The statistics of marking and re-marking

Dennis Sherwood, 21st June 2019, revised 10th March 2023

Author of *Missing the Mark – Why so many school exam grades are wrong, and how to get results we can trust*, published by Canbury Press, 2022

Contents

Why a statistical analysis is needed

This document[1] presents the mathematics and statistics of examination marking and re-marking, so providing an analytical explanation of much of the data on the reliability of school examination grades, as presented in two recent Ofqual reports[2]. 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 $m\dagger$ represents a specific mark (say, 59) given to a particular script.

[1] This document is a revised and improved version of a document originally posted on 14th July 2017; it is also the Appendix to a comprehensive study of grade (un)reliability, available here.

[2] *Marking Consistency Metrics* (November 2016), and *Marking Consistency Metrics – An update* (November 2018),

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 1.

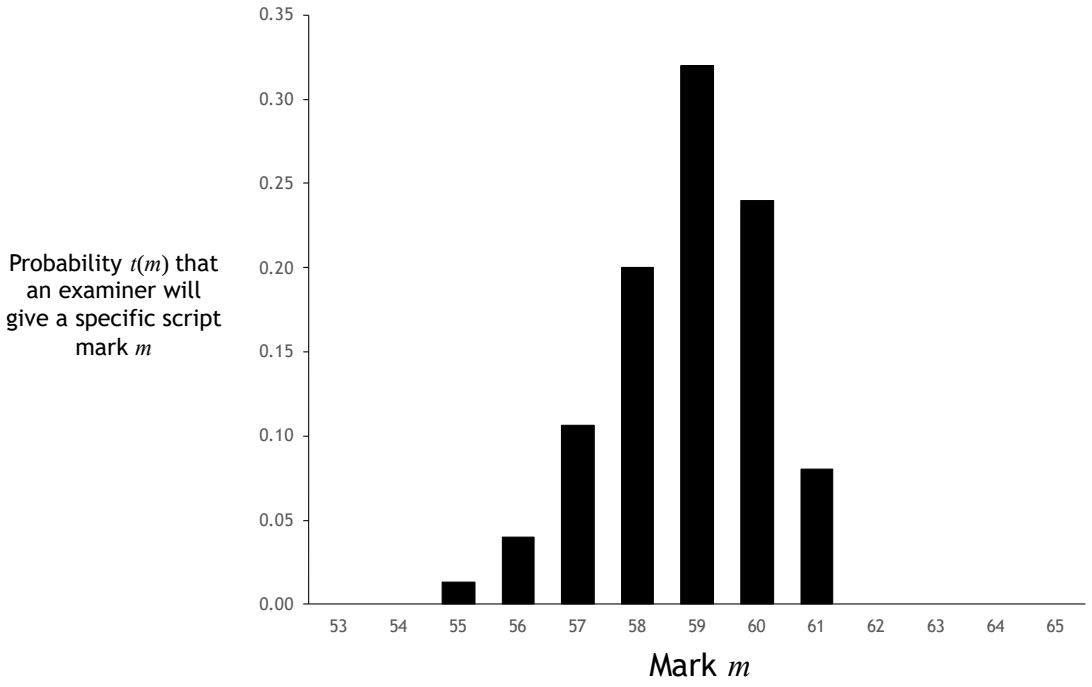
Table 1: Marks given by 75 different examiners to the same script

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 1 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 1:

Figure 1: The probability distribution $t(m)$ for the data shown in Table 1



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 1, and as represented by the height of the corresponding column in Figure 1. 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 1, and as illustrated in Figure 1 :

- 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 1 , there were 75 examiners: the 'middle' examiner is therefore the 38$^{\text{th}}$, and, as can be seen from the 'cumulative' column in Table 1 , among the 24 examiners who gave the script 59 marks. The median of the distribution shown in Figure 1 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 1), 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 1 , 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 1 , 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 $t(m)$

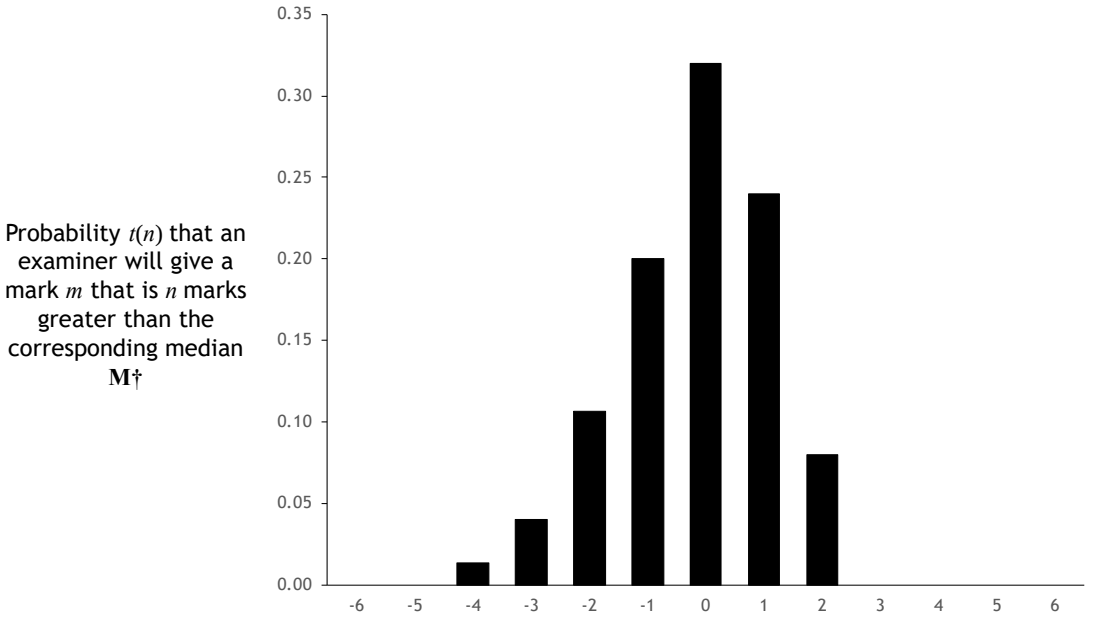
As shown in Figure 1 , the distribution $t(m)$ extends from $m_{min} = 55$ marks to $m_{max} = 61$ marks, with median $\mathbf{M}\dagger = 59$ marks, where the composite symbol $\mathbf{M}\dagger$ indicates that this median is specific, being the median of that particular distribution $t(m)$ of which the mark m is a member.

If a new variable n is defined such that

$$m = \mathbf{M}\dagger + n$$

then n represents the number of marks by which any mark m is greater than the median $\mathbf{M}\dagger$ 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 2.

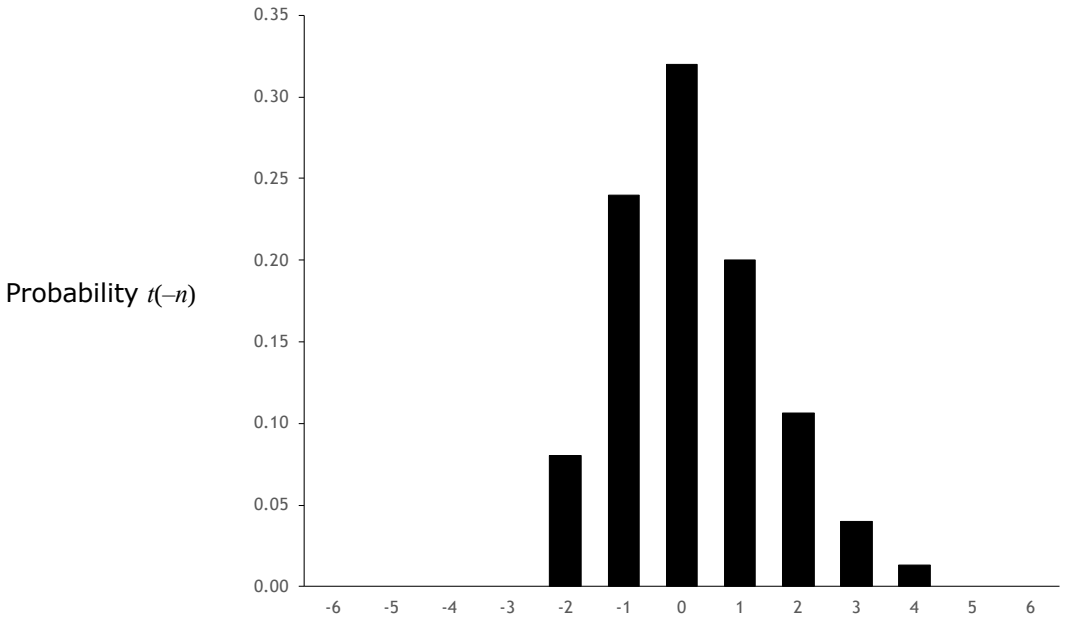
Figure 2: The distribution $t(n)$



To verify this, consider the specific value $n = -3$, for which, according to Figure 2, $t(n) = t(-3) = 0.04$. Since $m = \mathbf{M}\dagger + n$, then, for $\mathbf{M}\dagger = 59$, $m = 59 - 3 = 56$. According to Figure 1, $t(m) = t(56) = 0.04$, 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 1, and $t(n)$ straddling the median $\mathbf{M} = 0$, as shown in Figure 2.

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 2 (and Table 1), 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 3, which, as can be seen, is the left–right mirror image of the shape of the distribution $t(n)$ as shown in Figure 2.

Figure 3: The distribution $t(-n)$



Number of marks n by which the given mark m is greater than the median $\mathbf{M}\dagger$ of the specific individual panel distribution $t(m)$ of which the mark m is a member, such that $m = \mathbf{M}\dagger + n$

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 2), then $t(n)$ and $t(-n)$ can be distinguished, as shown by comparing Figures 2, for $t(n)$, and 3, 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 10 to 13, 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

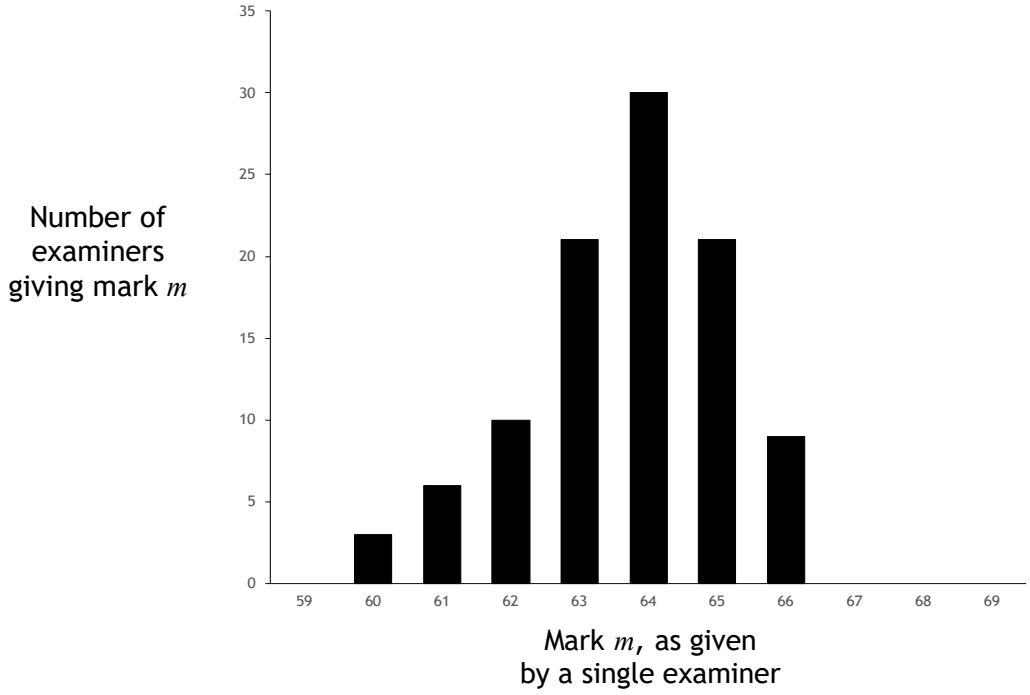
by a senior examiner; it is the ordinary re-mark distribution $r(h)$, however, that provides a realistic method of measuring grade reliability in practice.

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 4 (which is superficially similar in shape to the distribution shown in Figures 1 and 2, but is in fact different).

Figure 4: 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 4 shows the distribution of marks m given by each of 100 different examiners to the same script; strictly speaking, however, Figure 4 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 2 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 4 happens not to be left-right symmetrical, but in practice it often is. Whatever the shape might be, as discussed on pages 5 and 6, the distribution is always associated with a number of statistical characteristics, for example, for the distribution shown in Figure 4:

- 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 4 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 4 represents the marks given by 100 examiners, each marking the same script once. Following the same line of reasoning as on page 3, 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 4 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 4 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 4.

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, 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.[3] 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.

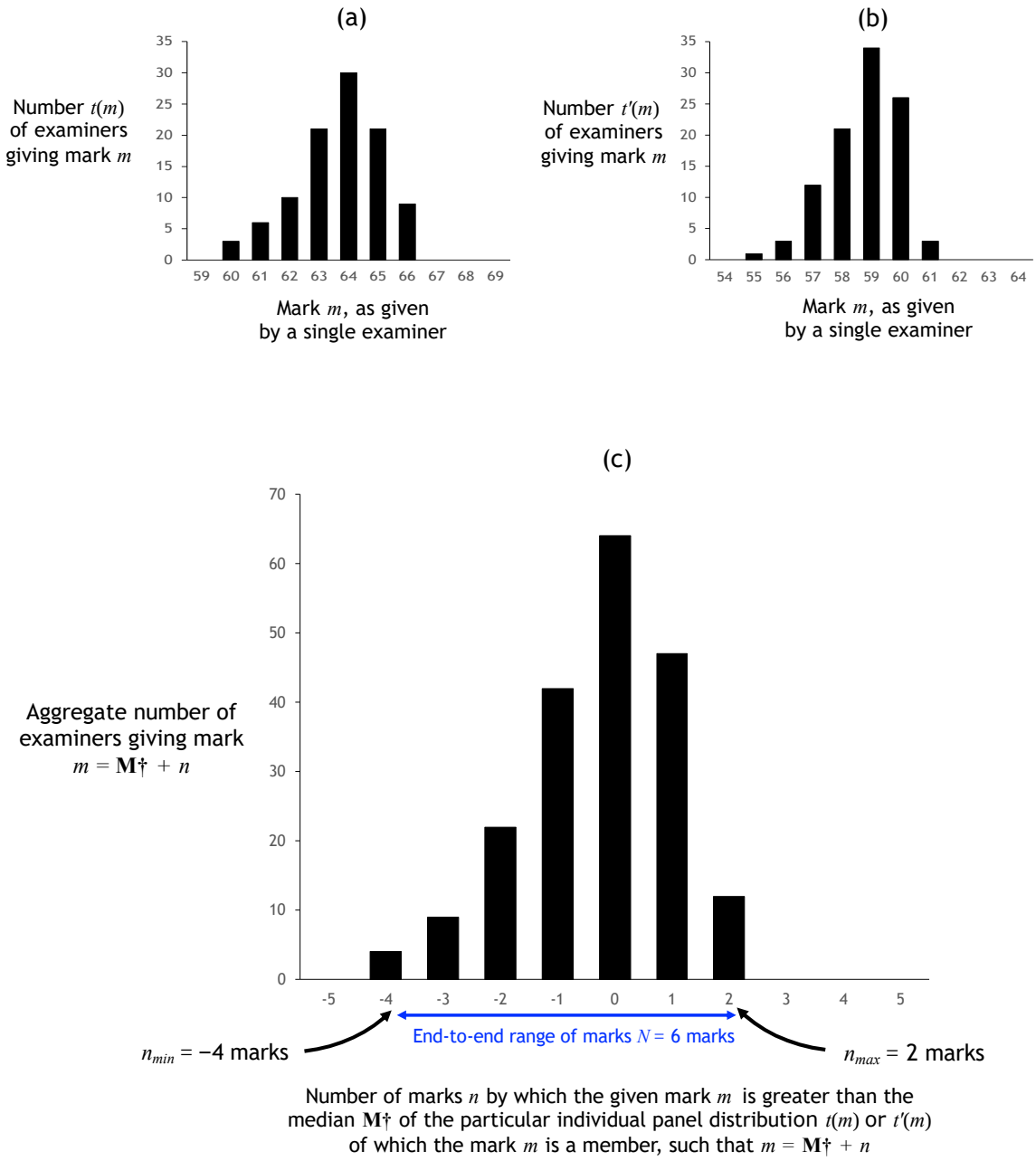
A central theme of Ofqual's November 2016 and November 2018 reports, however, is the use of a senior examiner's mark as a reference point, defining the 'definitive' mark and the corresponding 'definitive' grade – and Figures 12 and 13 of the November 2016 report even refer to the 'true grade'. In the 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 4 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 5:

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

Figure 5: Aggregating two individual panel distributions $t(m)$ and $t'(m)$



The individual panel distribution $t(m)$ shown in Figure 5(a) is the same as that shown in Figure 4, with a median $\mathbf{M}\dagger = 64$; 5(b) is the individual panel distribution $t'(m)$ for a second script, with median $\mathbf{M}\dagger = 59$, 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 5(c).

In Figure 5(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 5(a) and 5(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 5(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 5(c), would show a number of columns (say, seven, as in Figure 5), 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 6, with the corresponding numerical values in Table 2.

Figure 6: The generic panel distribution $T(n)$

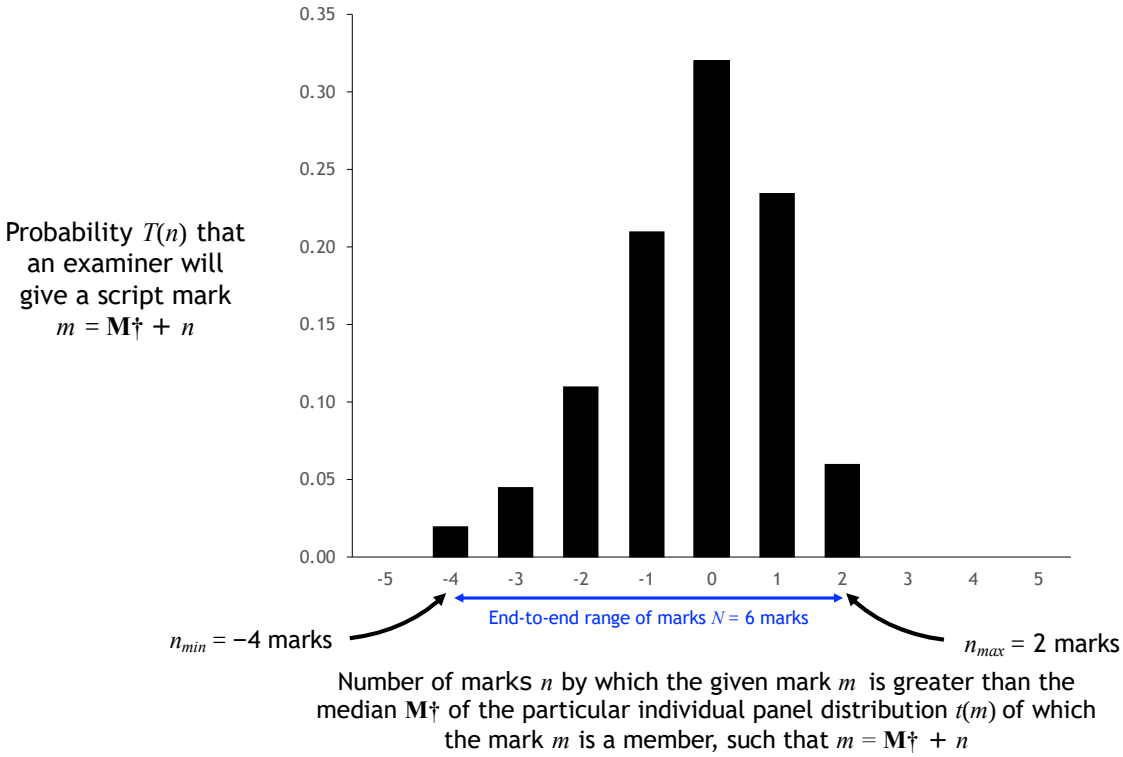


Table 2: Numerical values corresponding to the generic panel distribution $T(n)$ shown in Figure 6

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
n	Probability $T(n)$	

As can be seen from the total in Table 2, 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 6, 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 6:

- 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. It is, however, 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 5 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 6 and Table 2 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 $Q(p)$

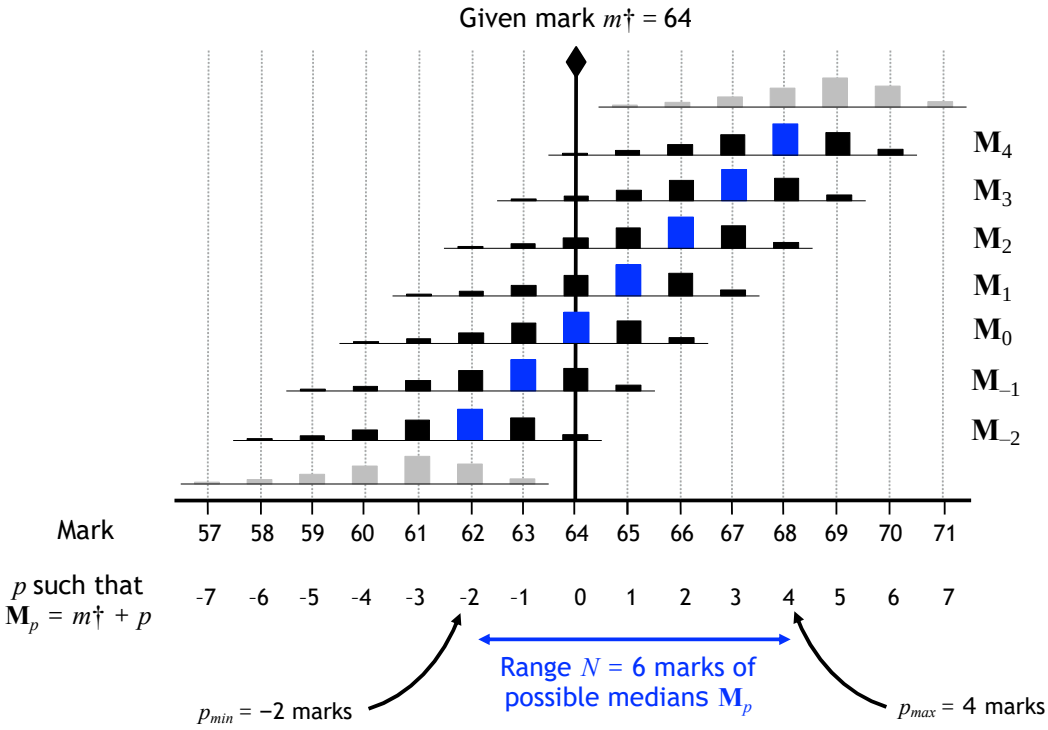
The medians $\mathbf{M}_p$

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 (as indeed it can, as described on pages 55, 89 and 92 here), 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 7:

Figure 7: The uncertainty of the medians $\mathbf{M}_p$ for the generic panel distributions, of the form shown in Figure 6, 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 6. 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 6

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 $m\dagger = 64$.

This is illustrated in Figure 7, which shows the given mark $m\dagger = 64$, and also (rather vertically compressed) representations of the all the generic panel distributions of the shape shown in Figure 6, and with medians $\mathbf{M}$ from $\mathbf{M} = 61$ to $\mathbf{M} = 69$.

Since the given mark $m\dagger = 64$ 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 $\mathbf{M}\dagger$ of the specific generic panel distribution of which the given mark $m\dagger = 64$ is a member lies in the range $62 \leq \mathbf{M}\dagger \leq 68$ 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 7 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 $m\dagger = 64$; 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 $m\dagger = 64$, and so it is possible that a script marked $m\dagger = 64$ might be a member of this distribution, in which case the 'right' mark for that script is $\mathbf{M} = 63$. As Figure 7 vividly shows, however, this is not the only possibility: the distribution $T(n)$ associated with the median $\mathbf{M} = 67$ also includes $m\dagger = 64$, 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 $m\dagger = 64$, 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 $\mathbf{M}_p = m\dagger + p$. Accordingly, when $p = 2$, $m\dagger + p = 54 + 2 = 56$, 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 6, which shows the generic panel distribution $T(n)$ on which Figure 4 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 7 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 7, 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 $Q(p)$

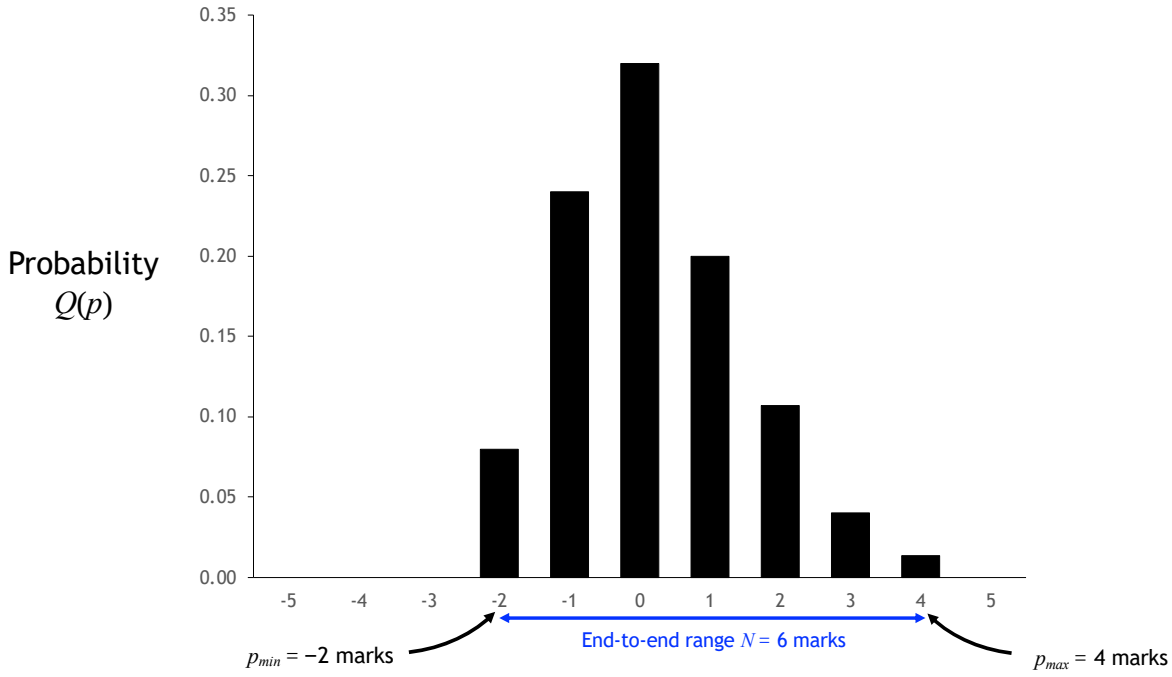
To determine $Q(p)$, consider an example of an examination subject for which the Figures 6 and 7 apply, and the particular case of a script marked $m\dagger = 64$ which is in fact a member of the generic panel distribution for which the median $\mathbf{M}\dagger = \mathbf{M}_2 = 66 = 64 + 2$, implying that $p = 2$. What is the corresponding probability $Q(2)$?

Reference to Figure 7 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 6, 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 7 which shows that the mark $m = 64$ lies two marks to the left of the median $\mathbf{M}_2$.

According to Figure 6, 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 6 and 7, 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 8:

Figure 8: The distribution $Q(p)$ of the medians $\mathbf{M}_p$ shown in Figure 7



Number of marks p by which the given mark m is greater than the median $\mathbf{M}_p$, such that $\mathbf{M}_p = m\dagger + p$, assuming that all values of $\mathbf{M}_p$ are equally likely.

Figure 8 is consistent with Figure 7, with each of the columns in Figure 8 corresponding to the equivalent median $\mathbf{M}_p$ as shown in blue in Figure 7. Figure 8, however, not only identifies the range of possible values of $\mathbf{M}_p$ (as does Figure 7), 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 8 and 6, 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 8, 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.

Quantifying grade reliability

In Figure 8, the height of the column for any value of p represents the probability that a script, given a single mark $m\dagger$ by a single examiner, is a member of the generic panel distribution characterised by the median $\mathbf{M}_p = m\dagger + p$. 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 $m\dagger$ by a single examiner, what is the probability that the 'right' mark for this script is $\mathbf{M}_p = m\dagger + p$?'. 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 $m\dagger$ 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 $\boldsymbol{m}^* = m\dagger + p$?' – where the bold symbol $\boldsymbol{m}^*$ indicates that the re-mark is done by a senior, and not by an ordinary, examiner.

The distribution $Q(p) = T(-p)$, as exemplified by Figure 8, is therefore very important as regards quantifying grade reliability. Assuming for the moment that the median $\mathbf{M}\dagger$ of a specific script's generic panel distribution $T(n)$, as illustrated in Figure 6, 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 8.
- For a script given any mark $m\dagger$, the probability that the 'right' mark is $\mathbf{M}_p$, such that $\mathbf{M}_p = m\dagger + p$, is given by the corresponding value of $Q(p)$, as exemplified by the height of the corresponding column in Figure 8.

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 $\mathbf{M}$ 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 6, then Table 3 shows the probabilities that the 'right' mark is one of the seven possibilities from 62 to 68.

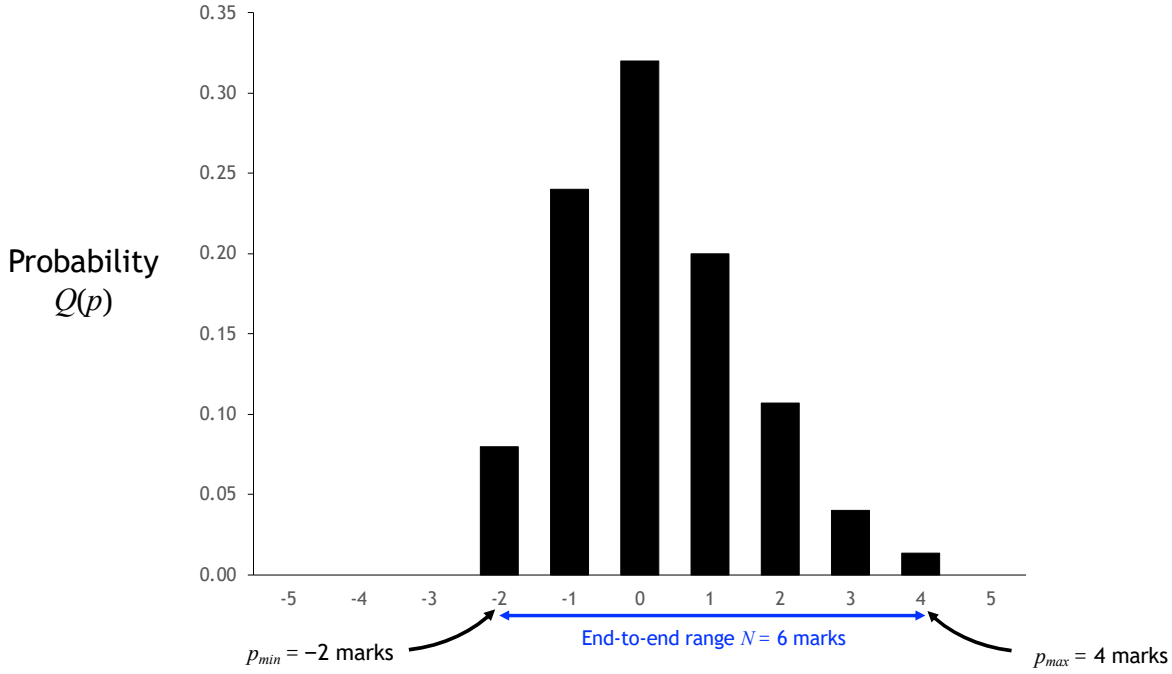
Table 3: The probability $Q(p)$ that a script originally marked $m\dagger$ = 64 is associated with a particular 'special' mark = $m\dagger + p$, these being the medians of successive individual panel distributions as illustrated in Figure 7

'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 3, 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 9: The special re-mark distribution, $Q(p)$ defining the probability that a script originally marked $m\dagger$ will be re-marked $\boldsymbol{m^*} = m\dagger + p$ by a senior examiner

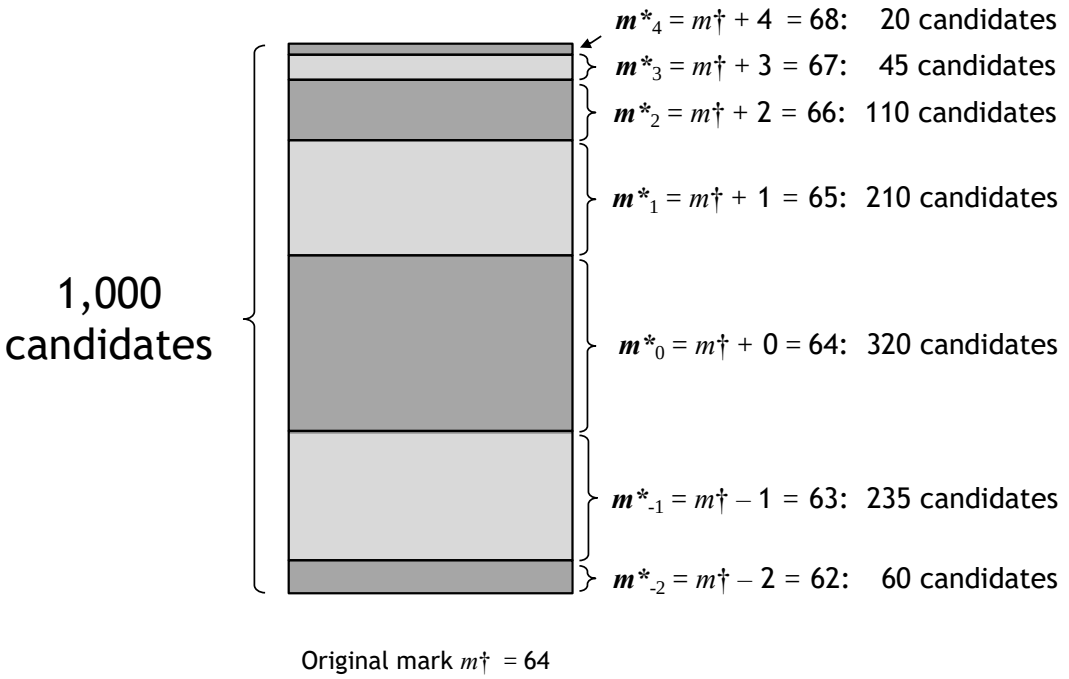


Number of marks p by which a re-mark $\boldsymbol{m^*}$ by a senior examiner is greater than the original mark $m\dagger$, such that $\boldsymbol{m^*} = m\dagger + p$.

$Q(p)$ and grade reliability

Suppose that the special re-mark distribution $Q(p)$ as illustrated in Figure 9, and the associated data as shown in Table 3, are known, and valid for a particular examination subject. Suppose further that 1,000 candidates are marked $m\dagger = 64$ 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 10:

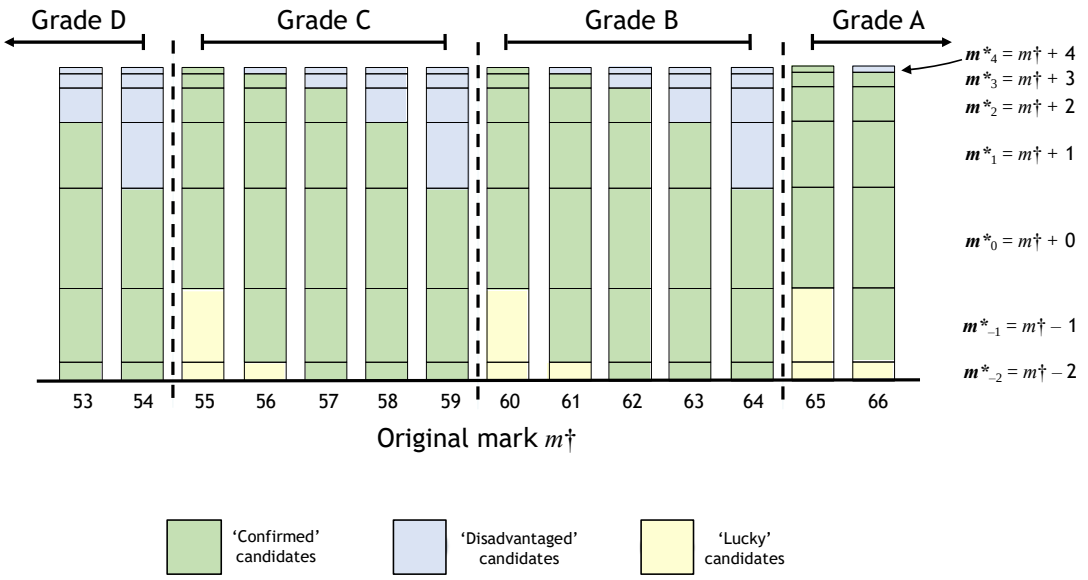
Figure 10: Re-marks $\boldsymbol{m}^*$ *by a senior examiner for a cohort of 1,000 candidates, all given an original mark* $m\dagger = 64$*, for an examination for which the special re-mark distribution* $Q(p)$ *as shown in Figure 9 is valid*



In Figure 10, 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 7; 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 11 brings together a series of representations of the type illustrated in Figure 10 for a sequence of marks from 53 to 66, for an examination with the grade boundaries as shown.

Figure 11: 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 $m\dagger$, 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 $m\dagger$ implies that any given layer has the same thickness across the diagram.

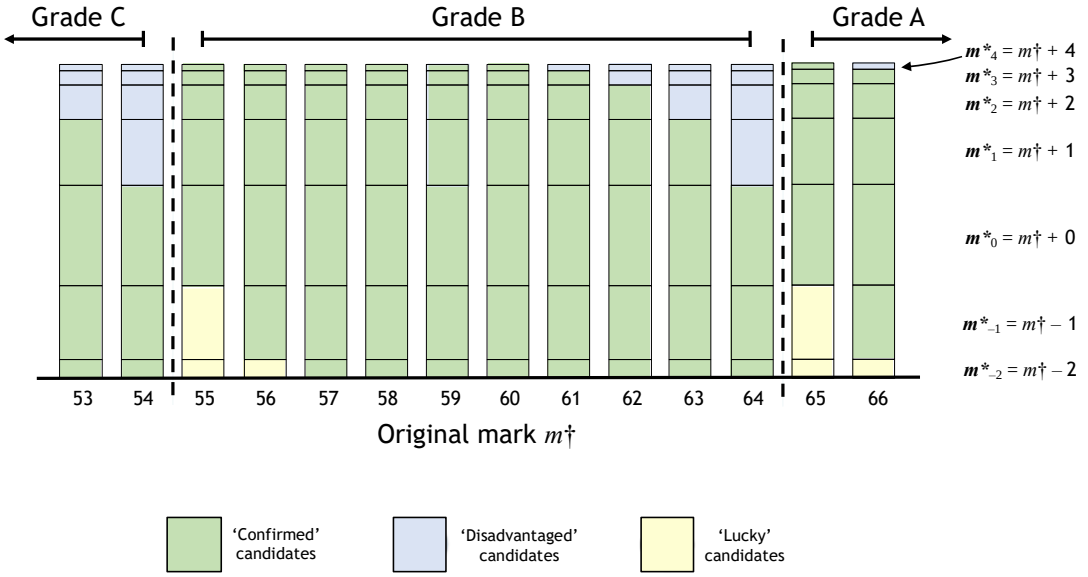
Taking as an example those candidates all originally given $m\dagger = 64$, and all awarded grade B, the bottom three layers represent the numbers of candidates whose scripts, if re-marked by a senior examiner, would be given 62 (60 candidates), 63 (235 candidates) or 64 (320 candidates), all of whom 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; these 385 candidates may therefore all be regarded as 'disadvantaged' (see pages 60 and 61 here). Of the 1,000 candidates originally marked $m\dagger = 64$, 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 $m\dagger = 64$ is therefore 61.5%.

For the 1,000 candidates originally given $m\dagger = 60$, 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 $m\dagger = 60$ is therefore 70.5%.

For the 5,000 candidates originally awarded grade B, Figure 11 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 $m\dagger$ from 55 to 64 inclusive, corresponding to a total of 10,000 candidates.

Figure 12: The effect on grade reliability of grade width



It is immediately evident visually that, compared to Figure 11, 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 12, 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 12 is therefore 90.1%, compared to an average reliability of 80.2% for the narrower grade B shown in Figure 11.

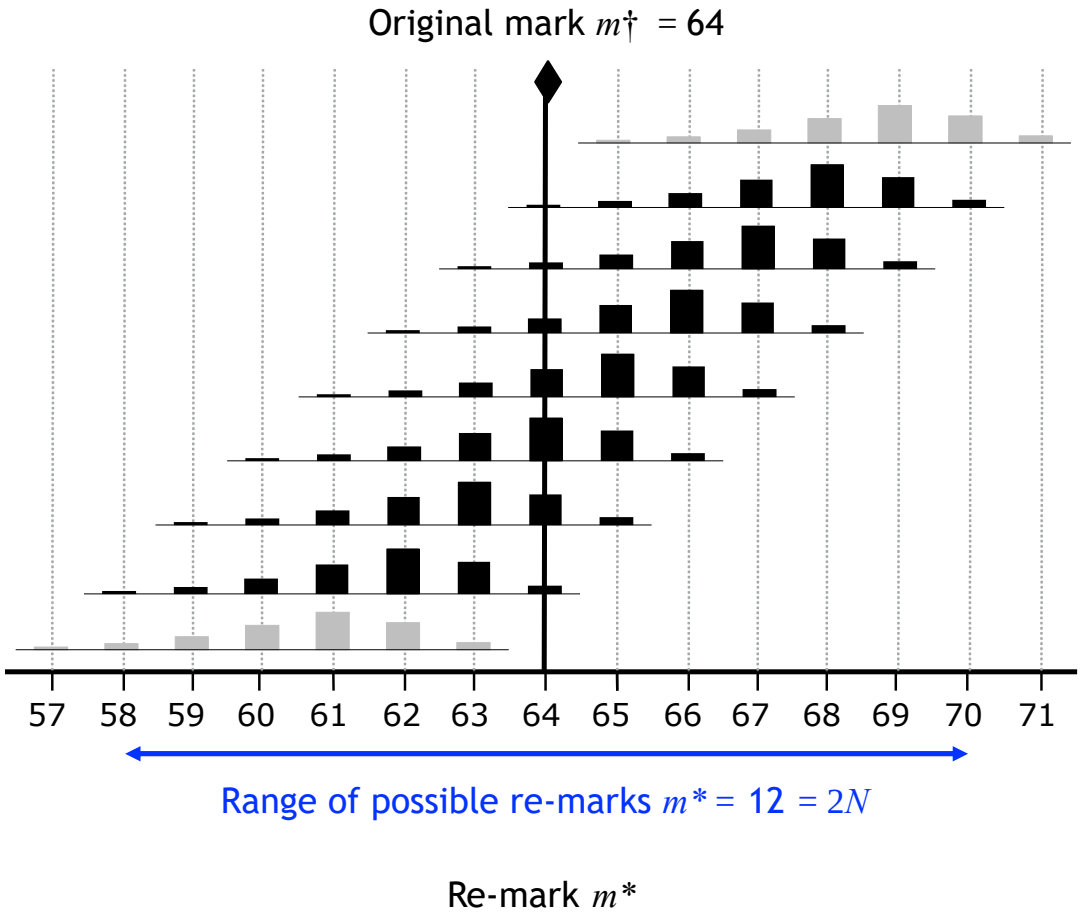
The ordinary re-mark distribution $r(h)$

The distribution $r(h)$

The special re-mark distribution $Q(p)$, as exemplified by Figure 9, 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 13.

Figure 13: Re-marking by an ordinary examiner



For an examination for which $T(n)$ is as shown in Figure 6, a script originally marked $m\dagger = 64$ can be a member of any of the seven distributions represented by the 'vertically squashed' representations of the distribution, as shown in Figure 13. 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 7.

As shown in Figure 13, for an examination characterised by the generic panel distribution $T(n)$ as illustrated in Figure 6, a script originally marked $m\dagger = 64$ 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 $m\dagger$ is re-marked $m* = m\dagger + h$ 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 37 to 39) that $r(h)$ is known mathematically as the 'convolution' of $Q(p)$ and $T(p)$, represented by the symbol $*$ as

$$r(h) = Q(p) * T(p)$$

Also, as discussed on pages 22 to 24, $Q(p) = T(-p)$, and so

$$r(h) = T(-p) * T(p)$$

this being known mathematically as the 'auto-correlation' of the underlying distribution $T(n)$, as shown in Figure 14, with the corresponding numerical values in Table 3.

Figure 14: The ordinary re-mark distribution $r(h)$

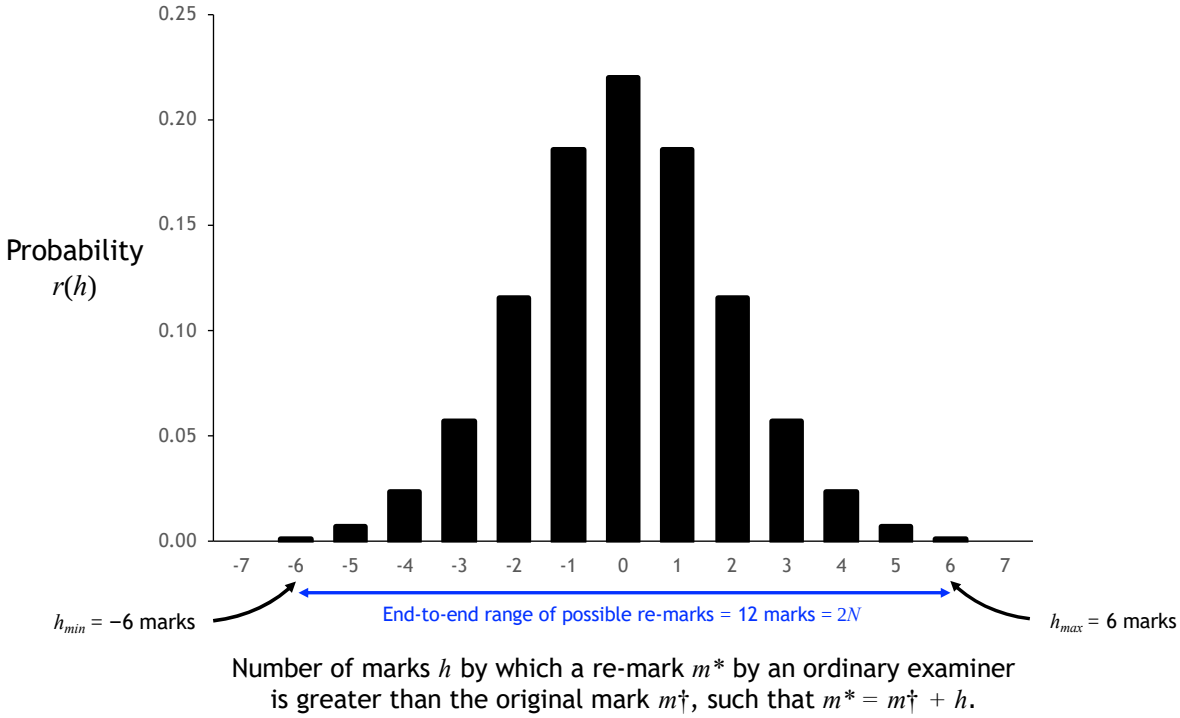


Table 4: The probability that a script originally marked $m\dagger = 64$ will be re-marked $m^* = m\dagger + h$, as shown in Figure 14.

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
$m\dagger = 64$	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 6, the distribution $r(h)$ shown in Figure 14, and the corresponding values in Table 4, 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 14 is a more informative representation of the same data as shown in Figure 13: Figure 13 shows separately the seven different individual panel distributions of which the original mark $m\dagger = 64$ is a member; in Figure 14, 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 14 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' – the fact that a candidate's grade is determined by the mark given by the examiner who happens to mark the script first.

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 9.

As already noted, however, any examination subject has a characteristic generic panel distribution $T(n)$. Furthermore, as discussed on pages 22 to 24, the distribution $T(n)$ can be used to determine the special re-mark distribution $Q(p)$ as

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

Also, as already noted and as will be proven on pages 37 to 39, 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 (as is indeed the case) the distribution $r(h)$ can be determined by statistical sampling, 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 $r(h) = Q(p) * T(p) = T(-p) * T(p)$ 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; 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 9, which shows a representative special re-mark distribution $Q(p)$, and Figure 14, 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. 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 6 and 9). 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, for example, here.
- 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 9 and 14), 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 13 – is

double that of the width of the corresponding special distribution (6 marks in the example shown in Figure 8); 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.

The double marking fallacy

A further feature of the ordinary re-mark distribution, as exemplified by Figure 14, 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 $m\dagger$, 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 $(m\dagger + m*)/2$ is a 'better' mark than either $m\dagger$ or $m*$. Is this true?

To explore this question, suppose that the generic panel distribution of the subject examination is as shown in Figure 6, 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 $\mathbf{M}\dagger = 66$, 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 $m\dagger = 64$. 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 $(m\dagger + m*)/2 = 63$. Which of the three marks 64 (the original), 62 (the re-mark) and 63 (the average) is right?

Figure 15: Double marking. A script is given an original mark $m\dagger = 64$. The 'definitive' mark for that script is $\mathbf{M}\dagger = 66$. How useful is a re-mark $m^* = 62$?

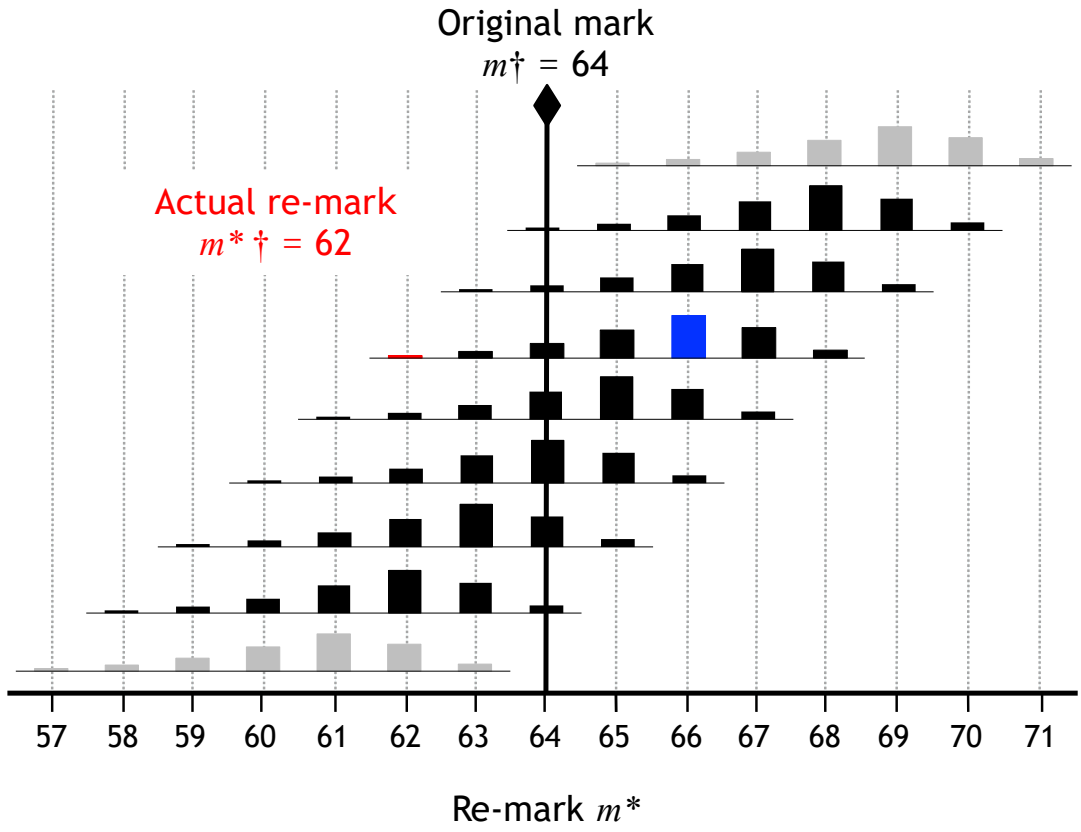


Figure 15, which contains much more information than is available to either of the two ordinary examiners, provides the context in which this example of double-marking is taking place. The original mark $m\dagger = 64$ is a member of any one of seven possible generic panel distributions, each with its 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 $\mathbf{M}_2 = 66 = \mathbf{M}\dagger$.

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 $(m\dagger + m^*)/2 = 63$ are even further from the 'definitive' mark than the original mark $m\dagger = 64$. In this instance, double-marking has made matters worse, not better.

Reference to Figure 14, and the data in Table 4, shows that there is a probability of about 12% that a script originally marked $m\dagger = 64$ will be re-marked two marks lower, $m^* = 62$, for which the parameter $h = -2$. Figure 14 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 $m\dagger = 64$. A re-mark $m^* = 64$ might be interpreted as confirmation that 64 is indeed the 'right' mark – but reference to Figure 15 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 $(m\dagger + m^*)/2 = 66$, which is equal to the 'definitive' mark $\mathbf{M}\dagger = 66$. The probability that this will happen, according to Figure 14, 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 $m\dagger = 64$ 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 14 shows that the probability that a script originally marked $m\dagger = 64$ 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 14 to determine this probability is wrong. Since it is known that both the original mark $m\dagger = 64$ 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 6 and Table 2: if a re-mark m^* is 2 marks lower than the original mark $m\dagger$, the probability is therefore about 11%. This number happens to be rather close to the probability of 12% as inferred from Figure 14, but this is a numerical co-incidence rather than an indication of a deeper truth; the fundamental question remains – which of Figures 6 and 14 is the correct one to use?

The argument in favour of using Figure 6, the generic panel distribution, is apparently compelling, for it as indeed true that both the original mark $m\dagger = 64$ and the re-mark $m^* = 62$ must indeed be members of the same generic panel distribution – and it is Figure 6, not Figure 14, 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 $m\dagger$ and the re-mark m^* must be members of the same generic panel distribution. But when the only information available is the original mark $m\dagger$, 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 $m\dagger = m^* = 64$, 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 10 to 13, the search for the 'right' mark is illusory (see also pages 81, 114 and 186 here).

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 8, and the ordinary re-mark distribution $r(h)$, as illustrated in Figure 11.

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 7, 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 6 is illustrated in Figure 8. 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, as discussed on pages 22 to 24, the left-right mirror image of $T(n)$ such that

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

As was shown by the comparison between Figures 7 and 13, 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 16: The variables p, n and h, showing that $h = p + n$

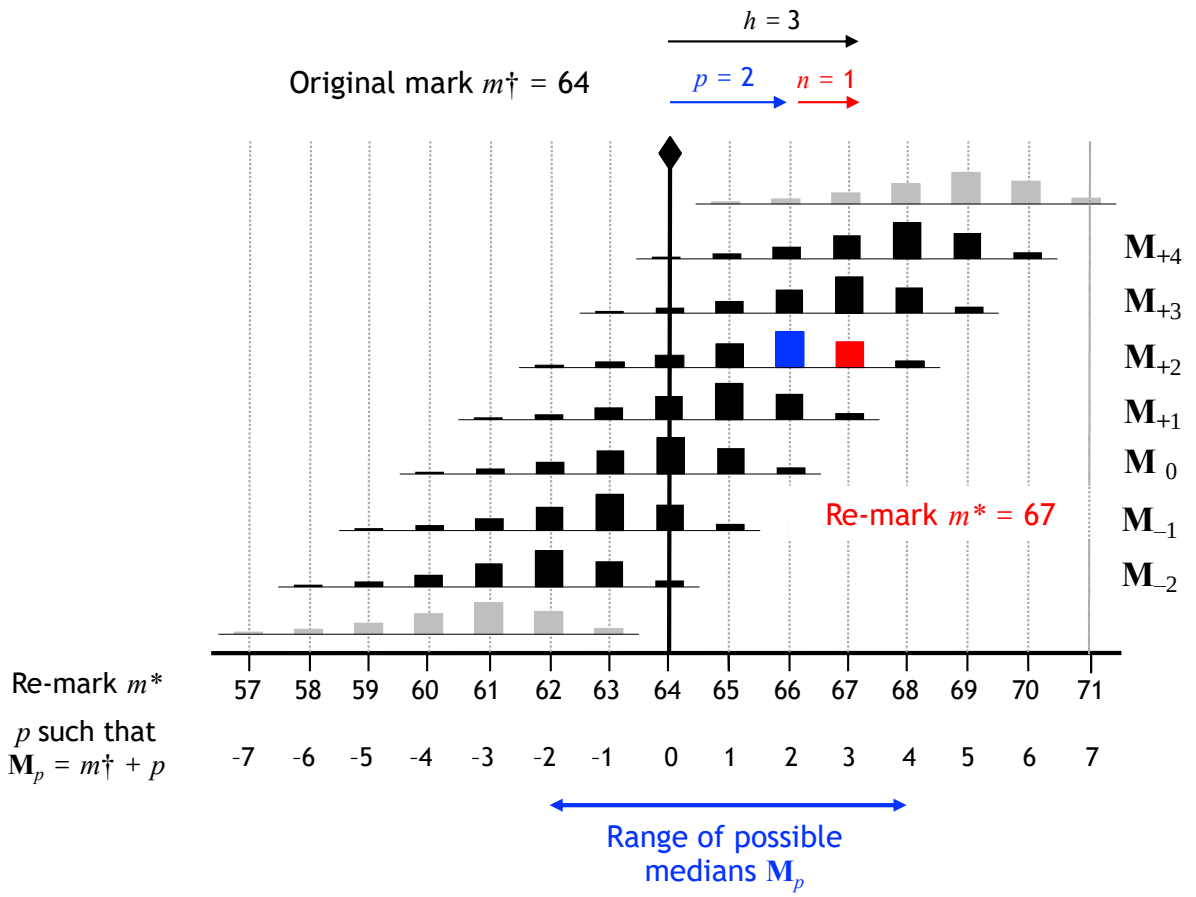


Figure 16 shows an example of a script originally marked $m\dagger = 64$, and subsequently re-marked $m^* = 67$. Since the parameter h is defined such that $m^* = m\dagger + h$, then $h = m^* - m\dagger$, which in this case implies that $h = m^* - m\dagger = 67 - 64 = 3$.

As illustrated in Figure 16, 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 $m = \mathbf{M}\dagger + n$, n marks greater than that distribution's median $\mathbf{M}\dagger$, 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 $m^* = m\dagger + h$, under the assumption that both the original mark $m\dagger$ and the re-mark m^* are members of the specific generic panel distribution of median $\mathbf{M}_p = m\dagger + p$.

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 $m^* = m\dagger + h$ 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 $\mathbf{M}_p = m\dagger + p$, 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 $Q(p) * T(p)$.

Since, as we saw on page 22,

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

then

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

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 7, and represents the number of generic panel distributions that include the original mark $m\dagger$, 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 6, 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 6 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 6), but 'spread' from left to right, as illustrated in Figure 17 (in which, for clarity, each individual distribution is shown by a continuous line rather than a sequence of columns).

Figure 17: The summation $\sum_{p} T(h-p)$

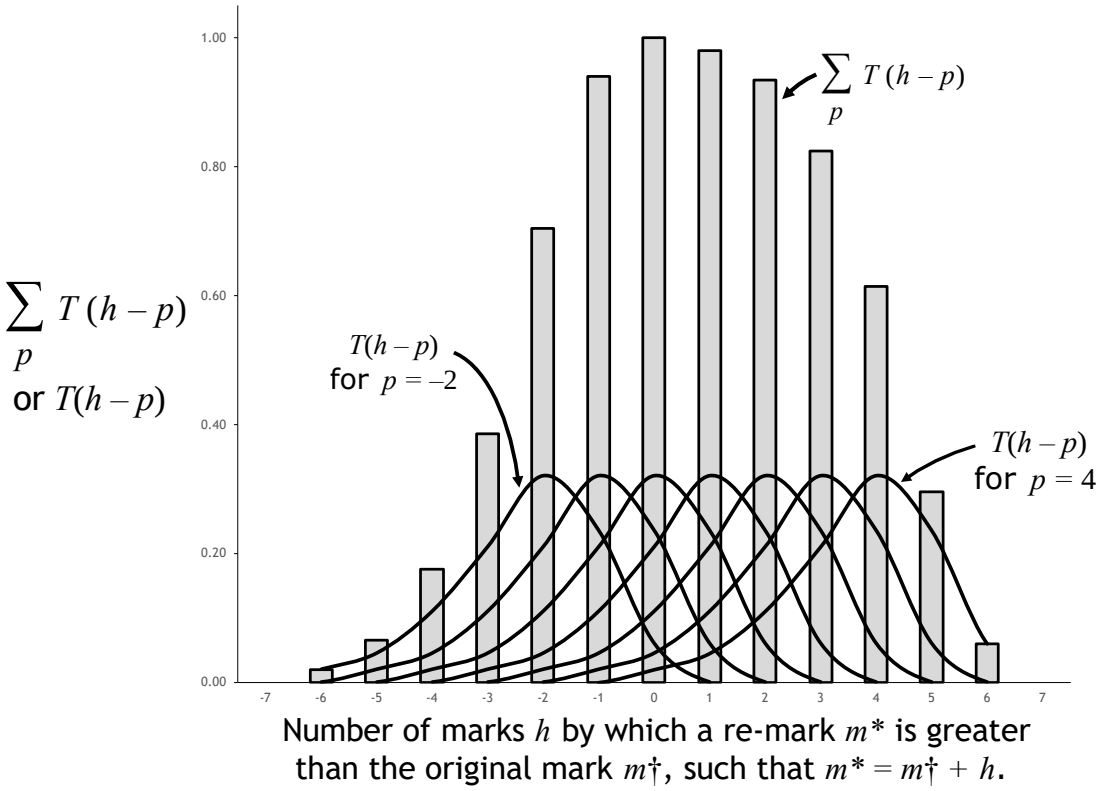


Figure 17 shows $N + 1 = 7$ generic panel distributions $T(h-p)$ of the general shape of $T(n)$ as shown in Figure 6, corresponding to each of the seven distributions shown in Figure 7, and then aggregated. When the values of each of these for any value of h are added, the result is as depicted by the histogram; the corresponding numeric values are shown an Table 5.

Table 5: Values of $T(h-p)$ *– the data corresponding to Figure 17, 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 5, the row identified as $T(h)$ shows values of $T(h)$, which are identical to those of $T(n)$ as given in Table 2, 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 7, and also shown in Figures 13, 15 and 16. 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 17; and the grand total in the bottom right-hand corner is 7.000.

In Figure 17, 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 $r(h)$

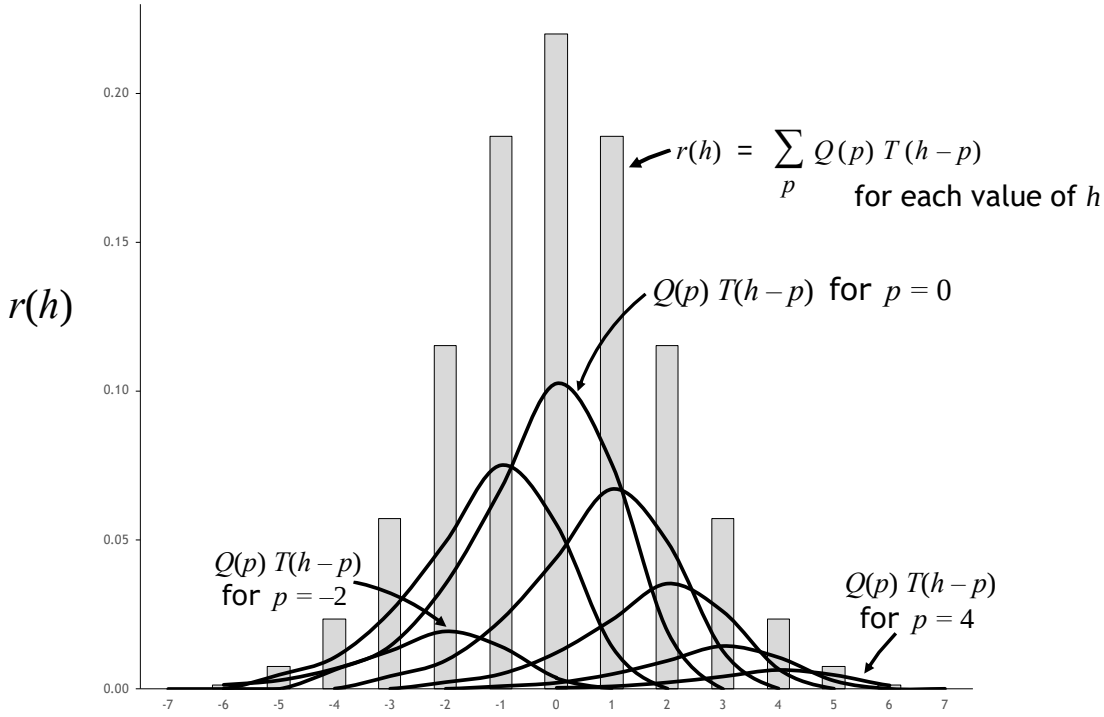
$$r(h) = \sum_p Q(p)\, T(h - p) = Q(p) * T(p)$$

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 18:

Figure 18: The ordinary re-mark distribution $r(h) = Q(p) * T(p) = T(-p) * T(p)$



Number of marks h by which a re-mark m^* is greater than the original mark $m\dagger$, such that $m^* = m\dagger + h$.

In Figure 18, 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 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 6, is asymmetrical.

Table 6: Values of $Q(p)\ T(h-p) = T(-p)\ T(h-p)$ – the data corresponding to Figure 18, 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 6 shows the data corresponding to Figure 18. 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 5. 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 6 corresponds to the 'row' in Figures 7, 13, 15 and 16 for the same value of p, and the row totals in Table 6 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 17, 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 6 therefore show the values of the special re-mark distribution $Q(p) = T(-p)$ for each value of p, as verified by Table 3.

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

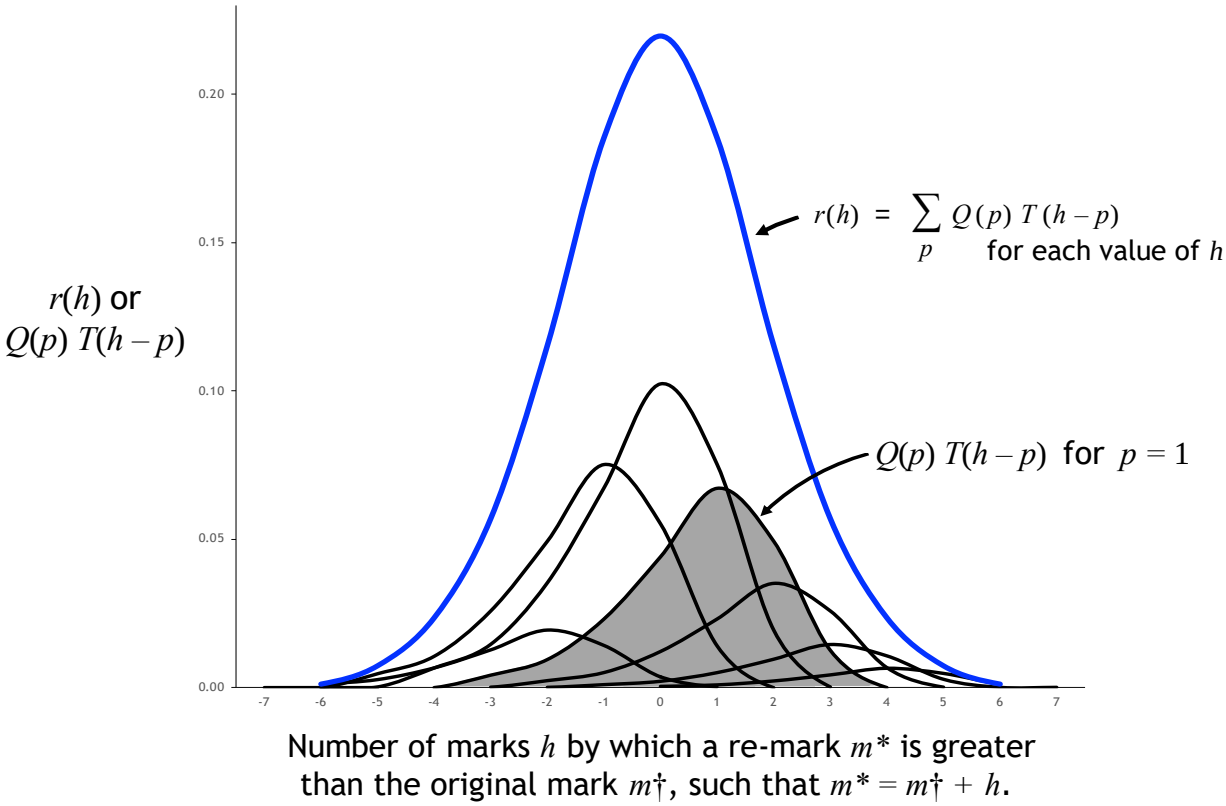
$$\sum_{p} Q(p)\ T(h-p) \ = \ \sum_{p} T(-p)\ T(h-p) \ = \ r(h)$$

This is the convolution $Q(p) \ast T(p) = T(-p) \ast T(p)$, and so the column totals give the numerical values of the ordinary re-mark distribution $r(h)$, as shown by the histogram in Figure 18.

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 19 corresponding to $p = 1$:

Figure 19: The probability that a script marked $m\dagger$ 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 $m\dagger$ 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 17 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 19.

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. 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 5. 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 13.

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.
$\boldsymbol{m^*}$	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 $p_{\min} = -n_{max}$ 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.
$\langle \mathrm{M} \rangle$	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.