Dennis Sherwood –Written evidence (EDU0007)

1. As the Committee appreciates, there is much interest in – and pressure for – reform of the educational system in general, and assessment in particular. Many important issues are on the table, including:

   - ensuring that the curriculum is well-suited to the needs of mid-21st Century society
   - the purpose of assessment
   - the role, if any, of high-stakes examinations, especially at age 16
   - the impact of technology and AI

   to name just a few.

2. These are all important and complex, and it will take many years for wise reforms to be formulated, agreed and successfully implemented.

3. But there is one issue that is both important and urgent, for it is doing much damage, now.

4. Whilst exams exist, the very least we should expect is that the outcomes of those exams – currently, the grades on candidates' certificates – should be fully reliable and trustworthy.

5. The fact, however, is that they are not.

6. As Ofqual themselves acknowledge, grades "are reliable to one grade either way".

7. Are grades "reliable to one grade either way" reliable enough? Especially in the context of Ofqual's statutory obligation under S128 of the Apprenticeships, Skills, Children and Learning Act 2009 "to secure that regulated qualifications give a reliable indication of knowledge, skills and understanding"?

8. There are many ways to award assessments that approach 100% reliability. None is 'perfect', but some are much better than others. As a matter of urgency, and as the first step along the reform path, an independently-led feasibility study should be conducted to identify the best solution for

subsequent implementation. In the meantime, all certificates, from summer 2023 onwards, should clearly show, to quote Ofqual's own words:

OFQUAL WARNING: THE GRADES ON THIS CERTIFICATE ARE RELIABLE ONLY TO ONE GRADE EITHER WAY

Context

9. I write this submission following the Committee's meeting of 30 March 2023, the focus of which was assessment.

10. As Lord Baker observed, the witnesses – Dr Michelle Meadows (Oxford University), Gavin Busuttil-Reynaud (AlphaPlus/AQA), Sharon Hague (Pearson), and Tim Oates CBE (Cambridge Assessment) – are all "believers in the *status quo*".

11. I too have a vested interest. To secure that assessments are fair for all young people. Which they currently are not. To ensure that the full truth is presented to the Committee. Which did not happen at the meeting.

12. Hence this submission.

Grade reliability – the evidence

13. Two important statements relating to the (un)reliability of GCSE grades were given in evidence to the hearing of the House of Commons Education Select Committee held on 2 September 2020:

    - that grades are reliable to one grade either way (Q1059, statement from Dame Glenys Stacey, at that time Ofqual's Chief Regulator)
    - that 96% of GCSE grades are accurate to plus or minus one grade (Q997, statement from Dr Michelle Meadows, at that time Ofqual's Executive Director for Strategy, Risk and Research).

14. Although these statements are slightly different, both highlight that a certificate showing, for example, "GCSE English, Grade 3" cannot be trusted. The grade the candidate truly merits might be grade 3 as awarded, but it might also be grade 2, or perhaps grade 4, or even grade 1 or grade 5. The difference, however, between grade 3 and grade 4 is potentially life-

changing, for grade 3 implies a 'fail' (as explored by Lord Knight), causing the student to lose a year of development, and to re-sit the exam, with possible adverse consequences on the student's mental health.

15. Unreliable grades do great damage, and these two statements – each given by an authoritative source – indicate just how unreliable grades are.
16. These statements were highlighted by Lord Watson, whose question invited the witnesses to comment on their implications.

The witnesses' responses

17. None of the witnesses pushed back on Lord Watson's question, or sought to refute the fundamental point that grades are (at best) reliable to one grade either way, and that about 200,000 GCSE grades are at least two grades adrift.
18. 200,000 is a very large number of extremely unreliable grades – especially given that the number of GCSE candidates is fewer than 800,000.
19. All the witnesses therefore accepted the truth that grades are indeed unreliable. Three did so tacitly; Dr Meadows did so explicitly when she said:

> "***It's really important that people don't put too much weight on any individual grade."***

20. Given the high-stakes nature of the exam system in England, that is a truly remarkable statement, especially when stated by Ofqual's former Executive Director for Strategy, Risk and Research.
21. Do students know that? Or parents? Teachers? Employers? Admissions officers?
22. Perhaps they should – so one recommendation might be to ensure that those words are printed, boldly, on every certificate. Or if not those, then perhaps Dame Glenys Stacey's near equivalent:

OFQUAL WARNING: THE GRADES ON THIS CERTIFICATE ARE ONLY RELIABLE TO ONE GRADE EITHER WAY

23. The witnesses made many other statements too, some of which I now examine.

Two 'wrongs' do not make a 'right'

24. Dr Meadows subsequently said, "*when you look at the whole [for example, all 8 grades on a student's GCSE certificate], you've probably got a very reliable indication [of a student's overall capabilities].*"

25. This brings to mind the possibility that an erroneously low grade in one subject is compensated by an erroneously high grade in another.

26. This does, of course, appear to be most reasonable.

27. Any such inference, however, is false.

28. For GCSE, AS and A level exams in England, each subject counts on its own. An erroneously low grade 3 in English still requires the student to re-sit, even if the same student is awarded an erroneously high grade 7 in History.

29. Similarly, an erroneously low grade 4 in Chemistry might deny a student entry to the A level Chemistry course, and so block any opportunity for that student to become a doctor, even if the student also achieved an erroneously high grade 9 in Physics.

30. Two wrong grades, one too high and one too low, do not 'cancel each other out'. Rather, they are both wrong.

"The reality of assessment"

31. Dr Meadows also said, "*I know, unfortunately, particularly for GCSEs in Maths and English, that a lot of weight is placed on [individual subject grades] … In English, that is problematic, but this is not a failure of our GCSE system – this is the reality of assessment. It's the same around the world. There is no easy fix I'm afraid. So I think it's how we use assessments*

that has to change, rather than creating a system of very lengthy assessments."

32. This statement claims that the unreliability of grades "*is the reality of assessment*", and that there are "*no easy fixes*" – a point emphasised by Dr Meadows in her earlier statement that "*to actually get 100% reliability would be technically pretty much impossible without the most extraordinarily long assessments*".

33. These words indicate that Dr Meadows believes that there is nothing that Ofqual and the exam boards can do to resolve the problem of unreliable grades, save for "*a system of very lengthy assessments*", which, by implication, nobody would accept. The problem is therefore insoluble, and so the only mitigation is for students, parents, colleges, universities and employers to change their behaviours as to "*how [they] use assessments*".

34. Dr Meadows's statement is half-true: this is indeed "*the reality of assessment*" as undertaken by Ofqual and the exam boards since at least 2010.

35. There are, however, many other ways of awarding assessments, some of which are by no means "*very lengthy*", but highly effective and easy to implement. Furthermore, assessments can be delivered that approach 100% reliability as closely as might be required. 22 possibilities are described briefly here, and in more detail in Chapters 14 and 15 of my book *Missing the Mark – Why so many exam grades are wrong, and how to get results we can trust*.

36. It could therefore be very worthwhile to examine these 22 possibilities (and any others that can be discovered) in detail, comparing them not only to each other but also to the *status quo*, so as to identify, wisely, the most suitable, and fairest, way to determine assessments from examinations. Such a feasibility study, however, should be led by an independent body – as Lord Baker pointed out, the current authorities have vested interests to protect.

37. Dr Meadows's claims that "*there are no easy fixes*", and that the only possible alternative is "*a system of very lengthy assessments*", are therefore false.

38. As Dr Meadows knows well, for I still have the notes of a meeting with Dr Meadows and her colleague, Dr Paul Newton, which took place at Ofqual's offices on 11 January 2017, at which I presented the key features of one of the easy solutions (similar to that described here).

"Assessing choreography"

39. The response from Sharon Hague included these words: "*The Ofqual research was referring specifically to subjects where examiners could have legitimately awarded different marks and the impact this eventually has on grades – so they were assessing choreography, for example, or creative writing…*"

40. This too is half-true and, if not half-false, then certainly misleading.

41. Yes, it is true that "*the Ofqual research was referring specifically subjects where examiners could have legitimately awarded different marks and the impact this eventually has on grades*". But the reference to "*choreography*" and "*creative writing*" implies that the "*subjects where examiners could have legitimately awarded different marks"* are somehow 'special' or 'exceptional'.

42. The Ofqual research to which Ms Hague referred is reported in Ofqual's November 2018 publication, Marking Consistency Metrics – An update. The key findings – which relate not only to GCSE but to AS and A level too – are shown in Figure 12, to be found on page 21 of that report, as reproduced here:

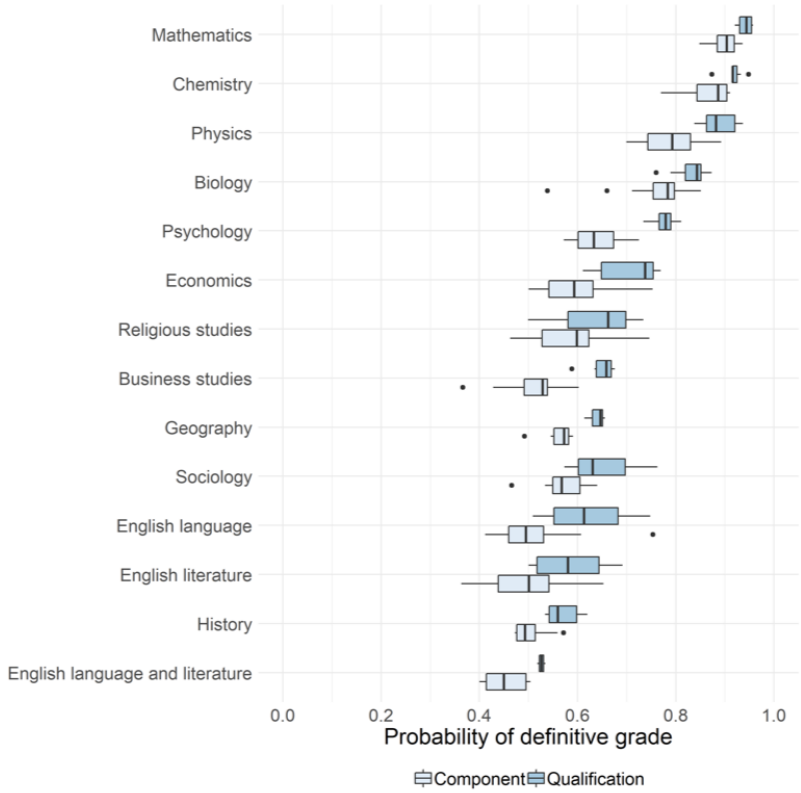

Figure 12. Boxplot showing the comparison of the probability of being awarded the 'definitive' grade at component and qualification level, for those GCSE, AS and A level qualifications for which we have full component data.[8]

43. As can be seen, the subjects studied range from Mathematics to (the A level subject only) Combined English Language and Literature.

44. Ms Hague's response implies that exams in only some subjects are associated with the possibility of "*legitimately different marks*", and that those subjects are especially subjective.

45. The truth is that exams in all subjects, including Mathematics, are such that "*legitimately different marks*" are possible, with the consequence that the resulting grades are all unreliable, to different degrees in different subjects.

46. The key measure of the reliability of the grades associated with each subject is shown by the heavy black line in the darker blue box. For example, for Geography, the reliability is about 0.65 (65%), implying that for every 100 students, about 65 are awarded what Ofqual refer to as the "definitive" or "true" grade (or more simply "right" grade), and the remaining 35 are awarded a grade which, presumably, is "non-definitive" or "false" – or, in plain English, "wrong".

47. For Mathematics, grades are about 96% reliable (96 out of 100 right, 4 wrong); History, about 56% (56 out of 100 right, 44 wrong – that's nearly half); the all-subject overall average is about 75%, or, in simpler language, about 1 GCSE, AS and A level grade in every 4 is wrong. That's the truth.

48. To make that real, in the summer 2023 exams:

    - of the 6 million grades awarded about 1.5 million were wrong …
    - … for every 10 students taking 8 GCSEs, only about 1 received a certificate on which all 8 grades were right; about 9 received a certificate on which at least one grade was wrong …
    - … all without (as I will describe shortly) any right of appeal.

"Less than 1% of grades are changed as the result of [the] challenge process"

49. Ms Hague then continued by emphasising the integrity of marking, which is true – the quality of marking is high. She further pointed out that "*less than 4% of GCSE grades were actually challenged last year*" which is also true. Her next statement was this: "*less than 1% of results were changed as a result of that challenge process*". This is true too. But it can be most misleading.

50. The (true) statement "*less than 1% of results were changed as a result of that challenge process*" can trigger in the listener's mind a thought such as "if only 1% are changed, that must mean that the remaining 99% are right", which appears to be somewhat reassuring. Ms Hague did not actually say that. But the Pearson website does: if you click here, you will see, referring to the summer 2017 exams, the claim that "99.25% of our grades were accurate on results day".

51. The (true) fact that only 1% of grades are changed does not imply that the remaining 99% are correct. A grade can be changed only if it is appealed – and as Ms Hague (correctly) said, about 4% of grades were appealed. 96% of grades were therefore not appealed, so the key question is "how many of those grades would have been changed had the corresponding appeals been raised?".

52. This is the question that Ofqual's research asked. And to answer it, Ofqual double-marked entire cohorts of scripts, once by an 'ordinary' examiner, and again by a senior examiner, giving the results shown in the chart on page 6 – results that show that only 75%, not 99%, of grades are correct.

53. Pearson should know this. And also know how misleading the statement "*less than 1% of grades are changed as the result of that challenge process*" can be.

54. One further point about appeals. Currently, a script can be re-marked only if a "review of marking" discovers a "marking error", such as a failure to comply with the mark scheme. Suppose, however, that a script is originally marked, say, 64, and that had a senior examiner marked the same script, the "definitive" or "true" mark would have been 66. The mark 64 is not in error, but a legitimate difference of academic opinion. To use Ofqual's language, the mark 64 is "within tolerance" of the senior examiner's "definitive" mark 66. This is an example of the well-acknowledged fact that different examiners can give the same script (slightly) different marks.

55. The grade on the candidate's certificate is based on the original mark, 64, so if the B/A grade boundary is 65/66, this is grade B. The senior examiner's mark 66, however, corresponds to the "definitive"/ "true"/"right" grade A, implying that the grade B shown on the candidate's certificate is "non-definitive"/"false"/"wrong".

56. But since there are no "marking errors", Ofqual's current rules prevent this grade error from being discovered and corrected. Of the 6 million grades awarded in summer 2022, the 1.5 million grades that were wrong therefore cannot be appealed, and remain wrong. For ever.

57. I believe this to be unjust.

58. I note that Ofqual changed the rules for appeals, introducing the "marking error" test, in May 2016, just six months before publishing the first evidence that 1 grade in 4 is wrong in November 2016.

Conclusion

59. For the last many years, and – as is quite likely – for some time into the future until more major reforms are implemented, assessment of young

people in England will continue to be based on high-stakes exams, with hard-edged grade boundaries and one-symbol grades. The very least we should expect is that those grades are fully reliable and trustworthy.

60. But they are not.

61. On average, about 1 grade in 4 is wrong, not only at GCSE, but at AS and A level too.

62. The key evidence was published in 2018, with preliminary evidence published in 2016. However, that grades are unreliable has been known for many years: here, for example, is an extract from page 70 of a report published by AQA in 2005:

    However, to not routinely report the levels of unreliability associated with examinations leaves awarding bodies open to suspicion and criticism. For example, Satterly (1994) suggests that the dependability of scores and grades in many external forms of assessment will continue to be unknown to users and candidates because reporting low reliabilities and large margins of error attached to marks or grades would be a source of embarrassment to awarding bodies. Indeed it is unlikely that an awarding body would unilaterally begin reporting reliability estimates or that any individual awarding body would be willing to accept the burden of educating test users in the meanings of those reliability estimates.

63. As shown on the report's cover, the lead author is Dr Michelle Meadows.

64. Lord Watson's question was precise, clear, and incisive.

65. I leave it to the members of the Committee to judge the quality, completeness, and integrity of the witnesses' answers.

66. In conclusion, I acknowledge that this submission makes many claims and allegations, as well as expressing some direct opinions. May I confirm my willingness to submit more material, or to appear before the Committee as a witness, so that my claims and allegations can be substantiated, and my opinions challenged.

6 April 2023