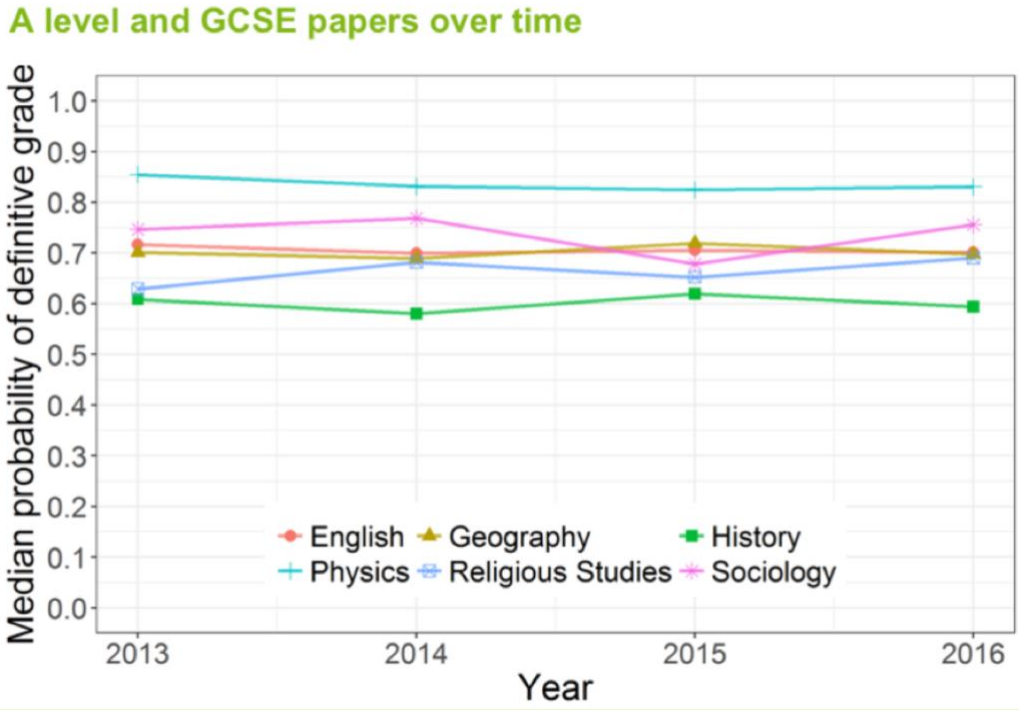
The grading system is broken – but marking is OK

Dennis Sherwood, dennis@silverbulletmachine.com

Here is a slide from a presentation given by Dr Michelle Meadows, Ofqual's Executive Director for Strategy, Risk and Research, at a symposium held on 29th June 2017:



Source: https://www.gov.uk/government/news/presentations-from-ofquals-summer-series-symposium-2017

The chart shows some results from an extensive research project (see https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/681625/Marking_consistency_metrics_-_November_2016.pdf) in which many tens of thousands of CGSE and A level papers were double-marked, with one mark given by an 'ordinary' marker and the other by a 'definitive' marker. So, to make that real, suppose that the first mark is, say 54, and the 'definitive' mark 56.

It's immediately apparent that these marks are different – but that's no surprise: to quote paragraph 5 of the blog entitled *GCSE, AS and A level marking reviews and appeals: 10 things you need to know*, posted on 3rd June 2016 by Ofqual's Julie Swan,

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

(see https://ofqual.blog.gov.uk/2016/06/03/gcse-as-and-a-level-marking-reviews-and-appeals-10-things-you-need-to-know/).

Both the marks 54 and 56 are therefore “different but appropriate”. One, however, is ‘definitive’ (that’s the 56), whilst the other (54) isn’t. So what?

So this. Suppose that grade C is defined as all marks from 53 to 57 inclusive. In this case, both of the marks correspond to a C, and the grade as determined by the mark (54) given by the first marker is the same as the ‘definitive’ grade, as determined by the ‘definitive’ mark (56).

But suppose that grade C is all marks from 51 to 55 inclusive, with Grade B from 56 to 60. In this case, the grade corresponding to the first mark, 54, is a C, but the ‘definitive’ grade, as determined by the ‘definitive’ mark 56, is a B. The grades are different. Despite the fact that both marks are “appropriate”.

This explains the chart: for each subject, the chart answers the question “Suppose that an entire cohort of subject entries is double-marked, with one mark being ‘definitive’. What percentage of that cohort would be awarded the same grade according to the marks given by both the first marker, and also the second, ‘definitive’, marker?”. This is important, for, in reality, a candidate’s script is marked only once, and it is the first marker’s mark that determines the grade that the candidate is actually awarded. If that first-awarded grade is the ‘definitive’ grade, that’s fine. But if it isn’t...

As an example, take physics, as shown by the light blue line at the top of the chart, with data points at about 85%. What that number means is this: for an entire physics cohort, at both GCSE and A level, and in each of the years from 2013 to 2016, about 85% of the candidates were awarded the ‘definitive’ grade. And, by inference, 15% weren’t - they were awarded a grade different from the ‘definitive grade’, perhaps a higher grade, perhaps lower. If we regard the ‘definitive’ grade as ‘right’, then any other grade must be ‘wrong’. So a more dramatic, but nonetheless valid, interpretation of the chart is “for GCSE and A level physics, in each of the years 2013 to 2016, about 85% of candidates were awarded the ‘right’ grade in the results as first published, and about 15% of candidates were awarded the ‘wrong’ grade”.

And for English Language, about 70% of the grades awarded were ‘right’, and 30% ‘wrong’; for history, 60% were ‘right’, and 40% ‘wrong’. To make that real: in summer 2016, 513,285 candidates sat GCSE English Language, of whom about 360,000 were awarded the ‘right’ grade when the results were announced that August. And about 150,000 candidates were awarded the ‘wrong’ grade.

You might like to read that paragraph again. And the same applies to A level.

The take-home message from all this is very straight-forward, if alarming. GCSE and A level grades have been hugely unreliable.

Why so?

Well, the immediate response is likely to be “because marking has been so bad”. And as we all know, there have been all sorts of issues around the recruitment, training, motivation and remuneration of markers.

But pause for a moment, and look at the chart again. For it contains two, significant, hidden messages.

The first is that *nothing that either Ofqual, or the Boards, have done over the period 2013 to 2016 has had any measureable impact on improving the reliability of grades*. Apart from the ‘wobble’ in 2015 sociology, the lines in the chart are pretty flat. So the reliability of grades has been unchanged over these four years, and the best that can be said is that any actions that Ofqual and the Boards have taken has stopped matters getting worse. Might matters have improved in 2017? Might the chart show a ‘kick’ up towards 100% in all subjects? I don’t think so...

The second message requires a ‘thought-experiment’. Let’s suppose that the grade reliability problem is indeed attributable to bad marking. If that is the case, then the chart tells us that, year-on-year, those who marked physics did a more reliable job that those who marked geography, who in turn did a more reliable job that those who marked history. Personally, I find that very hard to believe. I can see no reason at all why “all physics markers are conscientious” and “all history markers are sloppy”. Or why “those who exercise quality control over the marking of physics scripts are very strict” whilst “those who exercise quality control over the marking of history scripts don’t give a monkey’s”. It makes no sense. If the problem were attributable to poor marking, or flabby quality control, then I would have expected no particular correlation by subject: in some years, history would rank higher than physics; in others, not.

But the correlation by subject is there, stark, consistent. So the problem must be attributable to a feature of the *subject*, and not to a feature of the *people involved in marking that subject*.

What might that feature be? Let me call it ‘fuzziness’.

We all know that subjects like physics and maths are ‘tight’. Although, as Julie Swan’s blog tells us, it is quite possible for different markers to give the same question “different but appropriate” marks, in physics and maths, the range of those different marks will be narrow. History, though, is different: the range is likely to be much broader. To use my term, history is more “fuzzy” than physics, and so my interpretation of the position of the lines in the chart is in terms of “fuzziness”: physics is the least “fuzzy” subject and history the most, with the other subjects somewhere in-between. To me, that makes much more sense.

But if it does, there is a very important conclusion. The unreliability of grades is *not driven by ‘marking error’*: it’s driven by a broken grading system.

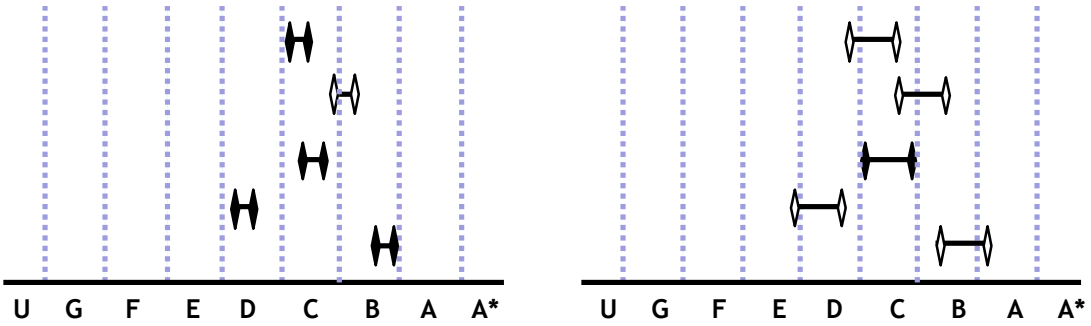
Let me go back to Julie Swan's blog, for it tells us everything we need to know to explain what's going on. Yes, there may be "*nothing wrong or unusual about*" the fact that "*it is possible for two examiners to give different but appropriate marks to the same answer*". But there can be one XXXX of a problem. If the range of possible marks that are "*different but appropriate*" are within the same grade width, then the grade is robust. But if that range straddles a grade boundary, then the grade is determined by the lottery of which mark from this range is given first. The broader this range, the greater the likelihood that the range will straddle a grade boundary, and so the more unreliable the grade. And if this range is even broader, then it is possible for two, or even three, grade boundaries to be straddled, which is a real mess.

This range is the "fuzziness" I have just referred. Physics is a relatively "unfuzzy" subject, so the range of "*different but appropriate marks*" is correspondingly narrow. The likelihood of straddling grade boundaries is therefore lower – but it still happens: the chart shows that physics grades are only 85% reliable. History is much more "fuzzy", and the range of feasible marks is wider. The likelihood of straddling a grade boundary is therefore greater, and so history grades would be expected to be less reliable than physics grades, as validated by the chart which shows history grades to be only 60% reliable.

The chart is important. It measures the reliability of grades for six subjects, over four recent years, and the implications of the numbers are startling. Thousands – or rather hundreds of thousands – of young people are being awarded the wrong grades, every year.

The chart further shows that the underlying problem is *not about marking*. Despite the problems with recruitment, training, motivation and remuneration, 'marking errors' have no significant effect: markers do a good job. But what does have a significant – very significant – effect is the straddling of grade boundaries, fundamentally attributable to an inherent feature of our examination system: that "*it is possible for two examiners to give different but appropriate marks to the same answer*", itself attributable to our use of open-ended essay-style questions, rather than closed, right/wrong, multiple choice.

Even if marking were absolutely perfect, *the straddling of grade boundaries will still happen*, for, as shown in the following figure, it is impossible to identify a single, unique, 'right' grade when a "fuzzy mark" straddles a hard-edged grade boundary. Marking is OK. It's the way that grades are determined that is broken. Badly broken. And it needs to be fixed.



"Fuzzy" marks, and hard-edged grade boundaries. On the left is a relatively "unfuzzy" subject, as shown by the narrow range of marks associated with each candidate's submission. For four candidates, this range lies within a single grade width, and so the grades awarded do not depend on which particular marks, within the "fuzzy" range, were given to each paper. For one candidate, however, the range straddles a grade boundary, and so that candidate's grade is unreliable. On the right is a subject associated with greater "fuzziness", but with grade boundaries in the same locations as the subject on the left. Now only one candidate has been awarded a reliable grade, whilst four have been awarded an unreliable grade.

How, then, might it be possible to solve the problem? That's very easy to answer. Since the problem is all about mapping "fuzzy" marks onto hard-edged grade boundaries, one way of solving the problem is to throw the grade boundaries away. Rather than giving grades, why don't we declare each candidate's mark, along with a measurement of the corresponding "fuzziness"? That is both clear, and fair. And there are some other possible solutions too. We need to have a project which looks into all this, and makes sensible recommendations. Does such a project have your support? I hope so. Remember that in 2016, about 150,000 young people were awarded the wrong grade for GCSE English Language. The numbers are similarly large in the other subjects then, and in summer 2017 too. This is unfair. It does harm. And it's totally unnecessary. This injustice can, and should, be fixed.