Written evidence submitted by Dennis Sherwood

Context

The hearings on 2nd and 16th September were wide-ranging, and, quite naturally, paid considerable attention to the past, so as to gain an understanding of the decisions that were made, and of the events that took place, between mid-March and the current time. This threw light on many matters, but some are still unresolved – such as what exactly happened between 16th and 20th March, and who, precisely, took the decision to design and build an algorithm "to produce a calculated grade for each student".

Furthermore, a number of documents and other forms of evidence remain to be disclosed, not least the release of schools' data to "a trusted third party … for a deep forensic analysis", as was agreed by Roger Taylor in his answer to Question 985.

A key remaining problem – on average, 1 exam grade in every 4 is wrong, and will continue to be wrong unless action is taken

This submission, however, looks forward to all future exams, whenever they might take place, to make the case that the Committee should take urgent action to resolve a fundamental and long-lasting problem: ***over recent years, on average, about 1 exam grade in every 4 has been wrong.***

To make that real: in each of the last several years, out of the approximately 6 million GCSE, AS and A level grades awarded, ***some 1.5 million have been wrong, more than 1.4 million of which were neither 'discovered' by the appeals process, nor resolved.***

When exams are resumed in October and November, and in summer 2021, and thereafter, ***unless action is taken now, 1 in every 4 of those grades will be wrong too.***

I believe that this problem must be solved with urgency.

The assessments resulting from all future exams must be reliable and trustworthy.

Action required

A number of possible ways of ensuring that exam assessments are reliable and trustworthy are briefly described here. There may also be others.

The action required is therefore to commission a study:

1. To identify all possibilities.

2. To evaluate each wisely, and to agree the best.

3. To implement that best solution.

How the required study might be delivered

The 'obvious' way to deliver the required study is for Ofqual to say, "This is our business, leave it to us".

I believe that Ofqual, in its current form, should not be entrusted to undertake this study.

So either:

1. Ofqual in its current form is disbanded, and the study carried out by Ofqual's successor; or

2. The study should be carried out by an independent panel, the members of which should be determined by the Committee.

My belief is that the second option will deliver the better result; it is also an action that can be taken, and implemented, as soon as the Committee sees fit, for its initiation depends on no other matters.

Why Ofqual in its current form should not do this

Firstly, because Ofqual have not already resolved this problem in the past. This is not because the problem was unknown; rather, it is because Ofqual have chosen, deliberately, not to address it.

As evidence for this statement, I cite two sources:

1. Ofqual's announcement of 11th August 2019 which states "This is not new, the issue has existed as long as qualifications have been marked and graded".

2. The absence of any reference in Ofqual's Corporate Plan 2020-21 to take action to ensure that grades are reliable and trustworthy.

Secondly, because Ofqual's track record in evaluating ideas wisely, fairly, and without bias is poor, as evidenced by Ofqual's paper dated 16th March 2020, entitled "Summer 2020 GCSE and A/AS level exam series: Contingency planning for Covid-19 – options and risks".

This paper documents the results of Ofqual's evaluation of 11 possible ideas as to how this year's grades might be awarded, with Annex 3 (pages 10 to 16) identifying the "advantages" and "risks" of the three short-listed options, and Annex 4 (pages 17 to 24) presenting Ofqual's evaluation of the eight options which were rejected on the grounds of being "less likely to meet our objectives".

As can be seen, the table in Annex 4 analyses the various rejected options in terms of "arguments for" and "arguments against".

Although the phrases "arguments for" and "arguments against" suggest that these "arguments" are intrinsic and absolute, this is not the case. Rather, what is portrayed as an "argument for" is in fact an argument used by a PERSON who is in favour of an idea; likewise, an "argument against" is an argument used by a PERSON who is against an idea. To evaluate ideas according to "arguments for" and "arguments against" is therefore a mask for a particular position, and so is inherently biased. This is bad practice: when evaluating ideas, it is essential that the process is done fairly and without bias.

An explicit example of this inherent bias is to be found on page 24, on which the analysis of one particular idea shows an "argument against" stated as "This would call into question the future of GCSEs".

There are many who do indeed "call into question the future of GCSEs", and who would consider that any possibility that might cause this to happen is a good idea, not a bad one. Ofqual's portrayal of this view as an "argument against" therefore proves Ofqual's bias.

This example is just one of many to be found in Ofqual's "Options" paper of 16th March. If this document is representative of how Ofqual identify and evaluate ideas, then this is proof that they should not be trusted to do so in general, and especially in the particular instance discussed in this submission.

Appendix – Verification that "on average, 1 exam grade in 4 is wrong"

My claim concerning the reliability of exam grades

In this submission, I have claimed that "on average, 1 exam grade in 4 is wrong".

Ofqual have never said this; on the contrary, they have often vigorously denied it, as, for example, in the complaint that Ofqual made to the Independent Press Standards Office in connection with an article that appeared in the Sunday Times on 11th August 2019.

The purpose of this Appendix is to validate my claim.

Ofqual's statements concerning the reliability of exam grades

Although Ofqual have never stated that "on average, 1 exam grade in 4 is wrong", they have made several statements that acknowledge that exam grades are not fully reliable and trustworthy – for example, at the Meeting of 2nd September:

1. In response to Question 1058, Dame Glenys Stacey stated that exam grades are "*reliable to one grade either way*".

2. In response to Question 996, Dr Michelle Meadows stated that she "*took solace*" that "*98% of A-level grades and 96% of GCSE grades are accurate plus or minus one grade*".

Furthermore, in a statement posted to Ofqual's website on 11th August 2019, in response to the Sunday Times article just mentioned, Ofqual state that "*…more than one grade could well be a legitimate reflection of a student's performance…*".

These three statements share a common feature. They are all vague and imprecise. They all hint that exam grades might not be as reliable as we might hope, and those numbers, 98% and 96%, sound most reassuring. But no statement declares, clearly and unambiguously, what, precisely, the reliability of exam grades actually is.

To do that, we need to dig deeper, and follow a trail signposted by Dr Michelle Meadows in her answer to Question 974:

"*Every year, we publish marking consistency metrics that report the extent to which grades would change if a different senior examiner had looked at the work. In fact, we looked at that work this year and took some comfort from it, in the sense that the levels of accuracy that we were seeing from the standardisation model were very similar to those levels of accuracy that we see each year through the marking process.*"

The evidence that, on average, 1 exam grade in 4 is wrong

This explicitly refers to "marking consistency metrics", and also links these metrics to the "levels of accuracy that we were seeing from the standardisation model" (of which more shortly).

Dr Meadows claimed that these metrics have been published "every year". I have been unable to trace these publications, but I have found one document that does indeed do this – an Ofqual report dated November 2018 entitled "Marking Consistency Metrics – An update".

The most important feature of this report is a diagram, Figure 12 on page 21, showing measures of the average reliability of the GCSE, AS and A-level grades for each of 14 subjects, reproduced here, and referred to in this text as 'Figure 12':

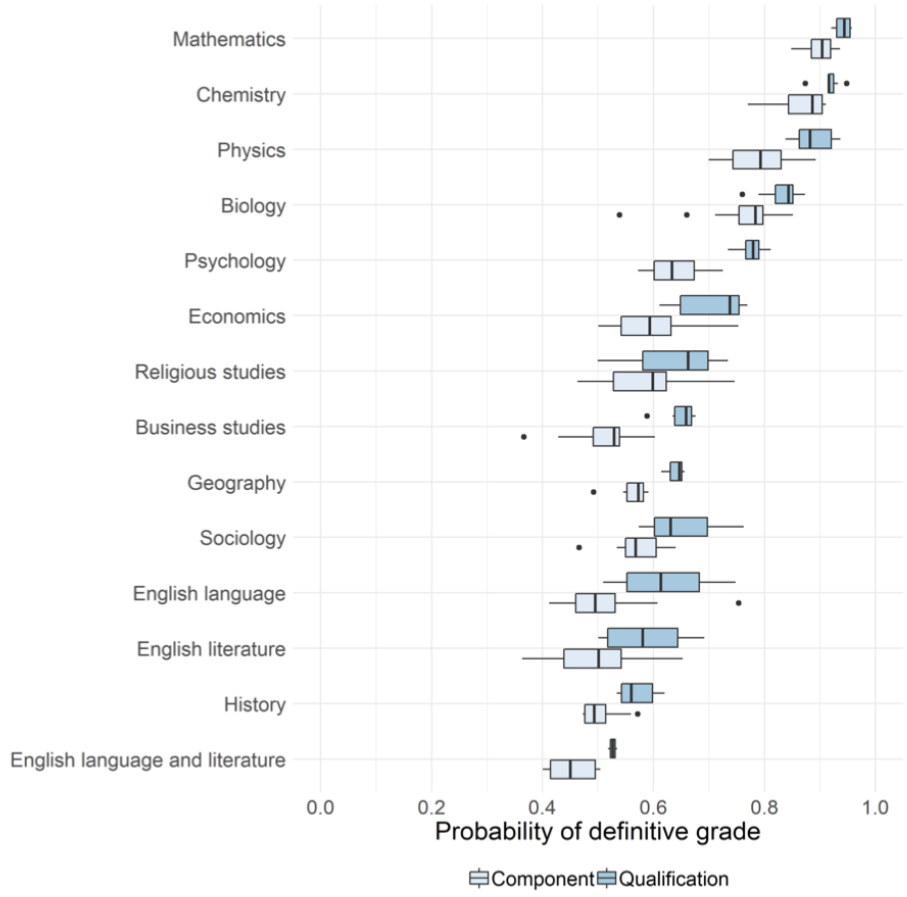

Figure 12. Boxplot showing the comparison of the probability of being awarded the 'definitive' grade at component and qualification level, for those GCSE, AS and A level qualifications for which we have full component data.

In this diagram, for each subject, the heavy line in the darker blue box answers this question:

"If the scripts submitted by an entire subject's cohort of students were to be fairly and independently marked twice, firstly by an 'ordinary examiner', and secondly by a 'senior examiner' (whose mark determines what Ofqual call the "definitive grade"), for what percentage of scripts would the originally-awarded grade be confirmed?"

The re-mark process is a 'second opinion' – a second opinion from an individual whose mark (and hence grade) is regarded by Ofqual as "definitive" (or indeed "true", as stated in Ofqual's earlier report, Marking Consistency Metrics, of November 2016).

If grades were fully reliable, then, for all subjects, the answer to the question would be "100% of originally-awarded grades are confirmed by a fair re-mark by a senior examiner".

Ofqual's chart, however, shows that this is not the case.

Rather, the average reliability – as measured by matching the senior examiner's 'second opinion' to the originally-awarded grade (such that a reliability of 100% corresponds in Figure 12 to a 'probability' of 1.0) – varies by subject from about 96% for (all varieties of) Mathematics to about 52% for Combined English Language and Literature.

If each of those subjects is weighted by the corresponding subject cohort, then the average reliability across all the 14 subjects shown is about 75% (I have the corresponding data, which is available on request).

I know of no data for other subjects, but I have made some estimates (available on request), and I am confident that the average reliability of all GCSE, AS and A-level grades, across all subjects, is about 75% – meaning a number unlikely to be less than 70% or greater than 80%.

In reality, that implies that if all 6 million scripts, as typically submitted for each summer's exams, were to be fairly re-marked by a senior examiner, some 4.5 million (about 75%) of the originally-awarded grades would be confirmed, and around 1.5 million (about 25%) grades would be changed, approximately half upwards and half downwards.

I think it unlikely that any student, whose grade is changed as the result of a fair re-mark, would say, "Ah! My originally-awarded grade must have been non-definitive!". More likely, I believe, would be, "That first grade must have been wrong!".

Hence my "headline" statement that, on average, 1 exam grade in every 4 is wrong.

Reconciling "1 grade in 4 is wrong" with Ofqual's statements

The statement that "on average, 1 exam grade in 4 is wrong" still needs to be reconciled with Ofqual's acknowledgement that "98% of A-level grades and 96% of GCSE grades are accurate plus or minus one grade".

To do this, I draw on my own simulation (details of which are available on request) of the exam results of 2019 A-level English Literature.

If the grades were 100% reliable, then on a fair re-mark by a senior examiner, all 40,824 candidates would be given the same grade as the original grade.

But according to Figure 12, only 58% of the entries – that's about 23,695 candidates – would have the original grade confirmed by a fair re-mark by a senior examiner, whatever that grade might be.

My simulation shows that a senior examiner's fair re-mark would result in a further 16,409 candidates being graded either one grade higher (8,085) or one grade lower (8,324) than the original grade. The total given either the original grade, or one higher, or one lower (that's "accurate plus or minus one grade" or "reliable to one grade either way") is therefore 23,695 + 16,409 = 40,104, this being 98% of the total cohort of 40,824, which is Ofqual's statement.

Finally, 402 candidates would be re-graded two grades higher, and 318 two grades lower, giving a total of 720 students two grades adrift.

The totals reconcile as

Grade confirmed	23,695	(58%)			
One grade adrift	16,409	(40%)	Sub-total	40,104	(98%)
Two grades adrift	720	(2%)	Grand total	40,824	(100%)

Ofqual's statements that "exam grades are reliable to one grade either way", "98% of A-level exam grades are accurate plus or minus one grade", or that "…more than one grade is a legitimate reflection of a student's performance" are all true.

But they mask a deeper truth.

That, on average, 1 exam grade in every 4 is wrong.

A fact that Ofqual have known for many years.

But done nothing to resolve.

This is not difficult to do. The problem is not primarily caused by "marking error"; rather it is a result of Ofqual's policy for determining grades – a

policy that fails to recognise the reality that "it is possible for two examiners to give different but appropriate marks to the same answer". If those two "different but appropriate marks" are within the same grade width, then there is no problem. But if they are on different sides of a grade boundary, then there is a very big problem indeed, for the grade that appears on the certificate is the result of the lottery of which examiner happened to mark the script, and which side of the grade boundary that mark happens to lie.

This is not an infrequent or rare event.

On the contrary.

It is a very frequent event indeed, and explains why 1 exam grade in every 4 is wrong.

The importance of Figure 12 to this year's process

A critically important step in the development of this summer's algorithm was the testing of its accuracy – as noted by Dr Michelle Meadows in her reply to Question 974, as already cited:

"Every year, we publish marking consistency metrics that report the extent to which grades would change if a different senior examiner had looked at the work. In fact, we looked at that work this year and took some comfort from it, in the sense that the levels of accuracy that we were seeing from the standardisation model were very similar to those levels of accuracy that we see each year through the marking process."

The details of how this was done are described in Ofqual's "Interim Report", on page 81 of which is this diagram:

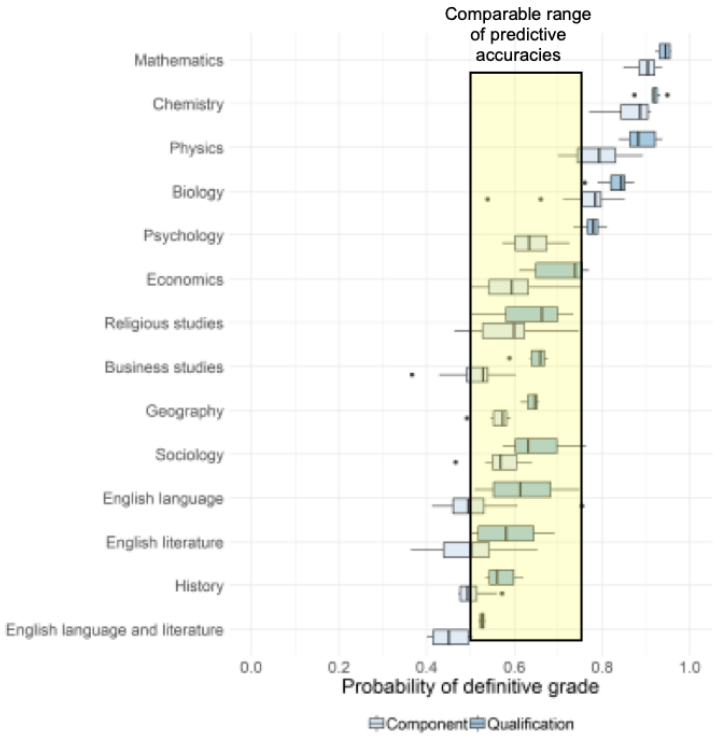

Figure 7.25 Probability of definitive grade being awarded based on an analysis of marking consistency reproduced from Ofqual 2018.

As can be seen, superposed on Figure 12 is a 'yellow block' representing the range of reliabilities ("predictive accuracies") resulting from the use of the algorithm. And it was this comparison in which Dr Meadows "*took some comfort… that the levels of accuracy that we were seeing from the standardisation model were very similar to those levels of accuracy that we see each year through the marking process*".

Figure 12 was therefore the 'benchmark' against which the results of the algorithm were tested; Figure 12 was the definition of 'good'.

Yet as this submission has shown, Figure 12 is the evidence that 1 exam grade in 4 is wrong, and has been wrong for years.

Which Ofqual have known for years.

But has done nothing to remedy. Nor has any intent to do so.

Hence the urgency of the Select Committee's action, as suggested in this submission.

September 2020