Which is better: one experienced marker or many inexperienced markers?

Tom Benton Research Division

Introduction

At present, the ultimate arbiter of the mark that should be awarded to any exam script is the principal examiner (PE). The reasons for identifying a single individual as the ultimate judge are mainly practical. For example, it means that when marking is reviewed, there is a single person with the authority to determine the correct mark in difficult cases where there is a disagreement over the marks that should be awarded. Furthermore, if the PE marks a number of scripts at the beginning of the marking process, then the extent to which other examiners are able to independently replicate these marks on the same scripts provides a good indicator of their suitability to begin marking. Marks from the PE are used in this way during both practice and standardisation of examiners at a time when no other marking has been completed. A similar process can also be used to monitor the quality of marking of different examiners right from the start of marking and throughout.

However, beyond these practical considerations, this approach also indicates a set hierarchy of markers. It implies a predetermined axiom that the best way to mark a given assessment is defined by the most senior marker (the PE). Therefore, the whole machinery of marking should be about communicating the PE's approach to marking to more junior markers, and them being evaluated with the regard to the extent to which their marks are in line with those that would be awarded by the PE.

At the top of the hierarchy is the PE. At the next layer of the hierarchy are team leaders – fairly experienced examiners who are responsible for supporting and monitoring teams of more junior markers. At the bottom of the hierarchy of markers are assistant examiners (AEs) who comprise the vast majority of the individuals involved in marking. Whilst junior, AEs are professionals who are appropriately trained and standardised before beginning marking. However, where marks awarded by AEs differ from those awarded by the PE, the default assumption is that it is the AEs that are wrong. Even where numerous AEs have a shared professional opinion of the mark that should be awarded a given script, the PE's mark is still assumed to be the correct one.

This article explores the evidence for the assumed supremacy of the PE over groups of more junior markers. Understanding the truth of this assumption has important implications both for the way in which markers are monitored, and for the way that quality of marking is reported at a national level.

Quality of marking, both for individual markers and at a national level, is usually evaluated via a marker monitoring process. This process typically collects data about each marker as follows. For each examination, before the main body of marking begins, a small number of scripts are pre-marked by senior examiners (usually including the PE), either in consultation with one another or acting alone. These scripts are known as "seed scripts" or "seeds", and the marks that the PE has agreed should be awarded to them are called the "definitive marks". During live marking, these seeds are placed at various random intervals in each marker's script allocation (Ofqual, 2016). Each marker marks these scripts "blind" (i.e., unaware that they have already been marked before and without knowledge of the pre-determined definitive mark). The marks awarded to seed scripts by individual markers are compared to the definitive marks and are continuously analysed for any signs that examiners' marking may be becoming inaccurate. After the end of an examination series, this same data is also used by The Office of Qualifications and Examinations Regulation (Ofqual) to produce reports on the quality of marking nationally (e.g., Ofqual, 2016). The implicit assumption both in the monitoring process, and in the post-series analysis, is that marks awarded by the PE (possibly in consultation with other senior examiners) can be treated as being absolutely correct (Suto, Nádas, & Bell, 2011) and that, as such, differences from these represent errors that must be eradicated.

With this backdrop in mind, this article aims to investigate the accuracy of definitive marks using empirical data. Specifically, it addresses the question of whether marks awarded by a PE are genuinely more useful than those derived by combining the marks of several, less experienced, AEs. If the PE's marks are truly "better" then they should be at least as predictive of wider achievement as the marks pooled from the AEs. This article will test whether this is in fact the case.

As noted above, definitive marks may, in some cases, be assigned by a team of senior examiners rather than by a PE working alone (Black, Suto, & Bramley, 2011). However, in order to simplify the language of the remainder of this article, we shall treat them as if they are generated by a single individual. For some definitive marks this is literally true, but no formal record is made of which senior examiners were involved in assigning particular definitive marks so the extent of this is not known. However, even if such marks involve some collaboration between senior examiners it is hard to believe that this would result in lower accuracy than would be achieved by a PE working alone. With that minor caveat in mind, we will proceed to the analysis.

Data and Method

This analysis makes use of seed scripts from all of OCR's GCSE, AS and A Level papers from summer examination sessions between June 2015 and June 2018. For the purposes of this research, seed scripts are valuable as they provide a small number of scripts for every examination that have both been marked by the PE, and have been marked by several AEs.

Analysis is restricted to all seed scripts that had been marked by at least five AEs. It is also restricted to those examination components (papers) where at least ten such seed scripts were available. This left a total of 724 papers for analysis across the four examination sessions. Further details regarding the data set used in analysis are given in Table 1.

Table 1: Details of data used in analysis

Number of seed scripts	10,786
Number of examination components	724
Number of AEs	20,827
Median number of seed scripts per component	15
Median number of markers per component	21

To begin with, the mark that should be awarded to each seed script was derived in three different ways.

1. **The definitive mark.** As mentioned earlier, such marks are usually decided upon by the PE themselves, although other senior examiners may also be involved in the discussion.
2. **The mean mark awarded across all AEs that marked the script.** Note that if an AE does a large amount of marking then they may mark all of the available seed scripts once, and then (having completed this), any further seed scripts they are assigned will be ones they have marked before. It is also possible that during marking, they may be given feedback based on the marks they award seed scripts. For this reason, to keep this measure as free from the influence of senior markers as possible, only the first attempt each AE made at marking each seed script was included in this measure. Note that, for the purposes of this research study, markers were included in the generation of this mean mark even if they were later deemed to be inaccurate and stopped from live marking[1]. This decision was taken to ensure that this research gave a pure idea of the accuracy of multiple junior markers without the influence of more senior markers in choosing who should be included. The mean was used as it is the simplest possible method for combining the marks of various AEs into a single score. Also, given that for any individual whole script, the marks assigned by different examiners tend to follow a roughly normal distribution (see Ofqual, 2016, Figure 11), the mean might reasonably be assumed to be the most efficient estimate.
3. **The median mark awarded across all AEs that marked the script.** As above, but using the median mark across all markers rather than the mean. Possible advantages of the median include resistance to outliers and the fact that it is more likely to result in a mark that is a whole number.

For the purposes of brevity within this article, the second and third methods of assigning marks to each script will be referred to as *collectively assigned* marks.

In order to evaluate the predictive power of different ways of generating marks for these seed scripts, I used the *external* ISAWG (Benton, 2017; Benton, 2018) for the candidates providing the seed scripts. The ISAWG is a standardised measure of each candidate's achievement that summarises their performance across all of the assessments they have completed within a given examination session. For further details on how it is calculated, see Benton (2017). The external ISAWG is based on the same calculation but only using assessments other than the one being studied. It can be interpreted as a very general measure of ability across different subjects. It was used in this analysis, as it was easily available for nearly all the candidates from whom seed scripts were selected.

The predictive value (or concurrent validity) of the three different ways of generating marks for each seed script was evaluated by the Pearson correlation of the marks with the external ISAWG. That is, the analysis identified which of the three different ways of generating the marks for each script was most predictive of candidates' wider achievement.

Note that, for each individual examination being studied, only 10 to 20 seed scripts were typically available for analysis. As such, the individual correlations, based on such a small number of candidates, were almost meaningless when considered individually. However, by analysing them as a group across the hundreds of examination papers included in the analysis, we hope to identify some clear trends with regard to whether definitive marks from a PE or pooled marks across many AEs have more predictive power.

Results

Comparison of predictive power of definitive marks versus consensus marks

Table 2 provides a summary of the results across the 724 components included in the analysis. The central result from this table is that collectively assigned marks for seed scripts (whether calculated using the mean or the median) were slightly more predictive on average of external achievement than the definitive marks.

The fact that the difference in predictive power between collectively assigned and definitive marks is small (just below 0.02) is not unexpected. Definitive marks are from the most senior marker and should be very accurate. As such, there is likely to be relatively little room for improvement. Nonetheless, the results show that pooled marks from several AEs have more predictive power than the marks from PEs. Whilst we can invent fanciful stories about how these results might have occurred (e.g., "assistant examiners are more influenced by signs of general ability whereas PEs recognise abilities specific to the assessment being studied"), the most straightforward interpretation of these results is that collectively assigned marks from multiple examiners are simply a more accurate indicator of a candidate's performance. That is, the many junior examiners outperform the single PE.

Table 2 splits the summarised results into components within Science or Mathematics, those within Modern Foreign Languages (MFL) and all of the rest including Humanities, Expressive Subjects and Classics. For all subject groups, the collectively assigned marks had more predictive power than the definitive marks. However, the gain in predictive power for Science, Mathematics and MFL (less than 0.01) was less than the gain for other subjects (close to 0.03). This result mirrors the general finding from most previous research that marking is most reliable in "exams containing structured, analytically marked questions" (such as are likely to occur in Science, Mathematics and MFL) and generally less reliable in "exams containing essays" (such as Humanities); see Bramley (2008). Similarly, the subjects where collectively assigned marks have the greatest superiority in predictive

1. Where markers are stopped from marking, all of the live scripts they have already marked are then allocated to a new marker to be remarked. However, for the purposes of this research, the marks awarded by such markers are available within our systems.

Table 2: Summary of correlations of different ways of assigning marks to seed scripts with external achievement

Subject	*Number of components*	*Mean correlation with…*			*Median correlation with…*		
		Definitive mark	*Mean mark*	*Median mark*	*Definitive mark*	*Mean mark*	*Median mark*
Science and Mathematics	302	0.788	0.793	0.793	0.812	0.823	0.822
MFL	33	0.715	0.725	0.722	0.747	0.753	0.755
Other	389	0.744	0.771	0.767	0.790	0.815	0.808
Total	**724**	**0.761**	**0.778**	**0.776**	**0.796**	**0.814**	**0.811**

power are also those where differences between markers are less likely to be "procedural" or "attentional errors" and more likely to be due to "inferential" or "definitional uncertainty" (Ofqual, 2018b).

It is worth noting that, although collectively assigned marks were allowed to be non-integer numbers (i.e., decimals), rounding the marks before calculating correlations had virtually no impact on the results in Table 2. The largest change from rerunning the analysis with rounded marks occurred for the mean of the correlations of mean marks with external achievement in MFL and was still only a change of 0.004 (a drop from 0.725 to 0.721). The median change of those figures in Table 2 that changed at all was just 0.0005. In other words, it is the fact that marks are derived from a collection of individuals that drives the good performance of collectively assigned marks and not simply the fact that they can include non-integer numbers.

As can be seen from Table 2, there was very little difference in predictive power regardless of whether collectively assigned marks were calculated using the median or the mean. For this reason, the remainder of this article will focus upon the use of the mean to generate such marks.

A visual representation of the difference in the predictive power of definitive and collectively assigned marks is given in Figure 1. For each of the 724 examination papers in the analysis, the chart plots the correlation of external achievement with the definitive marks against the correlation with the mean mark. Although the individual points in this chart are based on very small samples, and thus, nearly meaningless, it is the overall pattern that is of interest. To help discern this overall pattern, the plot includes a dotted line representing a line of equality and a solid blue line which represents a regression

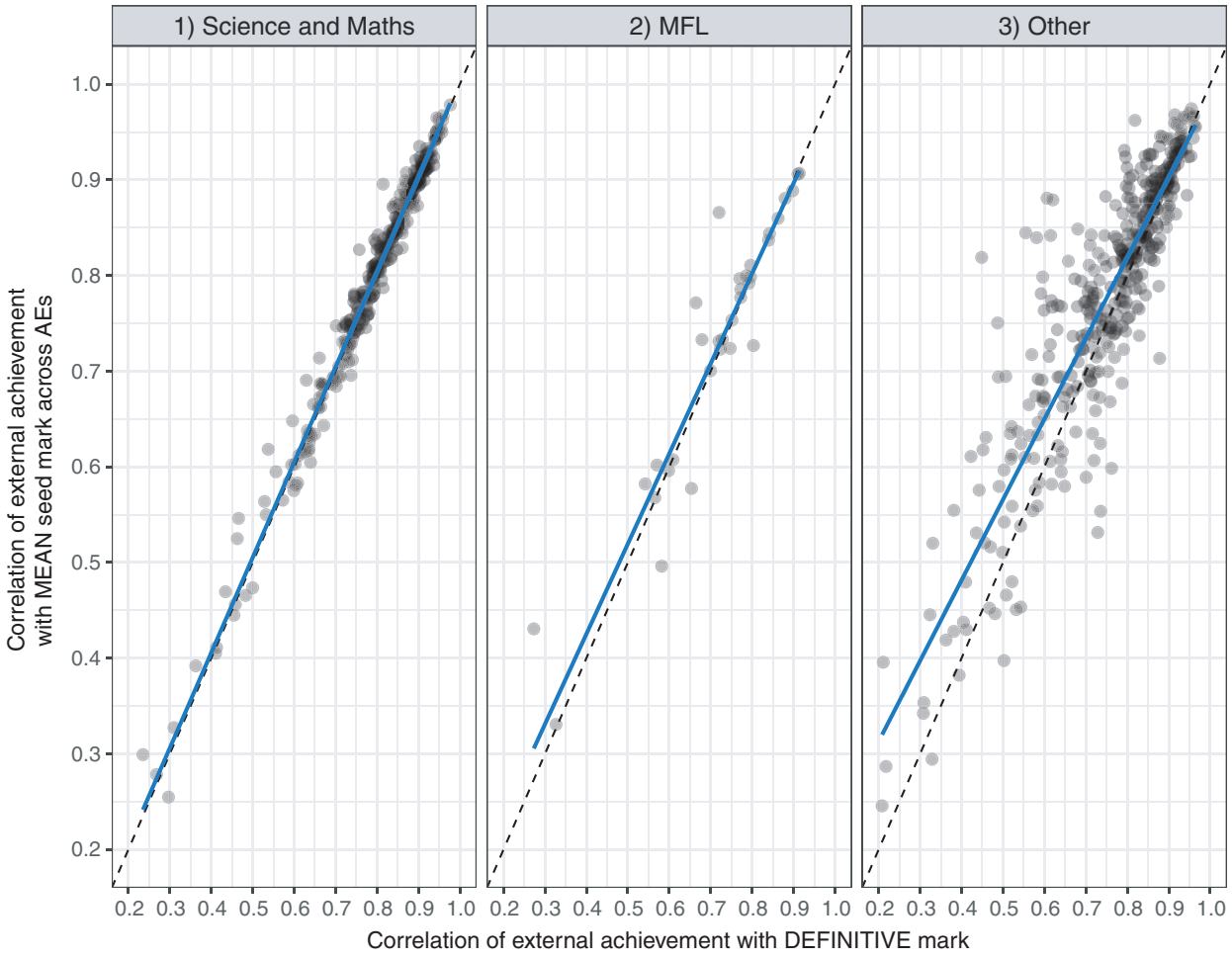

Figure 1: Comparing correlations of external achievement with mean mark awarded by AEs and with definitive marks awarded by PEs.

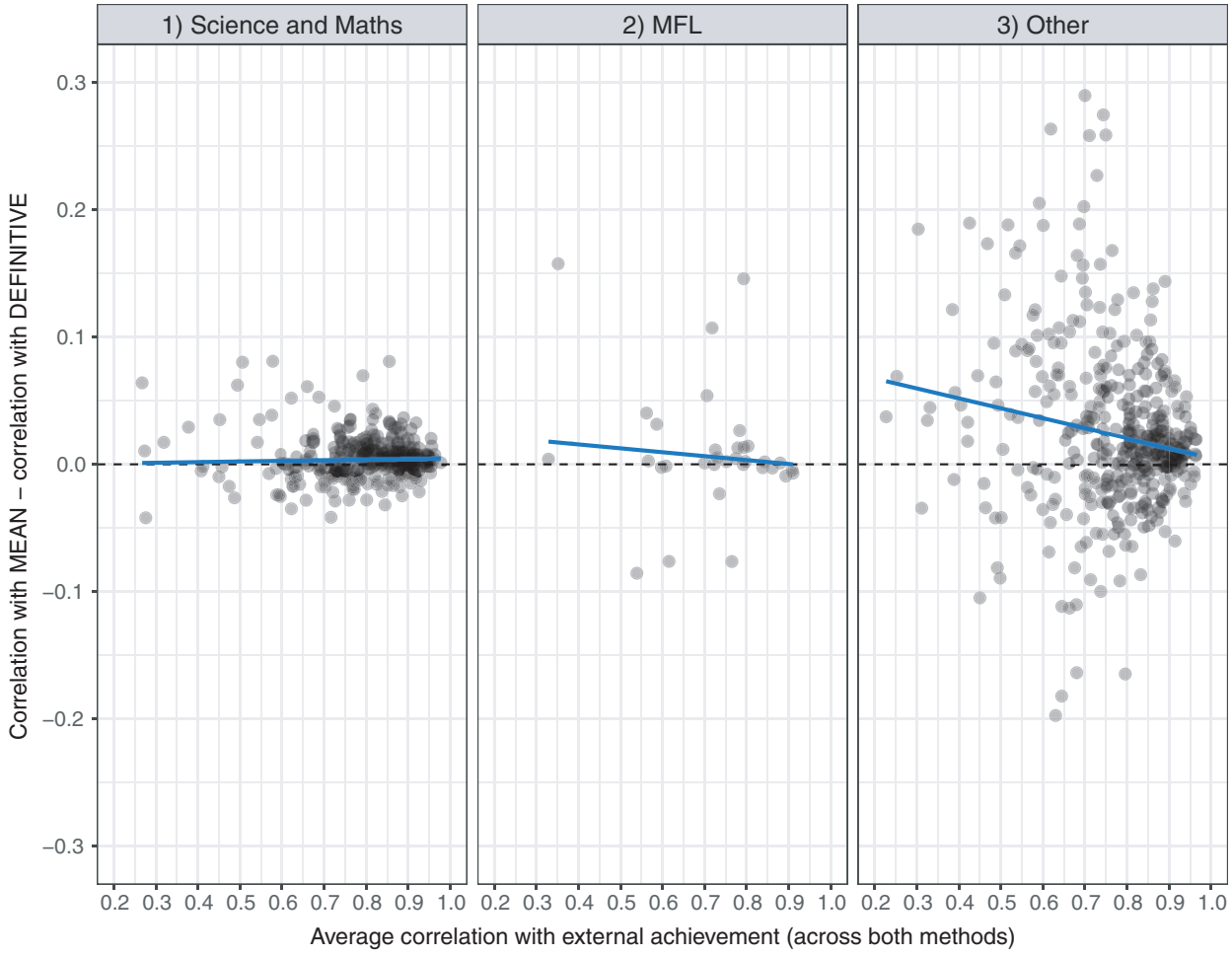

Figure 2: Mean-difference plot showing how the difference between the predictive power of collective and definitive marks relates to the average predictive power of both methods.

line[2]. As can be seen, for Science, Mathematics and MFL, there is barely any gain from using collective marks rather than definitive marks. However, for other subjects the blue regression line is clearly above the dotted line of equality. In other words, there are obvious gains from using collective marks rather than definitive marks.

Another way to view this same data is given in Figure 2. Using a Tukey mean-difference plot[3] allows a more detailed view of the differences between methods. Of particular interest is the right hand panel. This shows that the gain from using collective marks is largest when the predictive power of both methods is relatively low. Remember that each point in the chart is based on analysis of a very small number of scripts; as such, if the selected candidates happen to perform very differently in other examinations to the examination being studied, these correlations can be quite low. On the other hand, if the selected candidates happen to perform similarly well in other examinations then these correlations may be high and the potential for improvement is much lower. Nonetheless, regardless of the size of the gain, the central point from this analysis remains – the average mark from many junior examiners is superior to the mark from a PE.

2. Due to the very small sample sizes used to generate each correlation in the plots, and the associated increased risk of outliers, all regression lines in this article were created using robust regression. Regression lines were created using the function *rlm* from the R package *MASS* (Venables & Ripley, 2002).

3. Popularly known in biomedicine as a Bland-Altman plot.

The impact of the structure of exam papers

So far, we have seen that collectively assigned marks have more predictive power than definitive marks, and that this appears especially true in subjects that are likely to require more professional judgement in marking. However, this last point has only been crudely demonstrated by splitting assessments of Science, Mathematics and MFL from the remainder. This next section attempts to find a more universal measure of the likely level of subjectivity required to mark each exam paper.

The principle used to derive this measure is based on data on marking consistency across all awarding organisations combined reported in Ofqual (2016). Figures 4 and 5 of Ofqual's report (pages 10 and 11) show that (across many items) the scale of inconsistency between markers tends to increase in a way that is roughly proportional to the number of marks that are available on an item. That is, whilst it is accepted that there are many additional features of items and mark schemes that may affect marking consistency (Bramley, 2008; Black, Suto, & Bramley, 2011), to a large extent, the scale of marking consistency for markers can be explained by the number of available marks on the items they are marking.

More specifically, Ofqual's research shows that the standard deviation of marking differences (between individual markers and definitive marks) is roughly proportional to the number of marks available of the item. That is,

standard deviation of marking differences on item $i \approx k * Max_i$

where k is a fixed constant[4] and Max_i is the maximum number of marks available on item i. This formula gives us a straightforward way of estimating the likely scale of marker inconsistency at item level. To estimate the scale of marker inconsistency at the level of whole scripts we use standard statistical rules. These tell us that, if marker differences on different items are independent (and we would hope that they are), then the standard deviation of marker differences at whole script level should be the square root of the sum of the squared standard deviations of differences across each item. That is, using equations similar to some given on page 15 of Ofqual's report,

Standard deviation of whole script marking differences

$$= \sqrt{(SD\ of\ diffs\ on\ Q1)^2 + (SD\ of\ diffs\ on\ Q2)^2 + (SD\ of\ diffs\ on\ Q3)^2 + \cdots}$$

$$\approx \sqrt{(k * Max_1)^2 + (k * Max_2)^2 + (k * Max_3)^2 + \cdots}$$

$$= \boldsymbol{k * \sqrt{Sum\ of\ squared\ item\ maxima}}$$

The final term is the square root of the sum of the squared item maxima. From now on, we will refer to this as the RSSIM. Although, to some readers, the equations above may appear complex, the calculation of the RSSIM is relatively straightforward and provides a single metric which is proportional to the likely extent of disagreement between markers. For example, suppose an exam requires candidates to answer three questions in an exam with 10, 20 and 30 marks available respectively. In this case the RSSIM= $\sqrt{100 + 400 + 900} = \sqrt{1400} = 37.4$. On the other hand, if another 60 mark assessment consists of 60 items each with one mark available then RSSIM= $\sqrt{1 + 1 + 1 + \cdots} = \sqrt{60} = 7.7$. Thus, we would expect the amount of variation between markers to be roughly five times higher in the former 60 mark test than in the latter one.

Returning to our main research question, we can investigate how the gain in predictive power from using collectively assigned rather than definitive marks relates to the RSSIM. This analysis is provided in Figure 3. As expected, the greatest gain from using collectively assigned marks is found for those exams where marking consistency is hardest to achieve – that is, those with the largest RSSIM values. It can also be seen that the RSSIM also serves to explain the differences between the subject groupings used earlier. This is important as it means that the RSSIM can be used as a shortcut to classify exam papers in terms of the likely difficulty of marking without the need to manually decide upon how to group different subjects.

How many ordinary markers is the one good marker worth?

Having seen that collectively assigned marks have greater predictive power than definitive marks, the aim of this section is to determine how many AEs are needed in order for this to be the case. In order to answer this question the following procedure was used for each assessment within the study:

1. For each seed script in turn, randomly select *n* AEs to contribute to the collectively assigned mark.
2. Calculate the collectively assigned (mean) mark of each seed script based on the selected markers only.
3. Calculate the correlation of the (new) collectively assigned marks to all of the seed scripts with external achievement.
4. Repeat steps 1 to 3 500 times and record the average correlation across all of the replications.

4. Using Figures 4 and 5 of Ofqual (2016) we can see that k is approximately equal to 0.12.

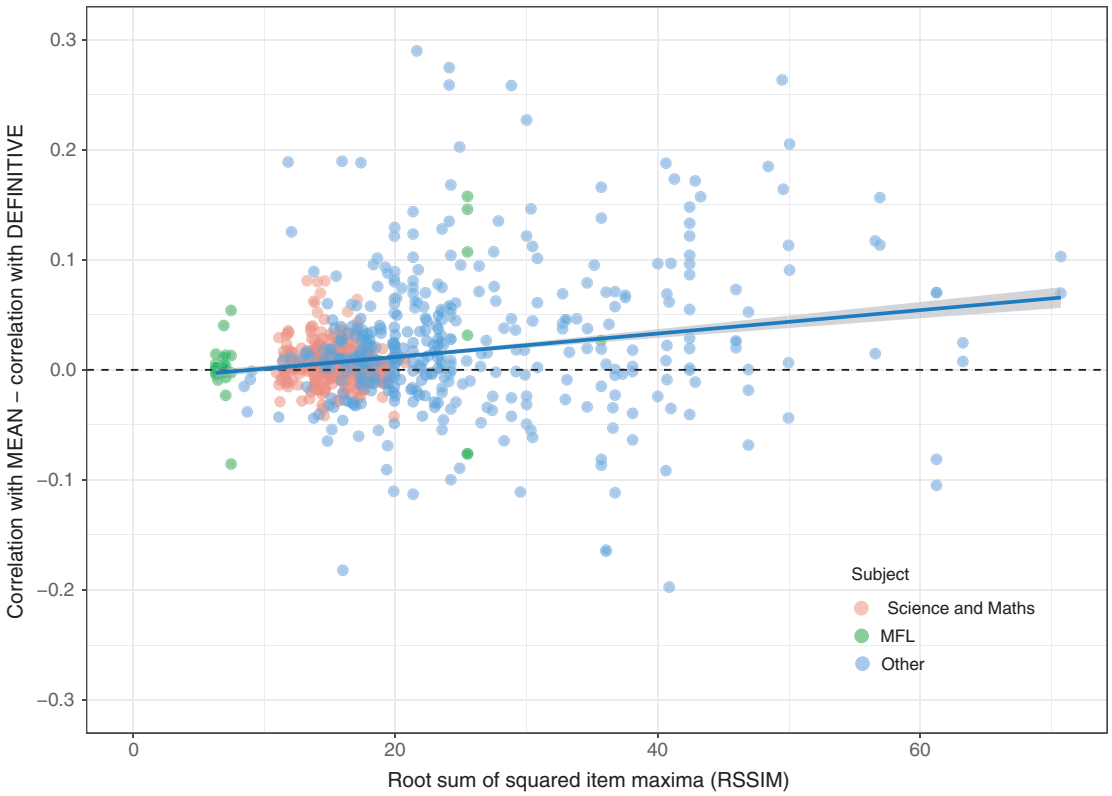

Figure 3: Relationship between the root sum of squared item maxima (RSSIM) and the gain in predictive power from using collectively assigned rather than definitive marks.

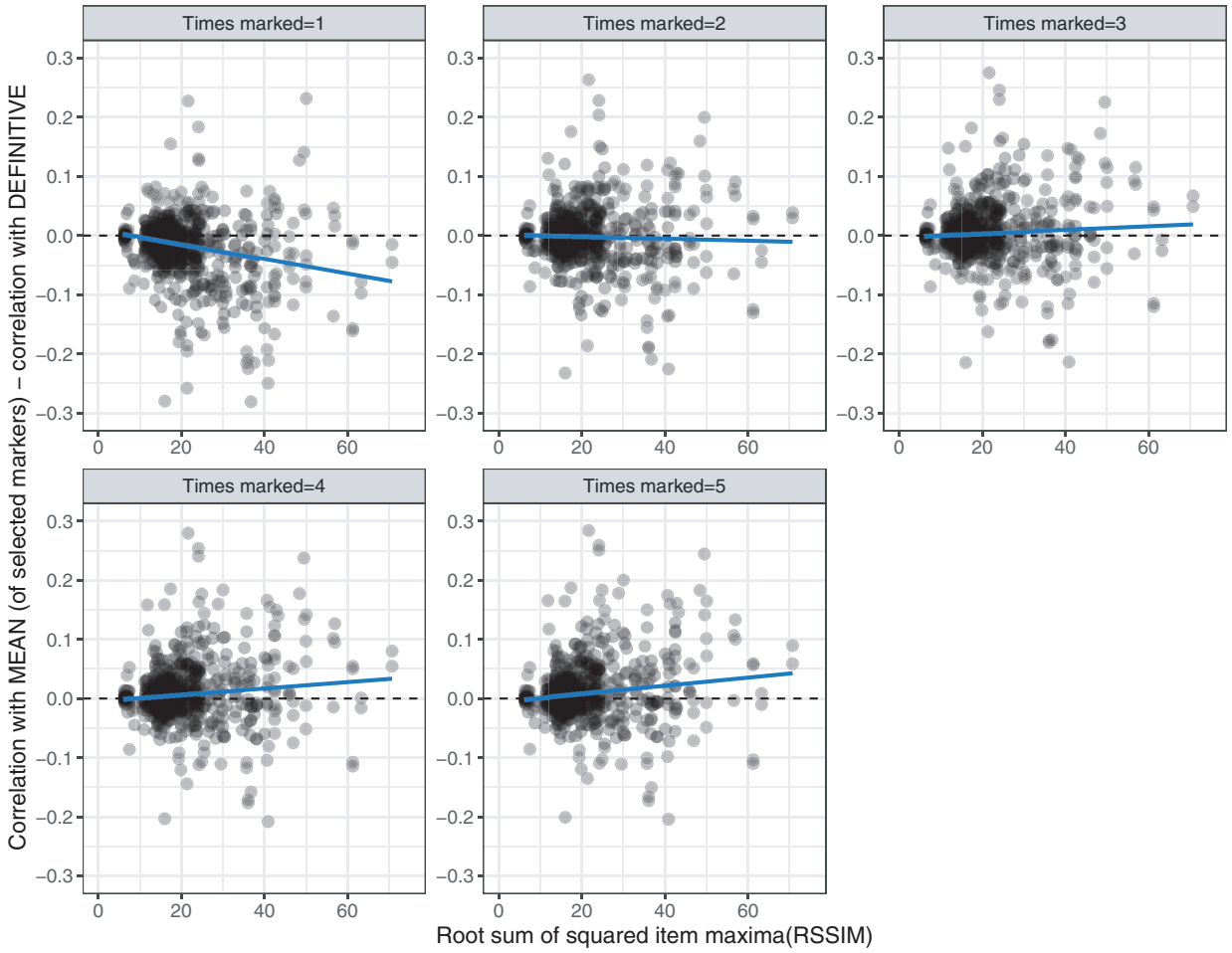

Figure 4: Relationship between the root sum of squared item maxima (RSSIM) and the gain in predictive power from using collectively assigned marks from between 1 and 5 AEs rather than definitive marks.

The above procedure was used to investigate the predictive power of collectively assigned marks from 1, 2, 3, 4 and 5 AEs (i.e., $n = 1, 2, 3, 4$ and 5).

The results of this analysis are shown in Figure 4. This chart shows that if only one AE marks each script (not really a collectively assigned mark at all) then the definitive marks tends to have noticeably more predictive power. If the mark awarded to each seed script is derived from two AEs (double marking) then definitive marks were still superior (but only just). However, marks collectively assigned by three or more AEs tended to have more predictive power than the definitive marks. Thus, it appears that a PE is worth between two and three AEs. This result is broadly consistent with results reported for marking of an English Language essay in Benton and Gallacher (2018).

How accurate are definitive marks?

For the final section of this article, we change direction slightly. We have seen in the previous section that definitive marks (awarded with the involvement of the PE) have greater predictive power than those awarded by individual AEs. In other words, PEs are very good markers. However, the preceding sections showed that the collectively assigned marks from multiple AEs have even greater predictive power – that is, they are better still. Taking these two facts together means that the data analysed in this article provides an opportunity to explore the accuracy of definitive marks themselves.

To gain some idea of the accuracy of definitive marks we look at the scale of differences between definitive marks and collectively assigned marks. Before doing this, it is important to put the two sets of marks on a common scale. Collectively assigned marks will tend to have a lower standard deviation than definitive marks. Broadly speaking, this is because it is harder for a candidate to convince two examiners that their essay is a work of genius than it is to convince one alone. Similarly, and on a more positive note, you would be unlucky to find two examiners in succession who both think your essay is entirely without merit. To account for this fact, before doing this analysis, the collectively assigned marks in each assessment were rescaled to have the same mean and standard deviation as the definitive marks.

This change in scale did not matter for the earlier analysis of correlations because correlations are scale free measures of association. However, in this analysis we are investigating the actual differences between methods in terms of marks. As such, accounting for changes in scale is important.

The results of this analysis are shown in Figure 5. Each point in this chart represents an assessment. The RSSIM of the assessment is plotted against the average across all seeds of the absolute difference between (rescaled[5]) collectively assigned marks and definitive marks. Once again, it is worth reiterating that each individual point is based on a very small number of scripts and, as such is not too meaningful in itself. Furthermore, it is worth remembering that all available AEs have been retained in this analysis including those that were ultimately stopped

5. Skipping the step of rescaling does not make an enormous difference to the results. The scale of differences between definitive and collectively assigned marks just becomes slightly larger.

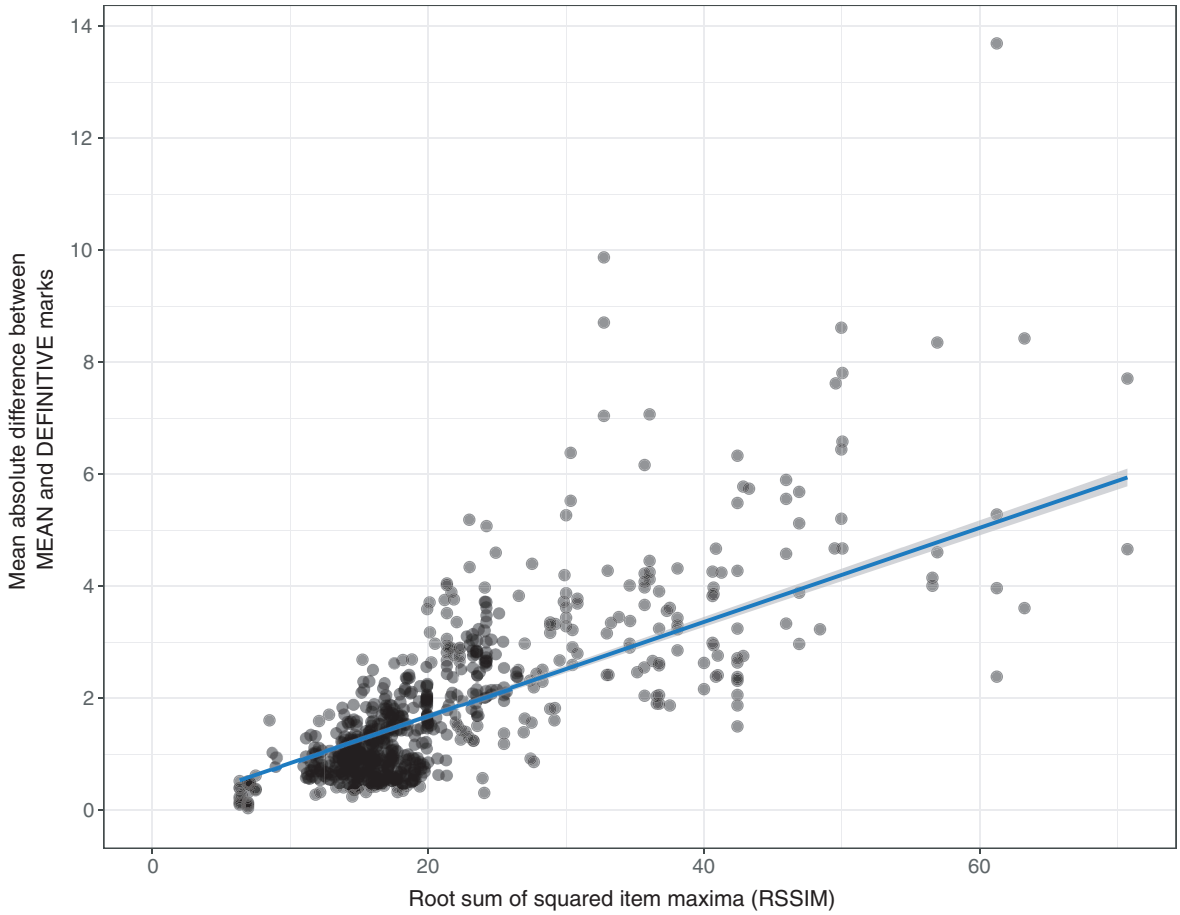

Figure 5: Relationship between the root sum of squared item maxima (RSSIM) and the mean absolute difference between collectively assigned and definitive marks.

from marking. For this reason, occasional markers who displayed very large discrepancies from definitive marks are included within the analysis leading to the possibility of outliers.

However, the main purpose of Figure 5 is to look at the overall trend. This is captured by the solid (robust) regression line. For the purposes of this chart only, this regression line was defined so that if extended it would pass through the origin. This decision was partly driven by common sense (if there are zero marks available on an assessment then it is impossible for any disagreement between markers to emerge), but was also made to allow us to derive a very simple rule for the likely accuracy of definitive marks. Specifically, the regression line shows that we expect the mean absolute difference between definitive and collective marks to be roughly one-twelfth of the RSSIM.

This result provides a lower bound for the level of consistency we can expect between markers. On the one hand, for each examination, we have taken our most experienced marker: the best of the best PE. On the other hand, we have a set of marks that the evidence suggests is even better than that – the collectively assigned marks. The analysis shows that even for these two most accurate sources of marks the difference between them tends to be about one-twelfth of the RSSIM. Thus, we cannot reasonably expect that the average difference between any two independent examiners, whether the PE or anyone else, would be less than this. For example, this implies that for an assessment with an RSSIM of 40 (as, for example, would occur for an exam out of 80 consisting of four 20 mark items), we cannot reasonably expect the average difference between the marks awarded to whole scripts by independent examiners to be less than 3.3 marks. It is also worth noting that this is an average difference. Broadly speaking, half of the differences between markers would be greater than this and half would be less.

Although the RSSIM is relatively easy to calculate, it is not particularly intuitive. To simplify matters Table 3 shows how, according to this analysis, the expected best possible absolute marking difference between examiners would vary, as a percentage of the paper maximum, according to the number of items in the paper if all items were worth the same number of marks[6]. This shows that for examination papers where candidates are asked to answer small numbers of items, the lower bound for the average difference between examiners is likely to exceed 5 per cent of the paper total. On the other hand, for papers where candidates are asked to respond to large numbers of items, the average difference may be less than 2 per cent of the paper total.

Also shown is an estimate of a lower bound for the standard deviation of marking differences. This metric is included to allow comparisons with results reported in Ofqual (2018a). By assuming that marking differences follow a roughly normal distribution, this is calculated by multiplying the mean absolute marking differences in the previous column by 1.25[7] The values in this column are relatively consistent with findings recently reported across all awarding organisations in Ofqual (2018a). For example, Figure 4 on page 14 of Ofqual (2018a) shows that the standard

6. This can be calculated as $100/(12 * \sqrt{Number\ of\ items})$.

7. According to the properties of the half-normal distribution the expected absolute value of a normally distributed variable is the standard deviation of that normal distribution divided by $\sqrt{\pi/2}$ – that is, roughly 1.25. See https://en.wikipedia.org/wiki/Half-normal_distribution for more details.

deviation (SD) of marking differences for Physics components tends to be at around 2 per cent of the paper total. Given that Physics papers typically include roughly 30 items[8], this value is very close to the relevant value reported in the final column of Table 3. Thus, for Physics at least, marking accuracy is already very close to the best that can reasonably be expected of markers.

Table 3: Expected differences between collectively assigned and definitive marks as a percentage of the total mark available for an exam assuming that all items are worth the same number of marks.

Number of items	*Mean absolute marking difference as percentage of paper total*	*SD of marking differences as percentage of paper total*
1	8.3%	10.4%
2	5.9%	7.4%
3	4.8%	6.0%
4	4.2%	5.2%
5	3.7%	4.7%
10	2.6%	3.3%
20	1.9%	2.3%
30	1.5%	1.9%
40	1.3%	1.7%
50	1.2%	1.5%
100	0.8%	1.0%

Discussion

The analysis in this article has shown that marks averaged across multiple junior examiners have greater predictive power than the definitive marks agreed by senior examiners. This is particularly evident in subjects such as Humanities where differences in professional opinion are more likely to arise. By extension, this implies that many junior examiners are better than a PE alone. Building on this evidence, the analysis has suggested a lower bound for the reliability of definitive marks themselves. These results have some practical implications.

Marker monitoring

Given the evidence that, particularly in subjects where professional judgement is an important element of marking, averaged marks are more accurate than definitive marks, it seems reasonable that, where possible, the former should form the basis for evaluating the performance of markers rather than the latter.

It is worth noting that the process of setting definitive marks remains useful as it may help to cement an idea of the marking standard that can be communicated to all markers. Furthermore, in practice, because exam boards require marker monitoring to be operational as soon as marking begins, and cannot wait until sufficient junior markers have marked seed scripts, marker monitoring would still need to work from a set of definitive marks to begin with. However, in rare cases where groups of junior examiners consistently disagree with the definitive mark, the evidence here suggests that the original definitive mark could reasonably be overwritten with the mean of the marks from AEs. At present where such situations occur, PEs are asked to review the original definitive marks to verify that they are the marks they intended. In practice, they can prove unwilling to shift their opinion of the correct mark for a script. The evidence in this article suggests that, since collectively assigned marks from AEs are generally better than those from a single PE, unless there was other evidence of a problem with the panel of AEs, the original definitive mark could be legitimately overwritten without the agreement of the PE. That is, if AEs as a group generally agree with the PE, then, in the rare cases where they disagree, the opinion of many AEs should take precedence.

Of course, if definitive marks were overwritten, it would be beneficial to ensure that any truly aberrant markers were excluded from this calculation. Alternatively, a robust mean, estimated using statistical methods that are resistant to outliers, could be employed.

Post-series evaluation

Recognising that marks from PEs are unlikely to provide an absolute truth with regard to the correct mark that each script should be awarded, and also noting the relatively small number of seed scripts used in analysis, has important implications for the use of such data for reviewing marking quality at a national level. In particular, it indicates that findings with regard to the overall accuracy of marking for individual exam components based on this data should be treated with some caution. Thus, although it may be desirable that analysis of such metrics will allow exam boards to "channel additional resource and support to those components or qualifications which most need improving" (Ofqual, 2018a, p.35) it is also possible that such metrics could be subject to random variation from year to year depending upon the exact selection of seed scripts.

Furthermore, since the evidence suggests that collectively assigned marks are superior to definitive marks, it would make sense for the post-award marking metrics generated by Ofqual to be based around the former rather than the latter. That is, rather than focusing on differences with definitive marks, these metrics should evaluate differences from the (more accurate) collectively assigned marks.

Marker accuracy

Finally, by looking at the difference between (very accurate) marks from PEs and (even more accurate) marks defined collectively by groups of AEs, we can derive a lower bound for the extent of difference we should expect between examiners. Furthermore, this article has shown how this lower bound for marking accuracy relates to the structure of the exam paper. The relevant formulae in this paper may help to set the tolerances for marker monitoring.

These results are important as they may help to manage expectations regarding the level of consistency between marks that it is possible to achieve through training and increased experience. After all, as we have seen, it is likely that even the best and most experienced markers, the PEs themselves, display some level of inconsistency for examination questions requiring professional judgement. If we are dissatisfied with the estimated levels of agreement between (principal) examiners suggested in this paper then the route to improve reliability is unlikely to lie in greater marker training or making slight tweaks to mark schemes. To dramatically improve reliability, far-reaching changes, such as altering the structure of assessments, or the number of markers who mark each exam, may be required.

Kathleen Tattersall, the very first chief regulator of Ofqual, warned against simplistic expectations that the marking system should be perfect (Clark, 2008). The research in this article provides rather more

8. Based on a quick review of a number of GCSE Physics papers available online.

detail about the likely scale of imperfection. Having an honest understanding of what can be achieved is important if we are to ensure that the demands placed upon an assessment system are realistic.

References

Benton, T. (2017, November). *Pooling the totality of our data resources to maintain standards in the face of changing cohorts.* Paper presented at the 18th annual AEA-Europe conference, Prague, Czech Republic.

Benton, T. (2018, November). *Exploring the relationship between optimal methods of item scoring and selection and predictive validity*. Paper presented at the 19th annual conference of the Association for Educational Assessment in Europe, Arnhem-Nijmegen, The Netherlands.

Benton, T., & Gallagher, T. (2018). Is comparative judgement just a quick form of multiple marking? *Research Matters: A Cambridge Assessment publication, 26*, 22–28.

Black, B., Suto, I., & Bramley, T. (2011). The interrelations of features of questions, mark schemes and examinee responses and their impact upon marker agreement, *Assessment in Education: Principles, Policy & Practice, 18*(3), 295–318.

Bramley, T. (2008, September). *Mark scheme features associated with different levels of marker agreement.* Paper presented at the annual conference of the British Educational Research Association, September 3–6, in Edinburgh, UK.

Clark, L. (2008, May 15). Pupils should not expect to receive 'accurate' results from exams, warns new watchdog. *Mail Online*. Retrieved 29th May, 2019, from http://www.dailymail.co.uk/news/article-1019938/Pupils-expect-receive-accurate-results exams-warms-watchdog.

Ofqual. (2016). *Marking consistency metrics*. Ofqual report, Ofqual/16/6121. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_da a/file/681625/Marking_consistency_metrics_-_November_2016.pdf.

Ofqual. (2018a). *Marking consistency metrics: An update*. Ofqual report, Ofqual/18/6449/2. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_da a/file/759207/Marking_consistency_metrics_-_an_update_-_FINAL64492.pdf.

Ofqual. (2018b). *A study of hard-to-mark responses: Why is there low mark agreement on some responses?* Ofqual report, Ofqual/18/6449/5. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_da a/file/759216/HardtoMark_-_FINAL64495.pdf.

Suto, I., Nádas, R., & Bell, J. (2011). Who should mark what? A study of factors affecting marking accuracy in a biology examination. *Research Papers in Education, 26*(1), 21–51.

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S. Fourth Edition*. New York: Springer.

"Learning Progressions": A historical and theoretical discussion

Tom Gallacher GL Assessment[1] and **Martin Johnson** Research Division

Introduction

Learning Progressions (LPs) are a relatively recent approach that aim to support three aspects of education: teaching and learning, assessment, and curriculum design. According to Schmidt, Wang, and McKnight (2005) the effectiveness of these three aspects of education may be increased by better coherence, and the LP approach claims to improve coherence by providing frameworks of knowledge and skills called "LP models". These frameworks describe the progression that can be expected of learners through their education (Gotwals & Songer, 2013). LP approaches are popular and influential across the fields of education and curriculum development, with discussion being carried out across a number of international contexts (Australian Council for Educational Research, 2018; E. M. Kim, Haberstroh, Peters, Howell, & Nabors Oláh, 2017; H. Kim & Care, 2018). This suggests that the consideration of the approach is topical.

This article outlines the specific objectives of the LP approach, the mechanism by which LP models may attain these objectives, and finally, the likelihood of this attainment (based on previous evidence). LPs should only be expected to achieve their aims if the assumptions of the LP approach are correct; however, our view is that the evidence suggests that the assumptions embedded within the frameworks are overly simplistic. Education is complex and the implementation of the LP approach to teaching and learning, assessment, or curriculum design may have unintended consequences when implemented without consideration of other possible approaches.

Proponents of the LP approach display a minimal engagement with previous theories of learning, and their ideas have been criticised as being "the latest manifestation of a much older idea, that of regularity in the development of students as they learn a certain body of knowledge or professional practice" (Wilson, 2009, p.716). This suggests that LP proponents should also consider the similarities of their theory with previous work to derive an approach that is most likely to attain its desired objectives.

Objectives of LPs

In order for LPs to benefit teaching and learning, assessment, and curriculum design, the approach needs to have a theory of learning that satisfies the practical and theoretical demands of the professionals involved in all three areas (Black, Wilson, & Yao, 2011).

1. The work was carried out when the first author was a member of the Research Division.