The Statistics of Examination Marking and Grading - Overview

Dennis Sherwood, 14th July 2017

Four important distributions

A candidate sits an examination, and the script is given a single mark m by a single marker. On what basis should that script's grade be awarded? It is possible that the single mark m is 'wrong', as may happen if the marker has not complied with the marking scheme, or as a result of a failure in quality control. The assumption made here, however, is that the single mark m is valid, and within the 'tolerance' allowed by all examination boards, and by Ofqual – a 'tolerance' that recognises that different markers might legitimately give the same script different marks.

Ideally, the grade is most fairly determined by the median $\mathbf{M}$ of the distribution of marks given by a panel of markers. In practice, however, this is not possible: any one candidate's script receives only a single mark m, as given by a single marker, and according to current policy, the grade is determined by m, not $\mathbf{M}$. How reliable[1] is m as an estimate of $\mathbf{M}$?

To answer this question, it is helpful to consider four distributions:

$T(n)$: For a script marked by a panel of markers, $T(n)$ - the 'panel distribution' - defines the probability that a single marker will give a mark $m = \mathbf{M} + n$, where $\mathbf{M}$ is the median of the panel distribution, and the fair mark for awarding the grade. This distribution defines the ***uncertainty in marking*** - uncertainty fundamentally attributable to the open-ended nature of GCSE and A level examinations, and as exacerbated by operational failures, such as non-compliance with the marking scheme, or quality control problems.

[1] Figure 14 in Ofqual's November 2016 paper *Marking Consistency Metrics* https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/568424/Marking_consistency_metrics_-_November_2016.pdf suggests that the answer to this question is "not very"...for example, for GCSE French, approximately 85% of candidates are awarded the grade they merit, and 15% are awarded the wrong grade; for GCSE English Language, the figures are about 70% correct, and 30% wrong; for GCSE English Literature, about 54% correct, and 46% wrong. Importantly, and worryingly, the narrowing of the grade widths, and the introduction of more grade boundaries, as associated with the imminent change in the grading system from A*, A, B, C... to 9, 8, 7..., is likely to increase the number of candidates awarded the wrong grade in each subject by some 50% - so it should be expected that about 45% of candidates in GCSE English Language will be awarded the wrong grade, and about 69% in English Literature. See also *Quality of Marking: confidence and consistency* in https://www.gov.uk/government/news/presentations-from-ofquals-summer-series-symposium-2017.

The Statistics of Examination Marking and Grading - Overview

Dennis Sherwood, 14th July 2017

$Q\,(p)$: the probability that the single mark m is a member of the panel distribution of median $\mathbf{M}_p$ such that $\mathbf{M}_p = m + p$. This distribution is important in determining the ***candidate's fair grade***, and also in estimating ***grade reliability***.

$H\,(\mathbf{M}_p)$: the probability distribution of medians $\mathbf{M}_p$, for all possible scripts of all possible medians, across the entire marking range. This distribution has a role in determining the distribution $Q\,(p)$.

$r\,(q)$: the probability that a script originally marked m is re-marked $m*$ (for example, on appeal), such that $m* = m + q$. This distribution is important as regards ***fair appeals***.

The distribution $T\,(n)$

Suppose that a single script is marked by a panel, each member of which gives that script a single valid mark m, generating a distribution $t\,(m)$. The mean $\mathbf{M}$ of this distribution fairly determines the script's grade. Suppose further that, say, nine more scripts are marked by the panel, so generating a total of ten distributions $t\,(m)$. Although the medians of each of these distributions will be different, it is likely that the shapes of the distributions will be broadly similar (this can – and must! - be tested empirically). Accordingly, each of these distributions can be shifted to the median $\mathbf{M} = 0$, and then aggregated and normalised.

The resulting distribution is defined as $T\,(n)$, the shape of which will be referred to as the 'panel distribution'. *This paper now assumes that this distribution is representative of the examination as a whole, and can be applied to any single script* such that $T\,(n)$ specifies the probability that the script will be given a mark $m = \mathbf{M} + n$ by a single marker, where $\mathbf{M}$ is the median of a panel distribution, were it possible for a panel to mark that script. This assumption can also be tested empirically.

A representative panel distribution $T\,(n)$ is shown in Figure 1.

The Statistics of Examination Marking and Grading - Overview

Dennis Sherwood, 14th July 2017

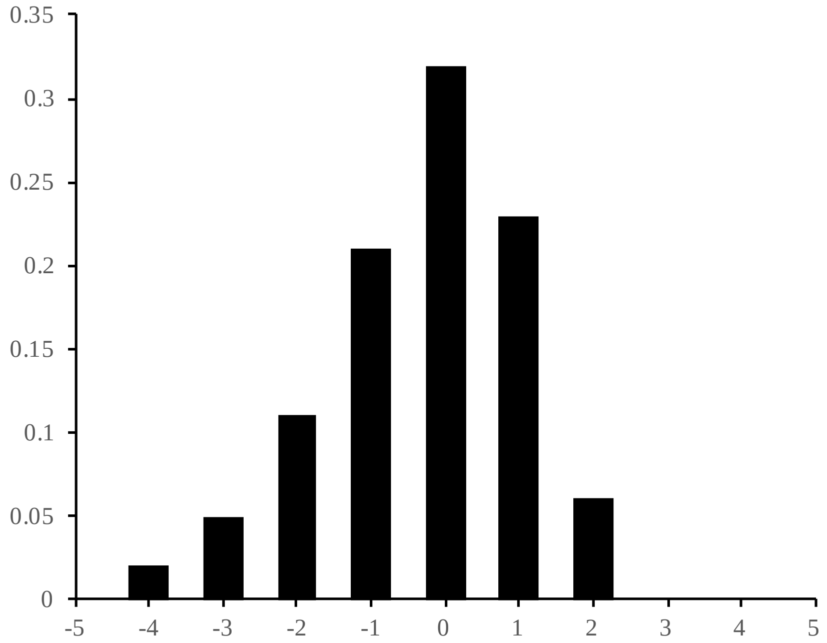


Figure 1: *A representative panel distribution* $T(n)$. This distribution is skewed; it can also be symmetrical.

This distribution has a total end-to-end spread N of 6 marks, extending from $n_{min} = -4$ to $n_{max} = +2$. If the distribution happens to be symmetrical, then the end-to-end spread can be represented as $N = 2f$, from $-f$ to $+f$.

The distribution $Q(p)$

The fundamental practical problem is that, for a single script given a single valid mark m by a single marker, there is no knowledge as to the specific distribution $t(m)$ of which that mark is a member, and therefore of the specific median $\mathbf{M}$ that defines the mark on which that script's grade should fairly be awarded.

But if $T(n)$ can be estimated by the sampling process just described, and if it is valid to equate the shape of $T(n)$ to that of $t(m)$, this limits the possible values of $\mathbf{M}$, as shown in Figure 2:

The Statistics of Examination Marking and Grading - Overview

Dennis Sherwood, 14th July 2017

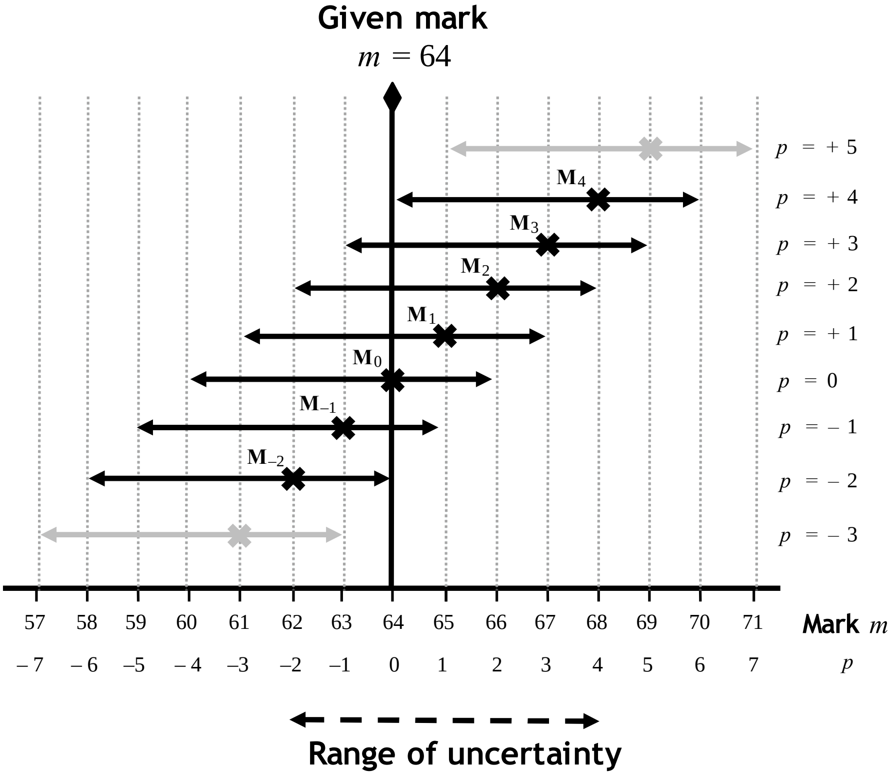


Figure 2: *The uncertainty of the medians* $\mathbf{M}_p$, For an examination for which $T(n)$ is as shown in Figure 1, a script marked $m = 64$ can be a member of only the seven distributions represented by the heavy arrows.

As can be seen, for a script given a single valid mark $m = 64$ by a single marker, and for an examination for which the distribution $T(n)$ shown in Figure 1 is valid, then it is extremely unlikely that the mark $m = 64$ is a member of a distribution for which $\mathbf{M} \leq 61$; likewise, for $\mathbf{M} \geq 69$. It is therefore almost certain that $\mathbf{M}$ is in the range $62 \leq \mathbf{M} \leq 68$ - a range of 6 marks, the same as the end-to-end spread N of the underlying distribution $T(n)$.

Not all of these possible values of $\mathbf{M}$ are equally probable, and the distribution $Q(p)$ defines the probability that the mark m is a member of the distribution of median $\mathbf{M}_p$ such that $\mathbf{M}_p = m + p$. As can be seen from Figure 2, in the particular case shown, the parameter p can take $7 = N + 1$ integer values from a minimum $p_{min} = -2$ to a maximum $p_{max} = +4$ (including $p = 0$), this being 'the other way around' as compared to the underlying distribution $T(n)$, which extends from $n_{min} = -4$ to $n_{max} = +2$. The end-to-end spread N, however, from $p_{min} = -2$ to $p_{max} = +4$, 6 marks, is the same as the end-to-end spread N from $n_{min} = -4$ to $n_{max} = +2$.

If all values of $\mathbf{M}_p$ are equally likely, or if there is no prior knowledge concerning the relative probabilities of the various values of $\mathbf{M}_p$, then

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

As shown in Figure 3, $Q(p) = T(-p)$ has the shape of the 'mirror image' of the distribution $T(n)$ as shown in Figure 1.

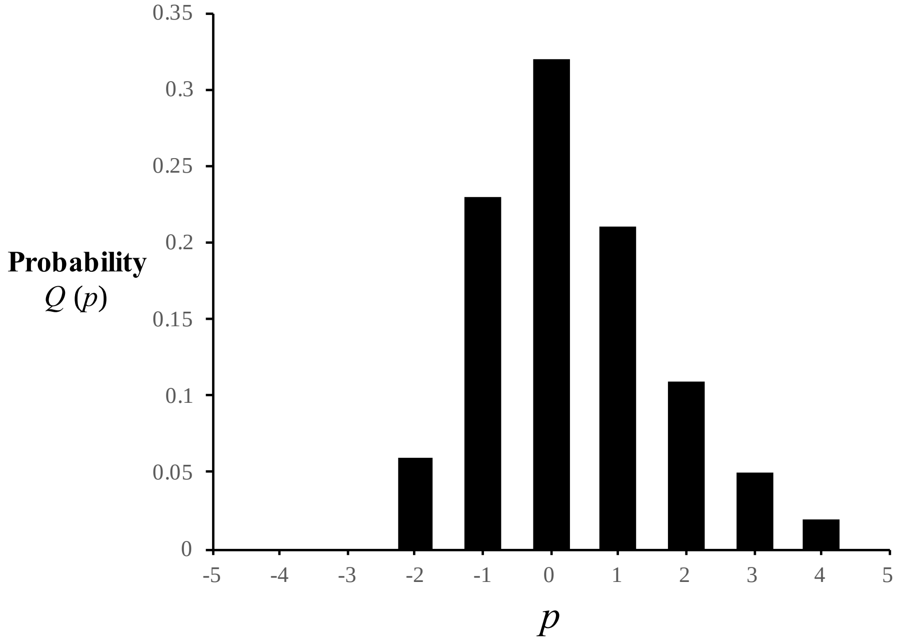


Figure 3: *The distribution* $Q(p)$. If $Q(p) = T(-p)$, then $Q(p)$ is the 'mirror image' of $T(n)$, as shown in Figure 1.

If, however, there is prior information as to the distribution $H(\mathbf{M}_p)$ of the medians $\mathbf{M}_p = m + p$, then, for any script given a single valid mark $m = 64$ by a single marker, Bayesian probability theory states that

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

which reduces to $Q(p) = T(-p)$ if all the values of $H(\mathbf{M}_p)$ over the values of p from p_{min} to p_{max} are equal, or approximately so.

In fact, $H(\mathbf{M}_p)$ can never be known – it is impossible for a panel to mark all scripts. 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. 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 will be shown on pages 8 to 11, for the purposes of fair grading, the important feature of the distribution $Q(p)$ is its end-to-end spread, and the value of p corresponding to the upper limit p_{max} of $Q(p)$. If all the associated values of $H(\mathbf{M}_p)$ are non-zero, then these characteristics of $Q(p)$ are determined directly from $T(n)$, regardless of the behavior of $H(\mathbf{M}_p)$. If only these values are to be estimated, then $Q(p)$ itself does 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 spread of $Q(p)$, making it less than the end-to-end spread of $T(n)$. However, if $H(\mathbf{M}_p)$ is zero, then the corresponding mark $\mathbf{M}_p$ can never be given. This is the case for all marks outside the marking range (say, below 0 or greater than 100), or perhaps for marks at the extreme ends (say, below 3 or greater than 98) – neither of which are important as regards the reliability of grades.

For the purposes of estimating the reliability of grades, however, the details of the distribution $Q(p)$ for all values of p are important, and so $Q(p)$ does need to be determined.

The distribution $r(q)$

Suppose that a script, originally given the single valid mark m by a single marker, is re-marked 'double-blind', or as the result of an appeal, by another marker, who gives the script a mark m^*. If both m and m^* are valid marks, and are not 'wrong' as a result, for example, of a failure to comply with the marking scheme, or quality control problems, then both m and m^* must be members of the same actual distribution $t(m)$ - but once again there is no knowledge as to which specific distribution this is. Reference to Figure 2 will verify that, if $m = 64$, and for an examination for which the distribution $T(n)$ shown in Figure 1 is valid, then m^* is limited to the range $58 \leq m^* \leq 70$, this being twice the end-to-end spread N of the underlying distribution $T(n)$.

Although, in this case, all marks m^* in the range 58 to 70 are possible, they are not equally probable. Accordingly, we may define a distribution $r\,(q)$ which specifies the probability that a script originally marked m is re-marked $m^* = m + q$. The distribution $r\,(q)$ can be obtained by weighting all possible distributions $T\,(p)$ by the probability of their occurrence as defined by $Q\,(p)$, implying that $r\,(q)$ is the convolution (see Appendix 1) of $Q\,(p)$ and $T\,(p)$

$$r\,(q) = Q\,(p) * T\,(p)$$

If, to an acceptable approximation, $Q\,(p) = T(-p)$, then

$$r\,(q) = T(-p) * T\,(p)$$

this being the auto-correlation of the underlying distribution $T\,(n)$, as shown in Figure 4.

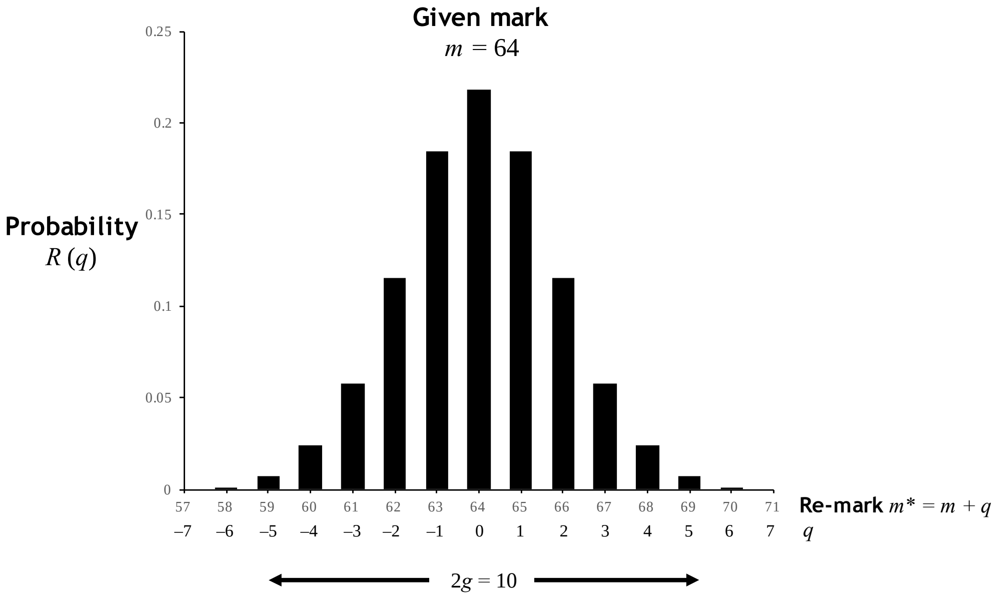


Figure 4: *The distribution* $r\,(q) = Q\,(p) * T\,(p) = T(-p) * T\,(p)$. This distribution defines the probability that a script originally marked m is re-marked $m^* = m + q$, where both m and m^* are valid marks.

If, in addition, the underlying distribution is symmetrical, then $T\,(p) = T(-p)$, and so

$$r\,(q) = T\,(p) * T\,(p)$$

this being the self-convolution of the underlying distribution $T\,(n)$.

As can be seen from Figure 4, even though the underlying distribution $T(n)$ is skewed, the distribution $r(q) = Q(p) * T(p) = T(-p) * T(p)$ is symmetrical, extending, as expected from Figure 2, from $m^* = 58$ to $m^* = 70$, corresponding to values of q from $q = -6$ to $= +6$. Detailed calculations (which are available from the author) show that, in this case, the actual values of $r(q)$ for $q = -6$ and $q = +6$ are $r(-6) = r(+6) = 0.0012$, which are very small - by comparison, for $q = -5$ and $q = +5$, $r(-5) = r(+5) = 0.0076$. Operationally, this implies that, for a script originally marked $m = 64$, there is a chance of about 1 in 1,000 that the script would be re-marked $m^* = 64 - 6 = 58$ or $64 + 6 = 70$, and a chance of about 1 in 100 that the script would be re-marked $m^* = 64 - 5 = 59$ or $64 + 5 = 69$.

As will be seen on pages 12 to 15, the measure of the end-to-end spread $2g$ of the distribution $r(q)$ is important both for fair appeals, and also, as discussed in Appendix 2, for the identification errors in marking attributable to, for example, failure of a marker to comply with the marking scheme, or quality control problems. Accordingly, a policy needs to be adopted as to what percentage of the distribution $r(q)$ should be used to define $2g$, subject to $p_{max} \leq g \leq N$. The remainder of this paper will assume that $2g = 10$, as shown in Figure 4.

Figure 4 also has implications as regards the inferences that can be drawn from a re-mark. Suppose that a script is originally given a single valid mark, say, $m = 64$ by a single marker, and that a re-mark m^* is also 64. Does the fact that the mark and the re-mark are the same 'prove' that the 'true' mark is indeed 64? 'Common sense' might suggest the answer "Of course it does!". But reference to Figure 4 shows (for an examination for which this figure is appropriate) that there is a rather better chance than 1 in 5 that any re-mark m^* will indeed equal the original mark m, and that the uncertainty associated with the corresponding value of $\mathbf{M}_p$ remains unresolved.

Making grading fair

Currently, grades are awarded according to the single valid mark m, as given to a single script by a single marker. Suppose that a particular examination, for which the panel distribution shown in Figure 1 is valid, has the C/B grade boundary at 64/65, and that seven candidates are all given the mark $m = 64$. All candidates are therefore awarded grade C.

Suppose further that each script has been marked by a panel, and that the corresponding panel distributions are known, as shown in Figure 5.

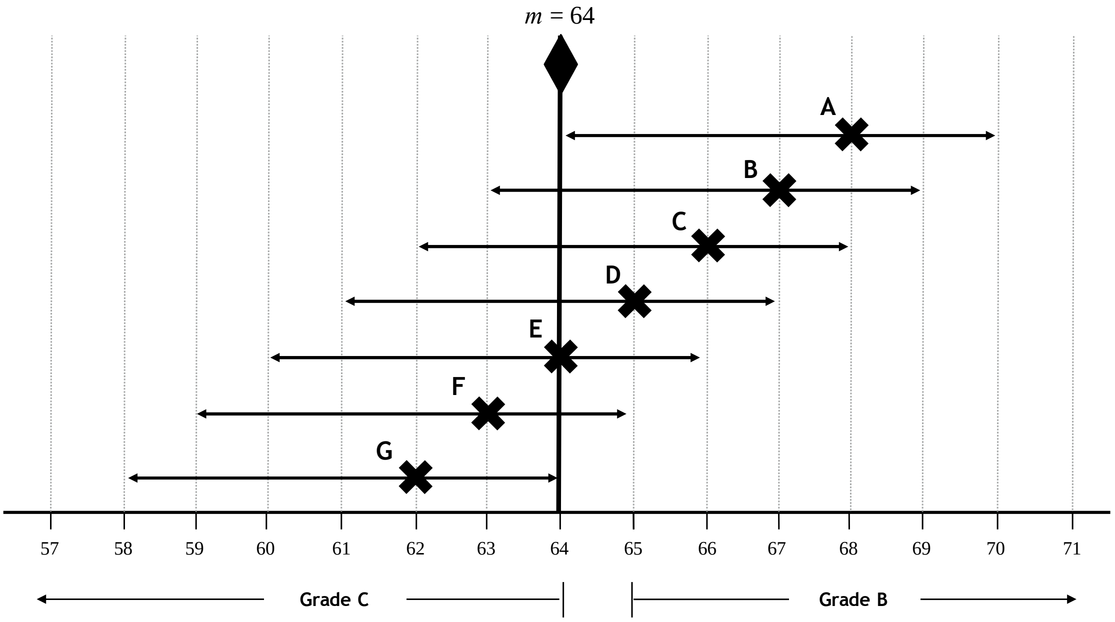


Figure 5: *Unfair grading.* If grades are determined by the mark $m = 64$, then all candidates are awarded grade C. This is fair for candidates E, F and G, but candidates A, B, C and D all merit grade B.

As can be seen, the mark $m = 64$ is a member of all seven distributions, and the award of grade C to candidates E, F and G is fair. For candidates A, B, C and D, however, the fair grade is B, implying that candidates A, B, C and D are 'disadvantaged'. This is an example of the current, unfair, misallocation of grades.

Suppose now that the policy on grading is changed: for the award of the grade, a script marked m is deemed to have the 'adjusted mark' $m + p_{max}$, where p_{max} defines the maximum positive extent of the distribution $Q(p)$, which in the current case is $p_{max} = 4$. If the underlying distribution $T(n)$ is symmetrical, so is the distribution $Q(p)$. Both $T(n)$ and $Q(p)$ therefore extend from a lower limit $-f$ to an upper limit $+f$, and so the 'adjusted mark' is $m + f$.

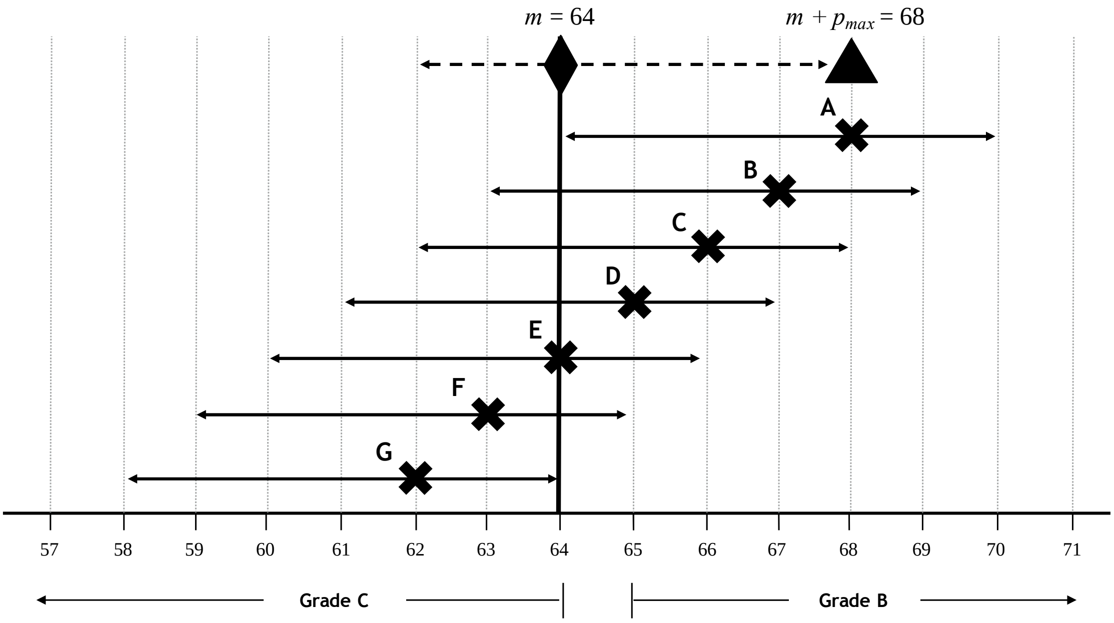


Figure 6: *Fair grading*. If grades are awarded according to the 'adjusted mark' $m + p_{max}$, then all candidates are awarded grade B. No candidates are 'disadvantaged', but candidates E, F and G are 'lucky'.

The impact of this policy is shown in Figure 6: all candidates are awarded grade B. No candidates are 'disadvantaged', but candidates E, F and G are 'lucky'. This demonstrates that although it is impossible to eliminate the uncertainty in marking, it is possible to control the resulting misallocation of grades: under the policy of grading according to $m + p_{max}$, the number of 'disadvantaged' candidates is reduced to close to zero, whilst the number of 'lucky' candidates is increased. Similarly, if grades are awarded according to $m - |p_{min}|$, then the number of 'lucky' candidates is reduced to close to zero, whilst the number of 'disadvantaged' candidates is increased.

The policy choice of grading according to m, $m + p_{max}$ or $m - |p_{min}|$, as illustrated in Figure 7, is important, and the policy debate should be held.

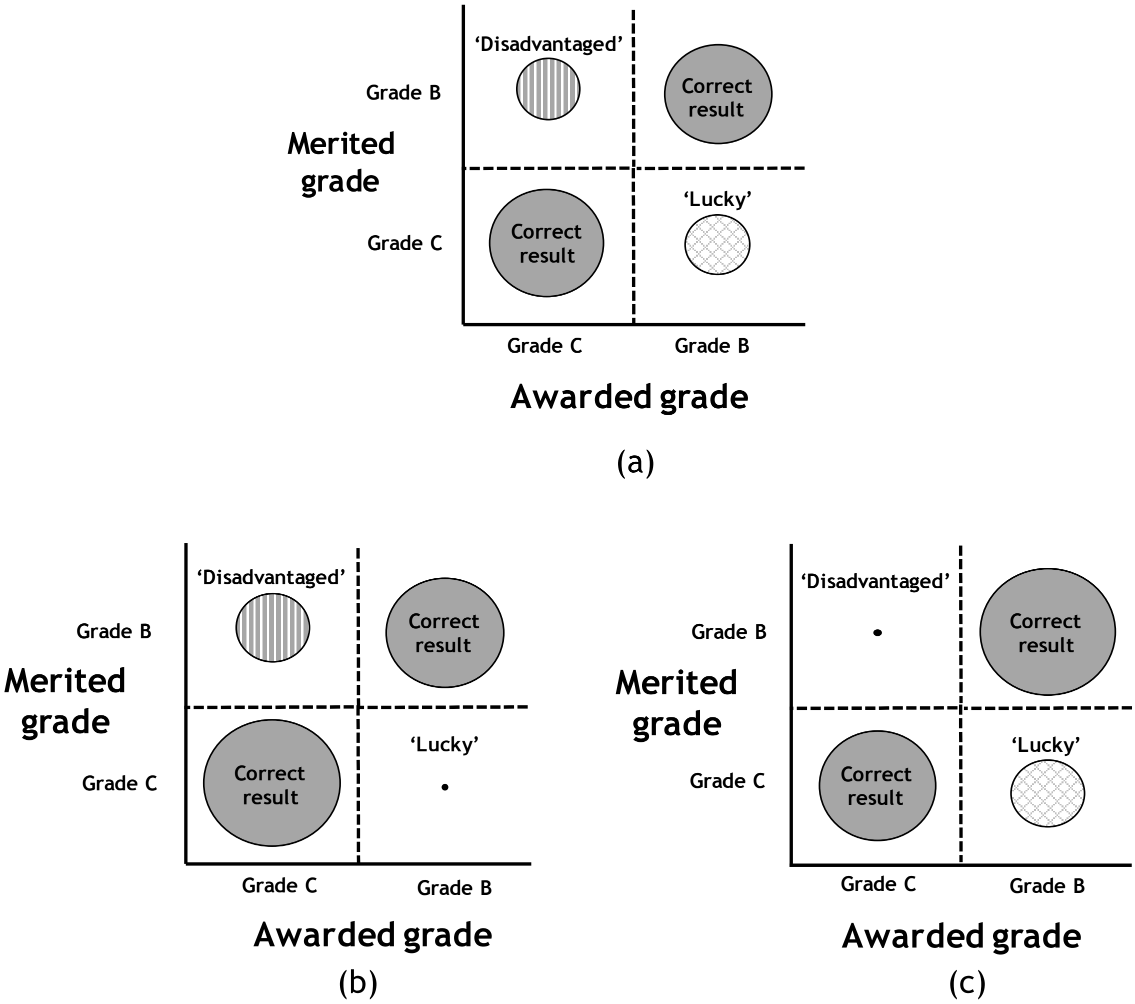


Figure 7: *A matter of policy*. An unavoidable consequence of mapping uncertain marks onto hard-edged grade boundaries is grade misallocation. Although grade misallocation cannot be eliminated, it can be controlled according to the policy by which grades are determined. A policy of awarding grades based on the given mark m results in a mixture of 'lucky' and 'disadvantaged' candidates, as in (a). If grades are awarded according to $m - |p_{min}|$ (or, for a symmetrical distribution $T(n)$, $m - f$), the number of 'lucky' candidates is reduced to close to zero, whilst the number of 'disadvantaged' candidates is increased, as in (b). If grades are awarded according to $m + p_{max}$ (or, for a symmetrical distribution $T(n)$, $m + f$), the number of 'disadvantaged' candidates is reduced to close to zero, whilst the number of 'lucky' candidates is increased, as in (b).

Which is the most appropriate policy for GCSE and A level examinations?

Fair appeals

The Statistics of Examination Marking and Grading - Overview

Dennis Sherwood, 14th July 2017

Even if grades are awarded according to $m + p_{max}$ (or $m + f$), errors in marking - such as failure of a marker to comply with the marking scheme, or a failure in quality control - can still occur. Ideally, these are all identified by the examination board, or by Ofqual, and corrected internally before the results are declared. Even in this ideal, however, it is appropriate that there is a process that allows a candidate to appeal, and that this process operates effectively and fairly.

Over recent years, if a candidate appeals, the paper originally marked m can be re-marked m^*. The re-mark m^* is then assumed to be 'better' than the original mark m, and so a new grade is assigned based on m^*. If both m and m^* are within the same grade width, then the grade is unchanged; if m and m^* are on different sides of the grade boundary, then the grade is changed accordingly. Over recent years, some 18% of appeals[2] have resulted in grade changes, almost all of them being up-grades.

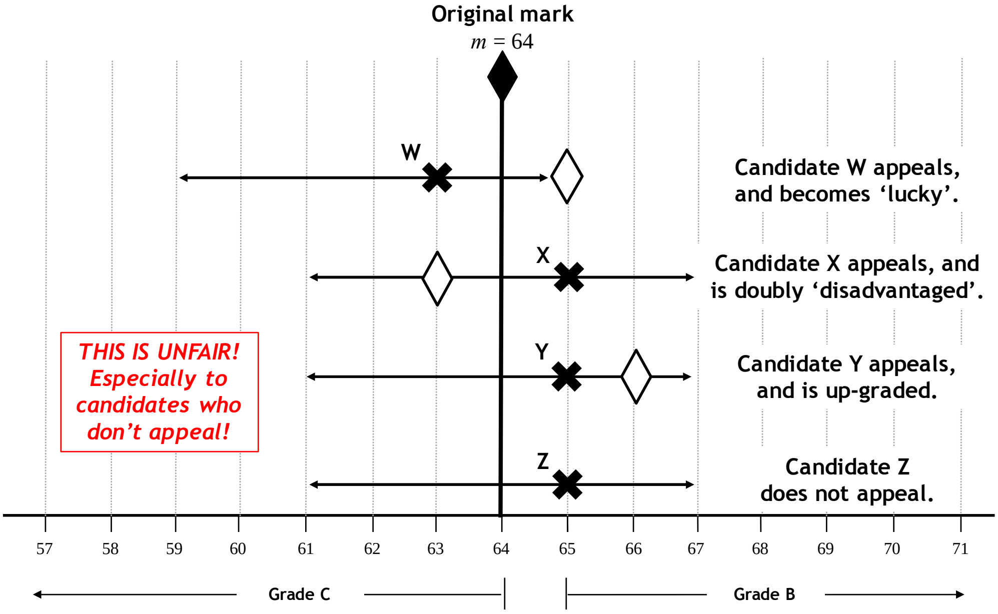

Original mark
m = 64
W
Candidate W appeals, and becomes 'lucky'.
X
Candidate X appeals, and is doubly 'disadvantaged'.
THIS IS UNFAIR! Especially to candidates who don't appeal!
Y
Candidate Y appeals, and is up-graded.
Z
Candidate Z does not appeal.
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
Grade C
Grade B

Figure 8: *Why the 'appeal and re-mark ◊' process is unfair*. Four candidates are given m = 64, and awarded grade C. Candidate W, who was originally given the grade merited, appeals and is up-graded, so becoming 'lucky'. Candidate X appeals, and is doubly 'disadvantaged'. Candidate Y appeals, and is correctly up-graded. Candidate Z does not appeal, and remains 'disadvantaged'.

[2] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/577360/Reviews_of_marking_and_moderation_for_GCSE_and_GCE_summer_2016_.pdf

Figure 8 illustrates the ‘appeal and re-mark’ process under the policy of awarding grades based on the given mark m. All four candidates are given $m = 64$, and all are awarded grade C. As can be seen, candidate W merits grade grade C, as awarded, whereas candidates X, Y and Z were ‘disadvantaged’. Candidate Z does not know this, and accepts the lower grade. The other three candidates, however, raise an appeal, and each script is re-marked $m*$, such that each re-mark $m*$ is within the same panel distribution as the original mark m. Given that the uncertainty in marking straddles the grade boundary, a number of different outcomes are possible, including

Candidate W is up-graded, and so becomes ‘lucky’, having been fortunate in ‘gaming’ the ‘appeal and re-mark’ system.

The award of grade C for candidate X is confirmed, and so that candidate is unfortunate in being doubly ‘disadvantaged’.

Candidate Y is awarded an up-grade, so correctly resolving the original misallocation.

For these four candidates, the ‘appeal and re-mark’ process is fair only to candidate Y. In general, the combination of the policies ‘award grades based on m’, and ‘on appeal, the re-mark $m*$ takes precedence’, is deeply flawed: flawed as a result of the disregard of the uncertainty of marking in the award of the grade based on m, and also in the ‘appeal and re-mark’ process.

The Statistics of Examination Marking and Grading - Overview

Dennis Sherwood, 14th July 2017

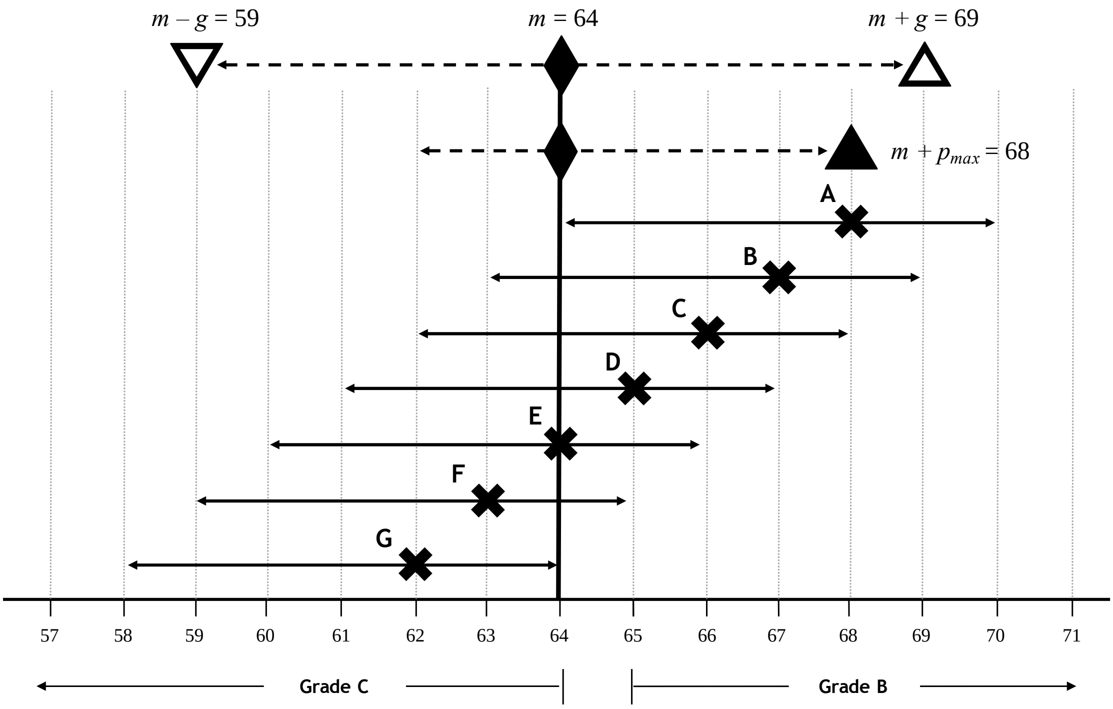


Figure 9: *Fair appeals.* Re-marks m^* within the range $m \pm g$ have already been taken into account by grading according to $m + p_{max}$. If a re-mark m^* is beyond the range $m \pm g$, then the script is re-graded according to $m^* + p_{max}$.

Figure 9 shows how a much fairer appeals process can be designed. Consider a script originally given a valid mark $m = 64$, and awarded a grade based on the 'adjusted mark' $m + p_{max} = 68$. If the script is then given a valid re-mark m^*, then m^* would comply with the probability distribution $r(q)$, where $m^* = m + q$, as shown in Figure 4. This distribution has an end-to-end spread of $2g = 10$, implying that m^* is likely to be within the range $m \pm g$, from a minimum of $64 - 5 = 59$ to a maximum of $64 + 5 = 69$.

Awarding the grade based on $m + p_{max} = 68$ takes this range into account, and so any re-mark m^* within the range $m \pm g$, from $m^* = 59$ to $m^* = 69$, is to be expected. Accordingly, if a script originally marked $m = 64$ is re-marked, say, $m^* = 67$, then that is not necessarily a cause for concern, for the candidate has already been given the 'benefit of the doubt' by being graded according to $m + p_{max} = 68$.

If, however, the re-mark m^* is less than $m - g = 59$, or greater than $m + g = 69$, then a revised 'adjusted mark' $m^* + p_{max}$ should be given, resulting in a grade change if $m^* + p_{max}$ is on the other side of a grade boundary as compared to $m + p_{max}$.

A wise policy for grading and appeals is therefore

Award the grade based on the 'adjusted mark', $m + p_{max}$.

If an appeal is made, and the re-mark is m^*, then if $m - g \leq m^* \leq m + g$, the original grade is confirmed, but if $m^* < m - g$ or $m^* > m - g$, the script is re-graded based on $m^* + p_{max}$. If both $m + p_{max}$ and $m^* + p_{max}$ are within the same grade width, the original grade is confirmed; if $m + p_{max}$ and $m^* + p_{max}$ are on different sides of a grade boundary, the grade is changed accordingly.

Some further points about fair appeals are discussed in Appendix 2.

Grade reliability

As has been shown, for an examination associated with the distribution $T(n)$ illustrated in Figure 1, a script marked $m = 64$ can be associated with a total of $N + 1 = 7$ medians $\mathbf{M}_p$ in the range $\mathbf{M}_{-2} = 62$ to $\mathbf{M}_{+4} = 68$, each of which has a probability given by the corresponding value of $Q(p)$. If an examination is such that 1,000 scripts are marked $m = 64$, then these scripts can be analysed as shown in Figure 10, which assumes that $Q(p) = T(-p)$.

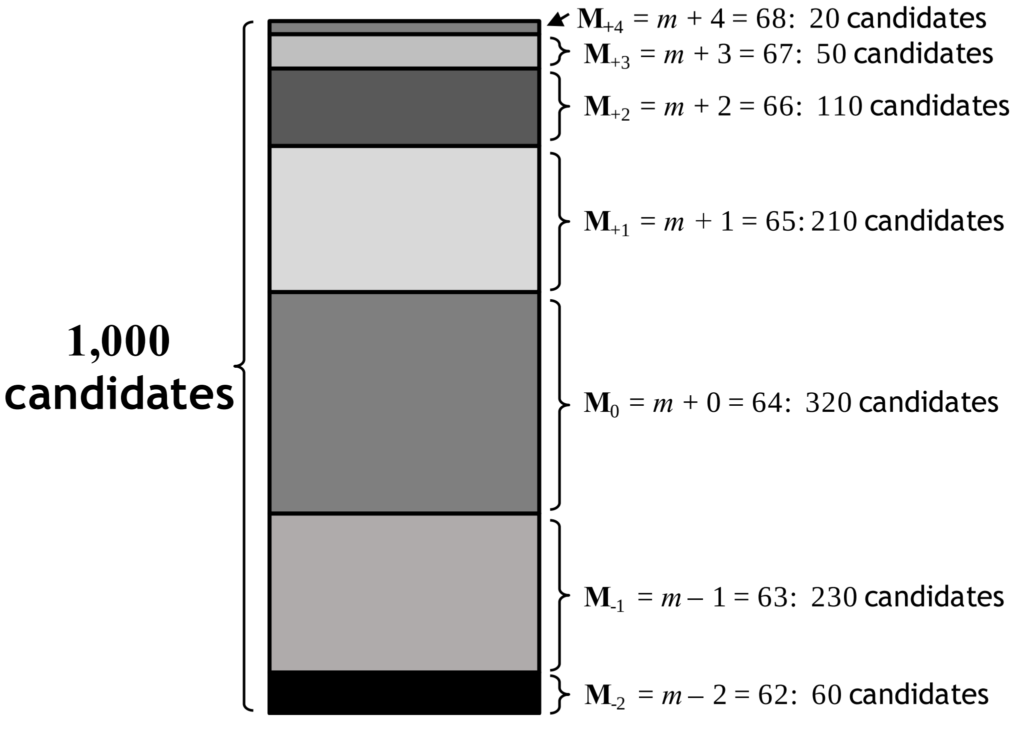


Figure 10: *An analysis of* 1,000 *scripts marked* $m = 64$. This analysis assumes that $Q(p) = T(-p)$, and that $T(p)$ has the shape shown in Figure 1.

An analysis of this type can be carried out for all marks *m*, and if this is done for a range of marks that cross a grade boundary, the result is of the type shown in Figure 11.

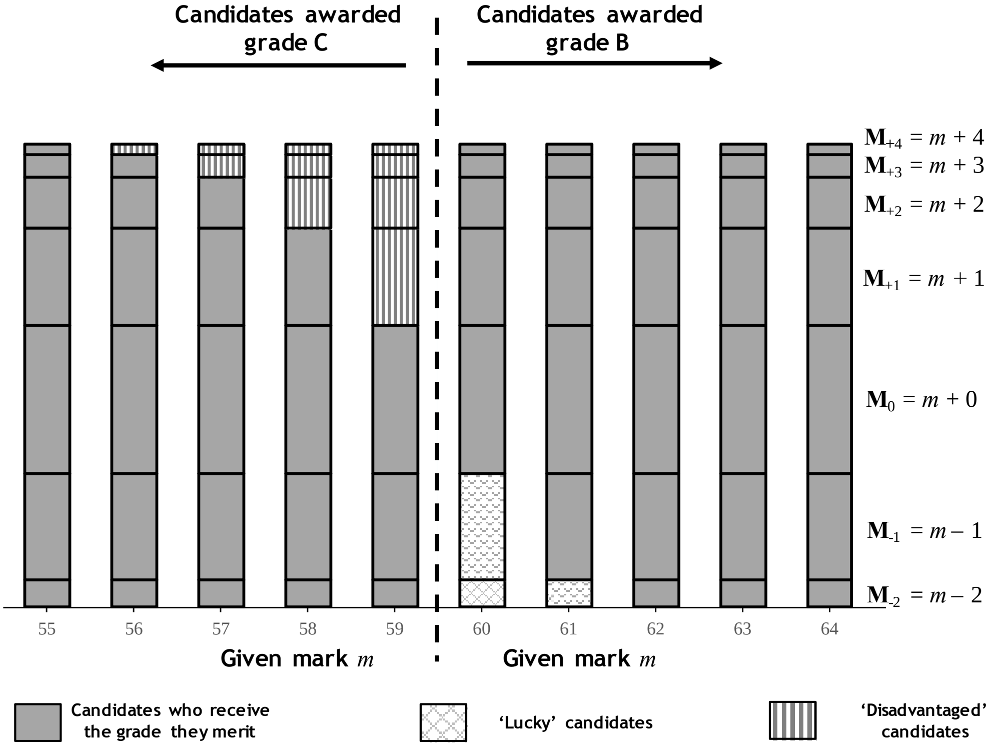


Figure 11: *Grade misallocation quantified*. This diagram applies the analysis shown in Figure 10 to a range of marks *m* across a grade boundary.

This enables the numbers of 'lucky' and 'disadvantaged' candidates to be estimated. Similarly, Figure 12 shows a similar analysis across a single grade width, for grades of two different widths *w*.

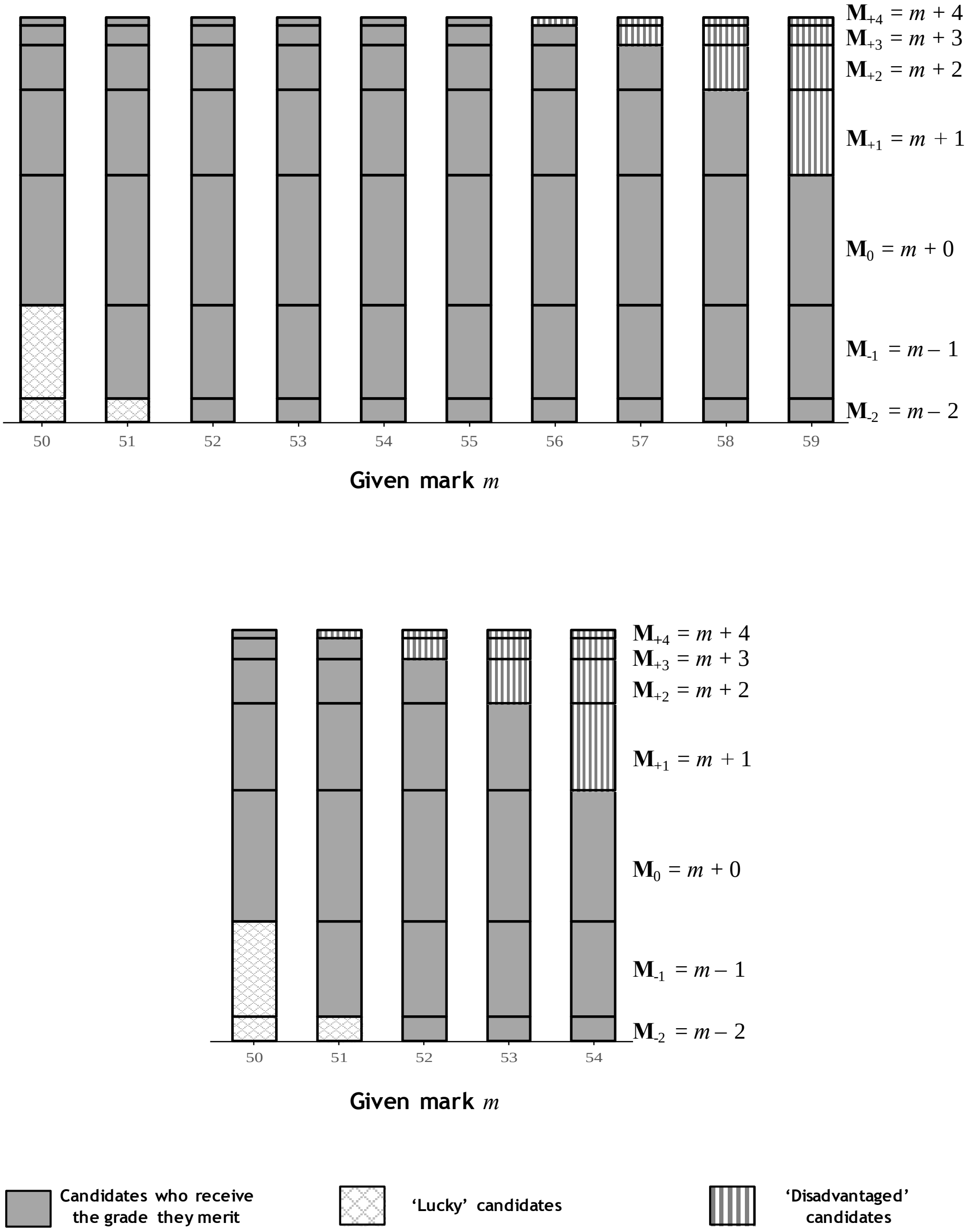


Figure 12: *Grade reliability*. For any given underlying distribution $T(n)$, the narrower the grade width w, the greater the probability that a candidate is awarded the wrong grade.

Knowledge of $T(n)$, and hence $Q(p)$, therefore allows grade reliability to be quantified, for example, as the ratio $w/(w + N)$. As can be seen, the numbers of 'lucky' and 'disadvantaged' candidates depend only on the end-to-end spread N of the distribution $Q(p)$, and are independent of the grade width w. Accordingly, as the grade width w narrows, the probability that a candidate is awarded the wrong grade increases – as is about to happen as a result of changing the grading structure from A*, A, B... to 9, 8, 7...

The Statistics of Examination Marking and Grading - Overview

Dennis Sherwood, 14th July 2017

It's not as difficult as it might appear

The narrative presented here – which attempts to be complete and accurate - does make things appear 'too difficult'. In fact, the key ideas suggested are easy to implement.

The two most important parameters are the end-to-end spread N of the distribution $T(n)$, and the lower limit n_{min}. To determine these, and use them for fair grading and fair appeals...

A small number of scripts are selected at random...

...and each is marked by a randomly selected team of markers (so as to represent the markers overall, rather than just selected experts).

This allows the distribution $T(n)$ to be determined, from which the end-to-end spread N can be estimated. If $T(n)$ is symmetrical, then this spread can be represented as $N = 2f$. If $T(n)$ is skewed, then it is the *lower limit* n_{min} that determines the parameter p_{max}: $|p_{max}| = |n_{min}|$.

All scripts are marked as normal, and each assigned a mark m.

All scripts are then associated with an 'adjusted mark' $m + f$ (or $m + p_{max}$)...

... and it is the 'adjusted mark' $m + f$ (or $m + p_{max}$) that is used for determining the script's grade fairly. This affects every candidate, and is operationally very easy to carry out.

To determine the parameter g, as required for fair appeals, the distribution $r\ (q)$ is the result of the convolution $r\ (q) = Q\ (p)\ *\ T\ (n)$. To a first approximation, $Q(p)$ can be estimated as $T(-\,n)$, the 'mirror image' of the known distribution $T(n)$; if necessary, the Bayes formula can be used (see page 5).

Finally, the reliability of each grade can be determined from the distribution $Q(p)$, as illustrated in Figures 10, 11 and 12.

The Statistics of Examination Marking and Grading - Overview

Dennis Sherwood, 14th July 2017

Appendix 1 - The distribution $r(q)$

A script is given a single valid mark m by a single marker, and a single valid re-mark m^* by another marker. Both m and m^* must therefore be members of the same distribution $t(m)$, which has a median $\mathbf{M}$. Operationally, however, there is no knowledge as to the precise nature of $t(m)$, and the corresponding median $\mathbf{M}$, and so we assume that the shape of $t(m)$ can be approximated as that of the panel distribution $T(n)$, implying that $\mathbf{M}$ is one of the values $\mathbf{M}_p = m + p$ for each of the allowed values of p.

Suppose that the original valid mark m, and a valid re-mark m^*, are in fact members of a particular $T(n)$ of median $\mathbf{M}_p = m + p$, associated with a specific value of p, as illustrated in the inset diagram (which has the same shape as Figure 1, but is 'vertically compressed'). Accordingly, the re-mark m^* must correspond to a value of n such that

$$m^* = \mathbf{M}_p + n = m + p + n$$

and the probability that a script with a (known) original mark m is given a re-mark m^* is the corresponding value of $T(n)$.

If we define q such that

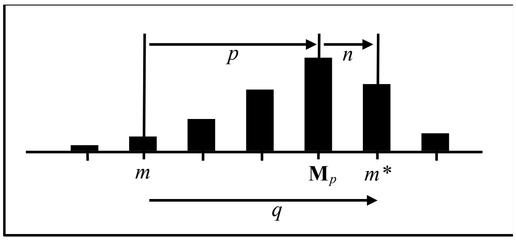
$$m^* = m + q$$

then

$$m + p + n = m + q$$

implying that

$$n = q - p$$



The distribution $T(n)$ of median $\mathbf{M}_p = m + p$ can therefore be expressed as $T(q - p)$.

The probability that a script with a (known) original mark m is given a re-mark $m^* = m + q$ is therefore the value of $T(q - p)$ for a specific value of p. But since the specific value of p is unknown, the probability $r(q)$ that a script given an original mark m will be given a re-mark $m^* = m + q$ is determined by weighting any particular $T(q - p)$ 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

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

This summation is the definition of the convolution $Q(p) * T(p)$. If Bayes may be ignored, $Q(p) = T(-p)$, and this expression becomes the auto-correlation of $T(p)$; if $T(p)$ is symmetrical, it further becomes the self-convolution of $T(p)$.

The Statistics of Examination Marking and Grading – Overview

Dennis Sherwood, 14th July 2017

Appendix 2 – Appeals, and the resolution of errors

This Appendix explores some further issues regarding appeals.

Throughout this paper, it has been assumed that the given mark m is 'valid', and is not 'in error', as might be the case if, for example, the marker has not complied with the marking scheme, or if there has been some failure of quality control. In reality, of course, errors can occur, and it is important that these errors are detected, and subsequently corrected. The examination boards take considerable care to ensure that their processes operate correctly, and that markers comply with the marking scheme. Nonetheless, some errors remain undetected, resulting in the award of the wrong grade.

The 'errors' referred to here are errors in compliance and process, resulting in a mark m that is 'invalid', and so outside of the distribution $T(n)$ that would result from the marks of a panel. Suppose, then, that a script is given as original mark m that is an 'error' in this sense, and that, perhaps as the result of an appeal, the script is re-marked m^*. If the re-mark m^* is valid, then it will be a member of the appropriate panel distribution $T(n)$, but the two marks m and m^* will *not* be members of the *same* panel distribution $T(n)$, and so the distribution $r(q)$, where $m^* = m + q$, will not apply.

If the re-mark m^* is such that $m^* < m - g$ or $m^* > m + g$, then this is good evidence that m and m^* do not belong to the same distribution $r(q)$, and hence the same panel distribution $T(n)$. If it is assumed that the re-mark m^* is valid, then this indicates that there is a high probability that the original mark m is 'in error' - hence the 'fair appeal rule' that the script should be re-graded on the basis of $m^* + p_{max}$.

There are, however, two other possibilities:

The original mark m is valid, and the re-mark m^* is 'in error'.
Both the original mark m and the re-mark m^* are in 'error'.

The common feature here is that the re-mark m^* is 'in error' – a feature that is unlikely if considerable care is taken over the re-mark, and if the re-mark is given by an experienced marker.

As has just been shown, the3re is a high probability that the original mark m is in error if the re-mark m^* is such that $m^* < m - g$ or $m^* > m + g$. There is, however, another question: "Is it possible for the original mark m to be 'in error', but for the valid re-mark m^* to be in the range $m \pm g$?". This question is important, for it raises the possibility that an original error might be 'masked' by the uncertainty in marking, and so remain undetected.

As illustrated in Figure 13, the answer to this question is "yes": this can happen, under particular circumstances.

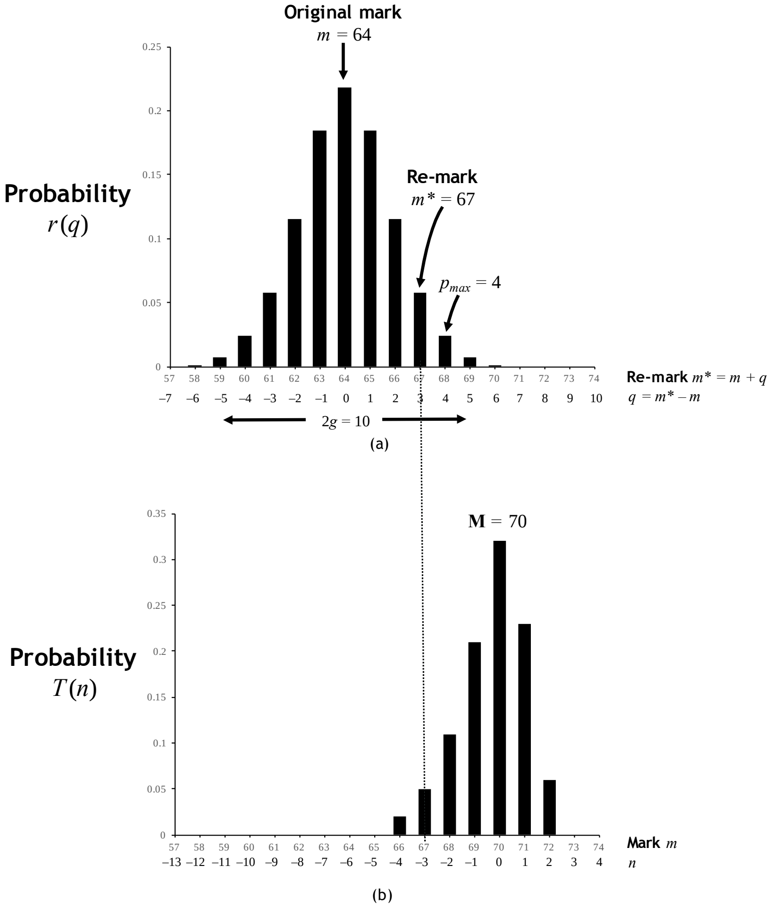


Figure 13: *Masking an original error*. A script is given an original mark m = 64. For an examination for which the distribution $T(n)$ as shown in Figure 1 is valid, the corresponding distribution $r(q)$ of re-marks $m^* = m + q$ is as shown in (a), which is the same distribution as shown in Figure 4, but with the horizontal axis extended towards the right. Suppose that, in fact, the original mark is 'in error', and that the correct panel distribution for the script is as shown in (b), which has a median $\mathbf{M} = 70$, beyond the upper limit $m + p_{max} = \mathbf{M}_{+4} = 64 + 4 = 68$ of the medians associated with the original, erroneous, mark $m = 64$. Suppose further that an appeal results in a re-mark $m^* = 67 = 64 + 3 = m + q$. As can be seen from (a), $m^* = 67$ is towards the right-hand tail of $r(q)$ for $m = 64$, and so is a plausible re-mark for the script. However, as shown in (b), $m^* = 67$ is also a member of the panel distribution for $\mathbf{M} = 70$. The overlap between the high-end tail of $r(q)$ and the low-end tail of a neighbouring $T(n)$ can therefore mask the detection of original marking errors, as can happen, in this case, for the three (= q) panel distributions $T(n)$ of medians $\mathbf{M}$ = 69, 70 and 71.

Figure 13(a) shows the distribution $r(q)$ for a script originally marked $m = 64$ for an examination for which the distribution $T(n)$ as shown in Figure 1 is valid. If the mark $m = 64$ is valid, then any re-mark $m^* = m + q$ of that script must be a member of the distribution $r(q)$, with the value of $r(q)$ giving the probability of its occurrence.

So, for example, a re-mark $m^* = 63 = 64 - 1$ has a probability of occurrence given by $r(-1) \approx 18\%$; a re-mark $m^* = 67 = 64 + 3$ by $r(3) \approx 6\ \%$; and a re-mark of $m^* = 64 = 64 + 0$ by $r(0) \approx 22\ \%$ - this last value showing that the fact that two markers might give the same script the same mark does not 'prove' that this is the 'right' mark, for the uncertainty implied by the shape of the distribution $r(q)$ remains.

Suppose that the script is re-marked $m^* = 67$. Since this is a member of the distribution $r(q)$, albeit of the rather low probability of about 6%, it is plausible that the original mark $m = 64$, and the re-mark $m^* = 67$, are members of the same distribution $r(q)$. This therefore confirms the original mark $m = 64$, and the corresponding grade based on $m + p_{max} = 64 + 4 = 68$ - one mark higher than the re-mark $m^* = 67$ - so everything appears to be fair.

Suppose, however, that the original mark $m = 64$ is 'in error', and that the script is actually associated with the panel distribution shown in Figure 13(b) of $\mathbf{M} = 70$ - a median beyond the upper limit $\mathbf{M}_{+4} = m + p_{max} = 64 + 4 = 68$ of the medians $\mathbf{M}_p = m + p$ associated with the original mark $m = 64$. As can be seen, the re-mark $m^* = 67$ is also a member of the panel distribution of median $\mathbf{M} = 70$. Since the re-mark $m^* = 67$ can *simultaneously* be a member of the distribution $r(q)$ associated with $m = 64$, and also the panel distribution $T(n)$ associated with the median $\mathbf{M} = 70$, there is an ambiguity: it is impossible, on the evidence of $m = 64$ and $m^* = 67$ alone, to distinguish between

Both the original mark $m = 64$ and the re-mark $m^* = 67$, are valid, so confirming the original mark $m = 64$ and the original grade based on $m + p_{max} = 64 + 4 = 68$.

The original mark $m = 64$ is 'in error', and the re-mark $m^* = 67$ is a member of the panel distribution $T(n)$ of median $\mathbf{M} = 70$, implying that the script should be re-graded on the basis of $m^* + p_{max} = 67 + 4 = 71$.

For any value of $m^* = m + q$, this ambiguity arises because the high-end tail of any distribution $r(q)$ overlaps the low-end tails of each of q neighbouring panel distributions $T(n)$ of medians from $\mathbf{M} = m + p_{max} + 1$ to $\mathbf{M} = m + p_{max} + q$; similarly, the low-end tail of any distribution $r(q)$ overlaps the high-end tails of each of q neighbouring panel distributions $T(n)$ of medians from $\mathbf{M} = m - p_{min} - 1$ to $\mathbf{M} = m - p_{min} - q$. Operationally, it is therefore possible that an error in the original mark might be 'masked' by the inherent uncertainty in marking. The policy that grades are based on $m + p_{max}$, and that appeals take into account uncertainty in terms of $m \pm g$, is therefore not 'perfect', for some original errors can remain undetected. This suggests that the policy can be made even more fair, and more likely to allow errors in the original mark to be corrected, by, for example, choosing smaller values of g (probably subject to a minimum of p_{max}; see also page 8), or by ensuring that those (few) scripts re-marked in the 'tails' of the distribution $r(q)$ are given a second re-mark m^{**}, so providing more information, and – hopefully – resolving the ambiguity.

The Statistics of Examination Marking and Grading - Overview

Dennis Sherwood, 14th July 2017

Acknowledgements

It is my pleasure to acknowledge the advice, suggestions and contributions relating to this article that I have received from a wide variety of very helpful and generous people, including Dr Marcus Allen, Richard Backhouse, Paul Baker, Duncan Baldwin, Claire Bracken, Mike Buchanan, Professor Robert Coe, Martin Collier, Rob Eastaway, Shaun Fenton, Peter Hamilton, James Handscombe, Dr Béatrice Harding, Chris Ingram, Stephen Jones, Peter Kent, Dr Michelle Meadows, Steve Miller, Bernard Minsky, Dr Paul Newton, Dr William Richardson, Jon Saunders, Alex Scharashkin, Neil Sheldon, Dr Penny Wickson and Dr Edward Wolfe. I, however, take full responsibility for the content.