, with blank cells* = 0

		h															Row total
		–7	–6	–5	–4	–3	–2	–1	0	1	2	3	4	5	6	7	
	$T(h)$				0.020	0.045	0.110	0.210	0.320	0.235	0.060						
p	5																0.000
	4								0.020	0.045	0.110	0.210	0.320	0.235	0.060		1.000
	3							0.020	0.045	0.110	0.210	0.320	0.235	0.060			1.000
	2						0.020	0.045	0.110	0.210	0.320	0.235	0.060				1.000
	1					0.020	0.045	0.110	0.210	0.320	0.235	0.060					1.000
	0				0.020	0.045	0.110	0.210	0.320	0.235	0.060						1.000
	–1			0.020	0.045	0.110	0.210	0.320	0.235	0.060							1.000
	–2		0.020	0.045	0.110	0.210	0.320	0.235	0.060								1.000
	–3																1.000
Column total		0.000	0.020	0.065	0.175	0.385	0.705	0.940	1.000	0.980	0.935	0.825	0.615	0.295	0.060	0.000	7.000

In Table A5, the row identified as $T(h)$ shows values of $T(h)$, which are identical to those of $T(n)$ as given in Table A2, but expressed in terms of the variable h rather than n. In particular, the median of $T(h)$ corresponds to the median value 0.320 for $h = 0$. Subsequent rows show the values of $T(h-p)$ for the various values of p defined by Figure A7, and also shown in Figures A13, A15 and A16. Across each row, the variable p is held constant, and the variable h takes successive values in principle from – 100 to + 100, but in practice only from $h_{min} = -6$ to $h_{max} = +6$, for it is only within this range that $T(h-p)$ has a non-zero value.

For values of p greater than $p_{max} = +4$ or less than $p_{min} = -2$, $T(h-p) = 0$ for all values of h; for values of p between $p_{max} = +4$ and $p_{min} = -2$ inclusive, values of $T(h-p)$ are shifted p marks to the right relative to $T(h)$ if p is positive, or p marks to the left if p is negative, with the median $\mathbf{M}_p$ of $T(h-p)$ corresponding to $h = p$.

The row totals $\sum_p T(h-p)$ are all 1.000; the column totals define the value of $\sum_p T(h-p)$ for each value of h as shown by the histogram in Figure A17; and the grand total in the bottom right-hand corner is 7.000.

In Figure A17, each of the distributions $T(h-p)$ has the same weight, implying that each distribution, and each corresponding median $\mathbf{M}_p$, are equally probable. In fact, this is not the case: the probability of any median $\mathbf{M}_p$ is determined by the corresponding value of $Q(p)$. Accordingly, when each of the $N + 1 = 7$ generic panel distributions $T(h-p)$ is weighted by the corresponding value of $Q(p)$, the result, mathematically is the ordinary re-mark distribution $r(h)$

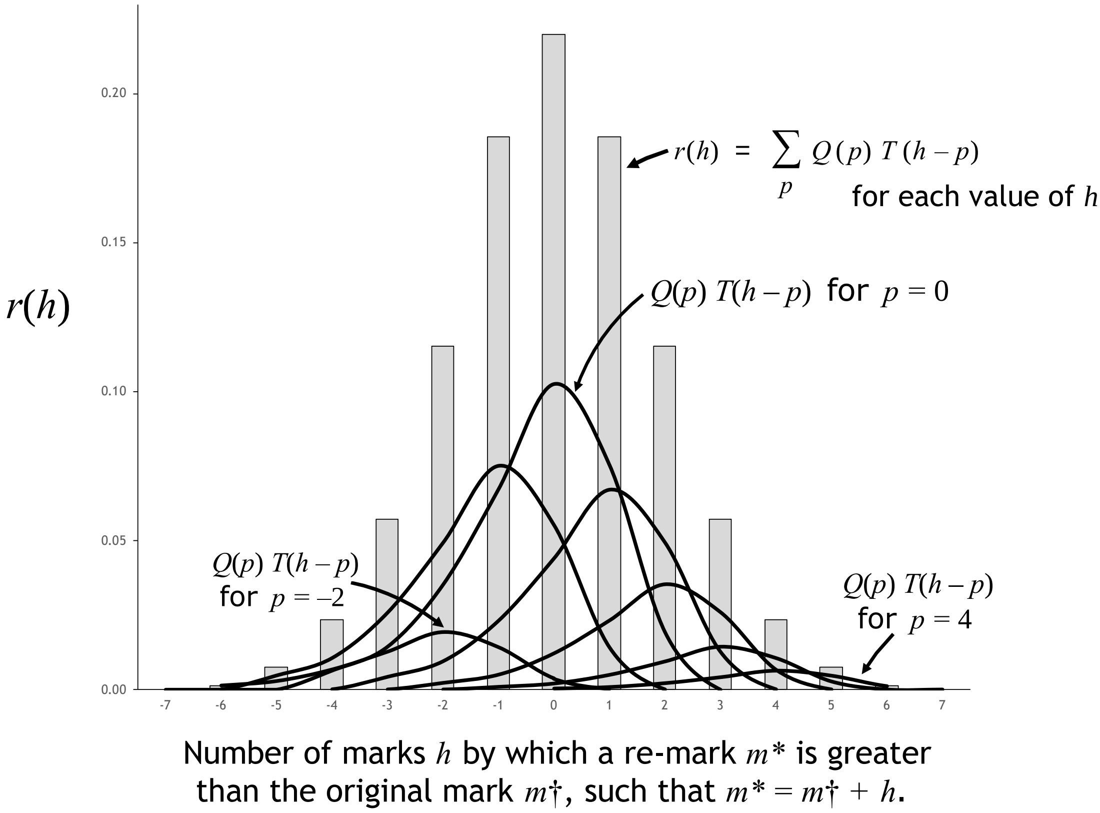
$$r(h) = \sum_p Q(p)\,T(h-p) = Q(p) * T(p)$$

If $Q(p) = T(-p)$, this becomes

$$r(h) = \sum_p T(-p)\,T(h-p) = T(-p) * T(p)$$

which may be represented graphically as shown in Figure A18.

Figure A18: The ordinary re-mark distribution $r(h) = Q(p) \ast T(p) = T(-p) \ast T(p)$



In Figure A18, the different sizes of the $N + 1 = 7$ generic panel distributions $T(h - p)$ of the general shape of $T(n)$ (compare Figure A14) are determined by weighting each $T(h - p)$ by the probability $Q(p)$ of its occurrence, with the distribution corresponding to the given mark $m\dagger$ (for which $p = 0$) having the heaviest weighting, and the remotest distributions ($p = 3$ and 4) the lightest. The summation, which represents the values of the ordinary re-mark distribution $r(h)$, is shown by the columns, and has the distinctive feature of being left-right symmetrical about $h = 0$, even though the underlying generic panel distribution $T(n)$, as shown in Figure A6, is asymmetrical. This explains the symmetry of the whiskers in Figures 3 and 4, and also the example of the ordinary re-mark distribution for the author's simulation of 2018 GCSE Geography as shown in Figure 14.

Table A6: Values of $Q(p)\ T(h-p) = T(-p)\ T(h-p)$ - *the data corresponding to Figure A18, with blank cells* = 0

			h															Row total $Q(p)$
			−7	−6	−5	−4	−3	−2	−1	0	1	2	3	4	5	6	7	
		$T(h)$				0.020	0.045	0.110	0.210	0.320	0.235	0.060						
		$Q(p)$																
p	5	0.000																0.000
	4	0.020								0.000	0.002	0.002	0.004	0.006	0.005	0.001		0.020
	3	0.045							0.001	0.002	0.005	0.009	0.014	0.011	0.003			0.045
	2	0.110						0.003	0.005	0.012	0.023	0.035	0.026	0.006				0.110
	1	0.210					0.004	0.009	0.023	0.044	0.067	0.05	0.013					0.210
	0	0.320				0.006	0.014	0.035	0.068	0.103	0.075	0.019						0.320
	−1	0.235			0.005	0.011	0.026	0.049	0.075	0.055	0.014							0.235
	−2	0.060		0.001	0.002	0.007	0.013	0.019	0.014	0.004								0.060
	−3	0.000																0.000
Column total $r(h)$			0.000	0.001	0.007	0.024	0.057	0.115	0.186	0.220	0.186	0.115	0.057	0.024	0.007	0.001	0.000	1.000

Table A6 shows the data corresponding to Figure A18. Each row represents a value for the parameter *p* from $p_{min} = -2$ to $p_{max} = +4$, and each column a value for the parameter *h* from $h_{min} = -6$ to $h_{max} = +6$. Across any row, for a given value of the parameter *p*, the numbers represent, for each value of the parameter *h*, the value of the product $Q(p)\ T(h-p) = T(-p)\ T(h-p)$. Since in any row the value of the parameter *p* is a constant, the value of $Q(p) = T(-p)$ is also a constant, corresponding to the probability that the original mark $m\dagger$ and the re-mark m^* are both members of the generic panel distribution of median $\mathbf{M}_p$. This value therefore acts as a (constant) weighting factor for each of the values of $T(h-p)$, this being a distribution of the shape of the generic panel distribution $T(n)$, but with the median $\mathbf{M}_p$ shifted to $h = p$, as shown in Table A5. Since, for all values of *p*, $Q(p) = T(-p) < 1$, the product $Q(p)\ T(h-p) = T(-p)\ T(h-p)$ will always be less than the corresponding value of $T(h-p)$, and will vary according to the value of *p*.

Each of the rows in Table A6 corresponds to the 'row' in Figures A7, A13, A15 and A16 for the same value of *p*, and the row totals in Table A6 each correspond to the summation $\sum_h Q(p)\ T(h-p)$. Since, across any row, the value of $Q(p)$ is a constant for all values of $T(h-p)$, then

$$\sum_h Q(p)\ T(h-p) \ = \ Q(p) \sum_h T(h-p)$$

As noted on page 104, the distribution $T(n)$ normalised, and so the distribution $T(h-p)$ is normalised too, implying that

$$\sum_h T(h-p) \ = \ 1$$

from which

$$\sum_{h} Q(p)\, T(h-p) \;=\; Q(p) \;=\; T(-p)$$

The row totals in Table A6 therefore show the values of the special re-mark distribution $Q(p) = T(-p)$ for each value of p, as verified by Table A3.

The column totals, which represent a summation over all values of p for each value of h, give successive values of

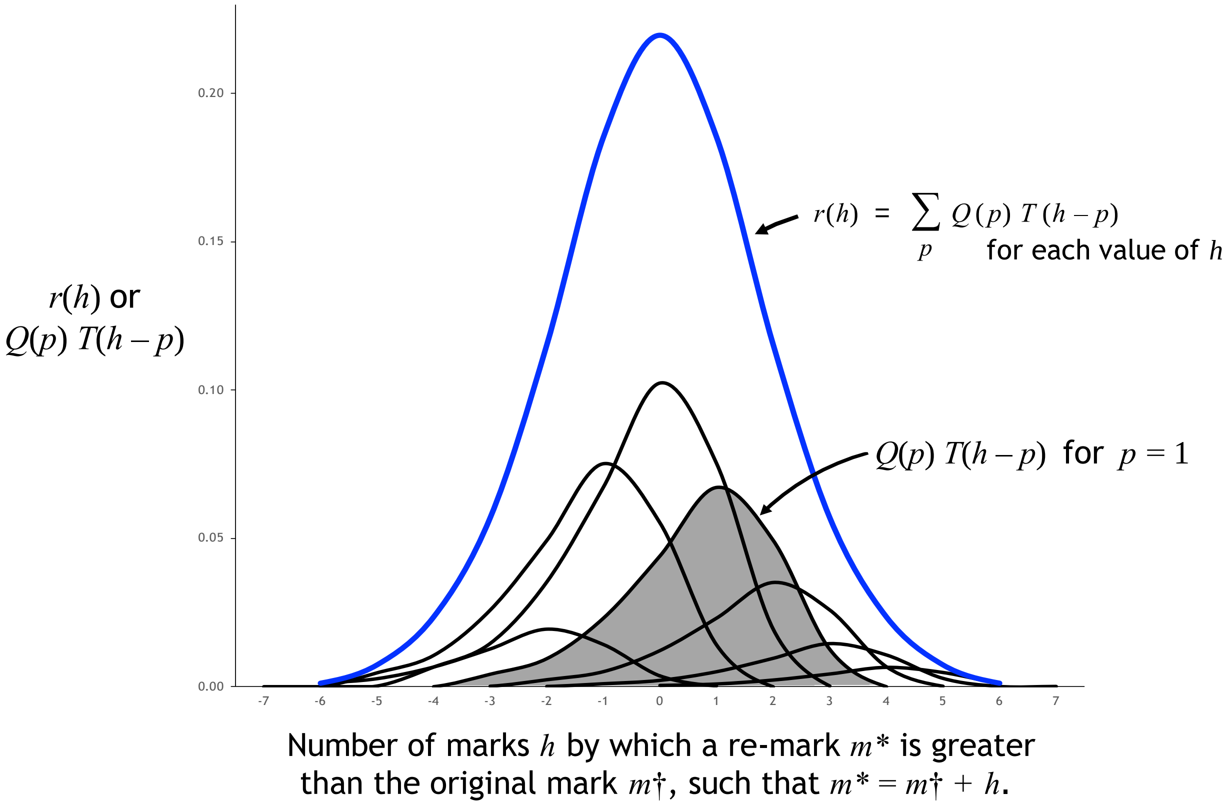
$$\sum_{p} Q(p)\, T(h-p) \;=\; \sum_{p} T(-p)\, T(h-p) \;=\; r(h)$$

This is the convolution $Q(p) * T(p) = T(-p) * T(p)$, and so the column totals give the numerical values of the ordinary re-mark distribution $r(h)$, as shown by the histogram in Figure A18.

Some properties of the ordinary re-mark distribution $r(h)$

Suppose that a script is given a re-mark $m^* = m^\dagger + h$ which is a member of the generic panel distribution of median $\mathbf{M}_p = m^\dagger + p$. The probability that the re-mark m^* is any one of the $N + 1$ marks associated with that specific generic panel distribution may be determined by calculating the total number of marks m^* associated with the corresponding value of $\mathbf{M}_p$, as given by summing the product $Q(p)\ T(h-p)$ over all possible values of h for any given value of p, as exemplified by the shaded area in Figure A19 corresponding to $p = 1$.

Figure A19: The probability that a script marked $m†$ will be re-marked m^* by an ordinary examiner, where m^* is any mark associated with the generic panel distribution for p = 1, with median $\mathbf{M}_1$



The shaded area measures the total number of marks associated with the generic empirical distribution for p = 1, the generic panel distribution of median median $\mathbf{M}_1 = m + 1$; this is also a measure of the probability that the given mark $m†$ is associated with the median $\mathbf{M}_1$. The area associated with any median $\mathbf{M}_p$ may be computed by summing the distribution $Q(p)\ T(h-p)$ over all possible values of h for a given value of p.

Mathematically, the shaded area in Figure A17 is given by the expression

$$\sum_h Q(p)\ T(h-p) = Q(p) \sum_h T(h-p)$$

in which the parameter p is a constant, for example p = 1 as shown in Figure A19.

Since the distribution $T(h-p)$ is normalised, the summation over all possible values of h must equal 1, and so the probability that any mark h is a member of the empirical distribution associated with the median $\mathbf{M}_p$ is given by

$$\sum_{h} Q(p)\, T(h-p) \;=\; Q(p) \sum_{h} T(h-p) \;=\; Q(p)$$

This is the corresponding value of $Q(p)$, the script's special re-mark distribution, in agreement with the result stated on page 33 of the main paper. The summation over h runs, in principle, from $h = -\infty$ to $h = +\infty$, but in practice from h_{min} to h_{max}.

This result may also be derived directly from the convolution function $\sum_{h} Q(p)\, T(h-p)$.

For a script associated with the generic empirical distribution $T(h-p)$, the median $\mathbf{M}_p$ of that distribution represents the 'right' mark as would be given if the script were re-marked by a senior examiner. Mathematically, the single mark $\mathbf{M}_p$ can be expressed by the Dirac δ-function $\delta(h - \mathbf{M}_p)$, which takes the value of 1 when $h = \mathbf{M}_p$, and the value of 0 for all other values of h (see page 33 here). The distribution $T(h-p)$ may therefore be replaced by the Dirac δ-function $\delta(h - \mathbf{M}_p)$, and so the convolution becomes

$$\sum_{h} Q(p)\, T(h-p) = Q(p)\, \delta(h - \mathbf{M}_p) \;=\; Q(p)$$

giving the result $Q(p)$, as before.

The total area under the $r(h)$ curve is given by

$$\sum_{h} r(h) \;=\; \sum_{h} \left[\sum_{p} Q(p)\, T(h-p) \right]$$

Reversing the order of the summations gives

$$\sum_{h} \left[\sum_{p} Q(p)\; T(h - p) \right] \;=\; \sum_{p} \left[\sum_{h} Q(p)\; T(h - p) \right]$$

from which

$$\sum_{h} r(h) \;=\; \sum_{p} Q(p) \sum_{h} T(h-p)$$

Since the two distributions $Q(p)$ and $T(h-p)$ are each normalised

$$\sum_{p} Q(p) \;=\; \sum_{h} T(h-p) \;=\; 1$$

from which

$$\sum_{h} r(h) \; = \; 1$$

so verifying that the function $r(h)$ is, as expected, normalised, as verified by the sum in the bottom right-hand cell of Table A5. Also, since in practice the summation over the $2N + 1$ values of h is from h_{min} to h_{max}, this implies that it is virtually certain that any re-mark $m*$ is within this range, as shown in Figure A13.

Index of mathematical symbols

f One-half of the end-to-end range $2N$ of the ordinary re-mark distribution $r(h)$.

h The number of marks between an original mark m and a re-mark m^* such that $m^* = m + h$.

h_{max} The maximum value of the parameter h for which the *ordinary re-mark distribution* $r(h)$ is non-zero. The end-to-end range of $r(h)$ is the difference $h_{max} - h_{min} = 2N = 2f$.

h_{min} The minimum value of the parameter h for which the *ordinary re-mark distribution* $r(h)$ is non-zero. The end-to-end range of $r(h)$ is the difference $h_{max} - h_{min} = 2N = 2f$.

m A first mark given by a single examiner to a single script.

m' An alternative first mark given by a single examiner to a single script.

m^* A fair re-mark given by a single ordinary examiner to a script originally marked m.

$\mathbf{m^*}$ A fair re-mark given by a single senior examiner to a script originally marked m.

$m\dagger$ The specific mark m as given to a particular script, against which, for example, a general re-mark m^* may be compared.

M The **mode** of any distribution.

$\mathbf{M}$ The **median** of any distribution.

$\mathbf{M}_p$ For an examination characterised by a *generic panel distribution* $T(n)$ of end-to-end range N, any script given a mark $m\dagger$ is associated with $N + 1$ *generic panel distributions*, each of median $\mathbf{M}_p = m\dagger + p$, where p can take any integer value from $p_{\mathrm{min}} = -n_{max}$ to $p_{max} = -n_{min}$, including 0.

$\mathbf{M}\dagger$ The **median** of the particular *individual panel distribution* $t(m)$ with which the mark m, given by a single examiner to a specific script, is associated.

$\langle \mathsf{M} \rangle$ The **mean** of any distribution.

n The number of marks by which the mark m given to any script by a single marker is greater than the median $\mathbf{M}\dagger$ of the generic panel distribution $T(n)$ of which that mark is a member, such that $m = \mathbf{M}\dagger + n$.

n_{max} The maximum value of the parameter n for which the *generic panel distribution* $T(n)$ is non-zero. The end-to-end range of $T(n)$ is the difference $n_{max} - n_{min} = N$; also, $n_{max} = - p_{min}$.

n_{min} The minimum value of the parameter n for which the *generic panel distribution* $T(n)$ is non-zero. The end-to-end range of $T(n)$ is the difference $n_{max} - n_{min} = N$; also, $n_{min} = - p_{max}$.

N The end-to-end range $n_{max} - n_{min} = p_{max} - p_{min}$ of both the *generic panel distribution* $T(n)$ and also the special re-mark distribution $Q(p)$. Also, one-half of the end-to-end range $h_{max} - h_{min}$ of the ordinary *re-mark distribution* $r(h)$.

$Q(p)$ The **special re-mark distribution**, defining the probability that a script, originally marked $m\dagger$, will be re-marked $\boldsymbol{m^*}$ by a senior examiner such that $\boldsymbol{m^*} = m\dagger + p$. The distribution $Q(p)$ is also the distribution of medians $\mathbf{M}_p$. The end-to-end range of this distribution is N marks, the same as the end-to-end range of the *generic panel distribution* $T(n)$, and one-half of the end-to-end range of the ordinary *re-mark distribution* $r(h)$. The distribution $Q(p)$ is normalised so that the sum

$$\sum_p Q(p) = 1$$

p The number of marks between an original mark $m\dagger$ and a re-mark $\boldsymbol{m^*}$ by a senior examiner such that $\boldsymbol{m^*} = m\dagger + p$, as associated with the *special re-mark distribution* $Q(p)$. The parameter p also defines the number of marks between an original mark $m\dagger$ and the median $\mathbf{M}_p$ of one of the $N + 1$ *generic panel distributions* of which the original mark $m\dagger$ is a member, such that $\mathbf{M}_p = m\dagger + p$.

p_{max} The maximum value of the parameter p for which the *special re-mark distribution* $Q(p)$ is non-zero. The end-to-end range of $Q(p)$ is the difference $p_{max} - p_{min} = N$; also, $p_{max} = - n_{min}$.

p_{min} The minimum value of the parameter p for which the *special re-mark distribution* $Q(p)$ is non-zero. The end-to-end range of $Q(p)$ is the difference $p_{max} - p_{min} = N$; also, $p_{min} = - n_{max}$.

$r(h)$

The **ordinary re-mark distribution**, defining the probability that a script, originally marked m, will be re-marked m^* by an ordinary examiner such that $m^* = m^\dagger + h$. The end-to-end range of this distribution is $2N$ marks, twice the end-to-end range of the end-to-end range of both the special re-mark distribution $Q(p)$ and the *generic panel distribution* $T(n)$. The distribution $r(h)$ is normalised so that the sum

$$\sum_{h} r(h) = 1$$

$t(m)$

The **individual panel distribution**, this being the probability distribution resulting from the marks m given by a panel of examiners to one specific script. The distribution $t(m)$ is normalised so that the sum

$$\sum_{m} t(m) = 1$$

$T(n)$

The **generic panel distribution**, formed by aggregating a sample of *individual panel distributions* $t(m)$, so determining a generic shape which can apply to all submissions within an examination. $T(n)$ has a median $\mathbf{M} = 0$. The end-to-end range of this distribution is N marks, the same as the end-to-end range of the special re-mark distribution $Q(p)$, and one-half of the end-to-end range of the ordinary *re-mark distribution* $r(h)$. The distribution $T(n)$ is normalised so that the sum

$$\sum_{n} T(n) = 1$$

$\delta(h - \mathbf{M}_p)$

The **Dirac δ-function**, which has the value of 1 when $h = \mathbf{M}_p$, and the value of 0 for all other values of h.