Cambridge Assessment

Research *Matters* / 31

A Cambridge Assessment publication

ISSN: 1755–6031

Journal homepage:
https://www.cambridgeassessment.org.uk/our-research/all-published-resources/research-matters/

On using generosity to combat unreliability

Tom Benton

To cite this article: Benton, T. (2021). On using generosity to combat unreliability. *Research Matters: A Cambridge Assessment publication, 31*, 22–41.

To link this article: cambridgeassessment.org.uk/Images/research-matters-31-on-using-generosity-to-combat-unreliability.pdf

Abstract:

Assessment reliability can be affected by various types of unforeseen events. In any such circumstances where a concern is raised that the reliability of assessment is lower than usual, our natural inclination is to allow extra leniency in grading to reduce the chances of students missing out on a grade they deserve. This article shows how, by focusing on the risk for individual students, we might logically approach this situation in deciding exactly how much additional generosity is required. In particular, it shows how making progress with this problem requires an acceptance that no assessment system is perfect and transparency about the level of reliability that is achievable. Having developed an approach, this article also shows how this may lead to different outcomes than the competing desire to maintain assessment standards so that the group of students in question are not unfairly advantaged relative to previous and future cohorts.

Cambridge Assessment

Cambridge Assessment is committed to making our documents accessible in accordance with the WCAG 2.1 Standard. We're always looking to improve the accessibility of our documents. If you find any problems or you think we're not meeting accessibility requirements, contact our team: Research Division, researchprogrammes@cambridgeassessment.org.uk

If you need this document in a different format contact us, telling us your name, email address and requirements and we will respond within 15 working days.

On using generosity to combat unreliability

Tom Benton Research Division

Introduction

This article concerns how we might think about acting in a situation where we know that, for whatever reason, an assessment is less reliable than normal. At the time of writing, the instance in mind is the cancellation of all exams in England in summer 2020 and the ultimate decision that grades should (largely) be awarded based upon teacher estimates in the form of Centre Assessed Grades (CAGs) instead.[1] However, the same principles may apply in other scenarios such as where, for practical reasons, it may be necessary to shorten the length of some assessments or alter their composition in some other way.

This article will not attempt to address all of the varied aspects of fairness that may be relevant to such a situation. A review of issues of fairness, such as whether qualification grades should simply be fair in terms of treating all students equally, or whether they must provide a "fair" reward for the effort put in by students is provided by Shaw and Nisbet (2021, this issue). Similarly, we will not be considering all of the different factors that may have affected different students to a greater or lesser extent over the past year and how assessment can be made fair in the light of this. Rather, this article will focus on a single issue—the positioning of grade boundaries. We will explore whether psychometrics can provide a rational approach to setting grade boundaries in a situation where an assessment system is less reliable than would typically be the case.

Once we know that the reliability of an assessment is lower than normal, the natural human inclination is to try to ensure that students do not lose out. Reduced reliability will increase the chances of a student ending up with a better grade than they would have achieved with a more reliable assessment. However, usually the greater concern is that it will also increase the chances of students being awarded a worse grade than would otherwise have been the case. Since the reduced level of reliability is due to circumstances that are beyond students' control, it seems unfair that some of them may end up with lower grades than they would have achieved under ordinary circumstances. As such, we may wish to ensure that assessments are graded leniently so as to reduce the number of students getting a lower grade than they would have done, while accepting this will mean more getting a higher grade than they would have done. It is clear that we would not know which particular individuals benefit and which are disadvantaged by the unreliability.

A straightforward way to incorporate leniency into assessment is during the awarding process when we select grade boundaries. Setting grade boundaries sufficiently low reduces the difficulty of achieving each grade and may mitigate the impact of reduced assessment reliability on students. This article considers this issue in more detail. In particular, the aim is to discover exactly how much generosity is needed during awarding

1 A discussion of why we believe such estimates may be less reliable than full-length exams will be provided in a later section.

to combat increases in unreliability, and how this depends upon the strategy to awarding that we take. Interestingly, we will show that, in some respects, the final distribution of grades in summer 2020 was similar to what might be expected from a logical application of giving students the benefit of the doubt from a position of uncertainty about how they would have performed in real exams.

True scores, true grades, observed scores, awarded grades, reliability

Before going any further, we need to define our terms. From now on, we define a student's *true score* (or *true mark*) as the score that they would achieve on a 100 per cent reliable assessment (i.e., an assessment that was very long and had completely reliable marking). True scores could be transformed to follow any statistical distribution, so, for the purposes of this article, we define *true grades* as being defined by students' percentile rank rather than by their score. For example, we might decide that the top 25 per cent of students should be awarded grade A (or above). Note that the concepts of "true scores" and "true grades" are entirely notional to help us think about issues relating to reliability. In reality we never observe the true score of any student. However, for the purposes of this research we imagine that we could in order to think about issues such as student misclassification.

Observed scores are the scores that are actually assigned to students including the measurement error inherent in any real assessment. Note that error, in this context, does not necessarily need to mean "mistake". Rather, it refers to anything that may lead to changes in students' scores between replications of an assessment process. For example, certain items will suit some students more than others, meaning that students' performances usually vary between different test versions. Similarly, there may be differences in the professional judgement of examination markers, meaning that re-marking may lead to changes in students' scores. We define *awarded grades* or *observed grades* as the grades awarded to students based on their observed scores.

Following usual practice in classical test theory, we define *reliability* as the squared correlation between true and observed scores. This is equivalent to the percentage of the variance in observed scores attributable to variance between students' true scores. It is also mathematically equivalent to the expected correlation between observed scores on two parallel tests (very roughly speaking, these are tests made to the same specification so that, in particular, they consist of equally difficult items and are of equal length).

Ensuring that no student is disadvantaged

In a perfect world we would ensure that, even with reduced reliability, absolutely no student ends up with a lower grade than they deserve. In the midst of a crisis regarding an assessment, statements to this effect are often released by the organisations responsible for them. What is usually left unsaid is that we would, under the circumstances, perhaps accept some students being awarded higher grades than they deserve in order to ensure that no student misses out. This initial section considers exactly how generous we would need to be in order to achieve this goal.

Figure 1 provides an imaginary example to introduce these ideas further. The plot shows the relationship between simulated true marks for 1,000 students on the horizontal axis and their observed marks from a sub-optimal assessment process on the vertical axis. For the purposes of this simulation, both true marks and observed marks were simulated to be between 0 and 100, to have a mean of 50 and a standard deviation of 16 (see footnote[2]). The correlation between true and observed scores was simulated to be 0.8—that is, the reliability of observed scores was 0.64 (=0.8^2).

Next, to begin with, imagine that we have initially decided that the top 25 per cent of students should be awarded grade A. For example, we may have initially considered that, despite the increased unreliability, we would like to preserve the grade distribution from previous years so that the grades awarded to students were indistinguishable from those awarded under the normal assessment process. As such, we identify grade boundaries on both the true and observed marks that achieve this. These are represented in Figure 1 by the solid black lines and are 61 out of 100 in each case.

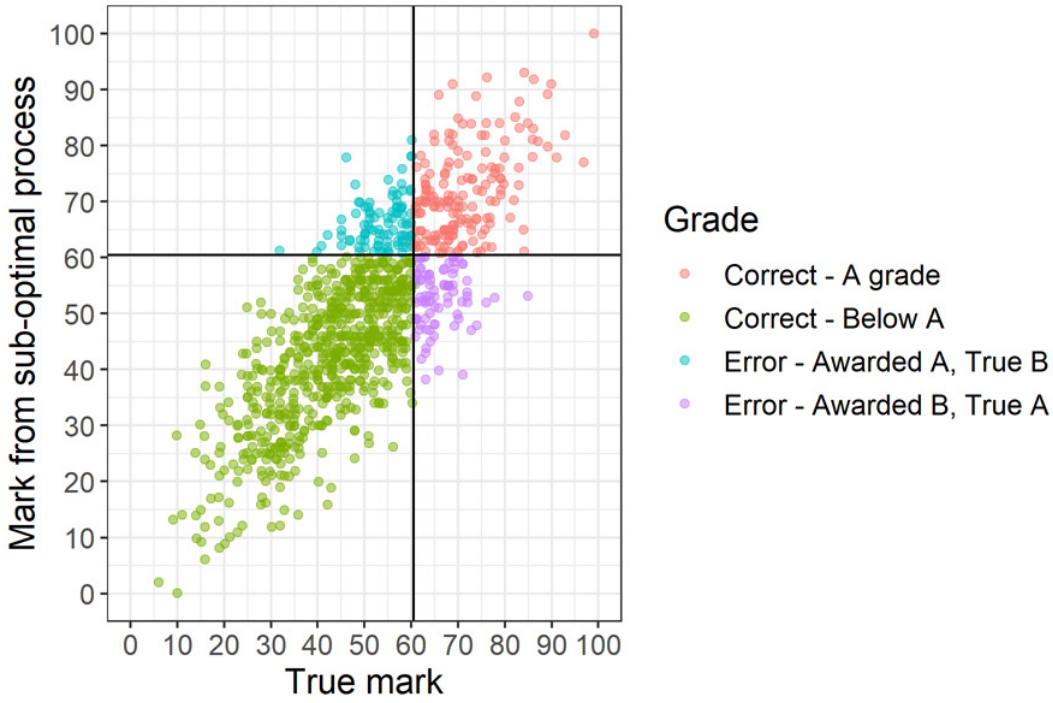

Figure 1: Relationship between true marks and grades, and marks and grades from an unreliable assessment if we choose to award the same percentage of top grades.

Suppose next that, having realised the possibility of students missing out on the grade they deserve, we decide to lower the grade boundary to eradicate this risk. This process is illustrated in Figure 2. For the purposes of this figure, the grade boundary on observed scores has been lowered to the point where no student is awarded a grade lower than they deserve. This has required lowering the grade boundary from 61 out of 100 to 38 out of 100. Furthermore, in total, only 464 students are now awarded the correct grade. The good news is that all 254 students that deserved a grade A have been awarded this (the red points in Figure 2). However, this has come at a cost. Lowering the grade boundary

2 These are reasonably typical score distributions for examinations. For example, see Table 1 of Wheadon and Stockford (2010).

has led to 536 students being awarded a grade A when, according to their true scores, they did not deserve this (the blue points).

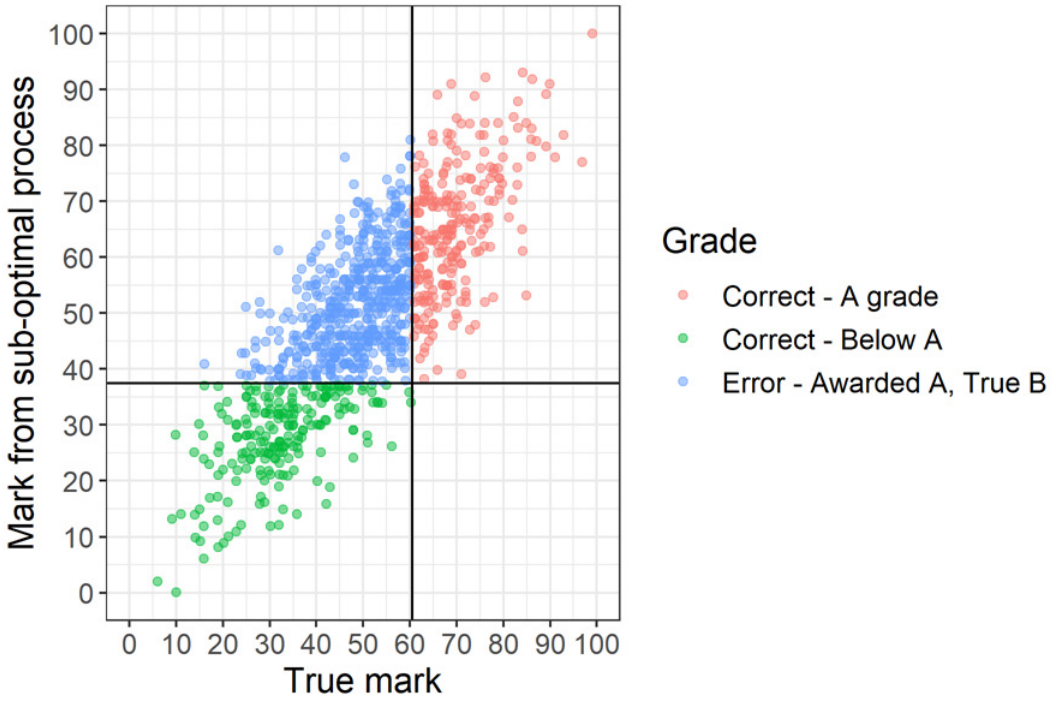

Figure 2: Relationship between true marks and grades, and marks and grades from an unreliable assessment if we lower the grade boundary for the unreliable set of marks.

Figure 2 demonstrates the difficulty of ensuring no student is disadvantaged by unreliability. Measurement error can affect students both positively and negatively and to differing amounts. In the presence of unreliability, with a sufficiently large number of students, there will always be at least one person whose observed score is far below their true score. Lowering grade boundaries far enough to mitigate this risk in its entirety would require an enormous dose of generosity. This in turn risks undermining the meaning of having achieved a particular grade, devaluing the currency of the qualification in question, and undermining public confidence. As such, ensuring that absolutely no student is disadvantaged is not something that can realistically be achieved through the awarding process alone.

Figure 2 imagines that we are trying to ensure that no student is disadvantaged relative to their true score. However, true scores are never observed and so in reality "an assessment will never be perfectly reliable" (Wheadon & Stockford, 2010, p.32). As such, "it is inevitable that some candidates with true scores in one grade will, on some occasions, achieve a score just outside that grade" (Wheadon & Stockford, 2010, p.32). With this in mind, while we might accept that we cannot lower grade boundaries to a point where no student achieves a grade below the one they deserve, we may be able to lower them to an extent where the risk of a student being under-graded by an abnormally unreliable assessment is no higher than the risk we accept under normal circumstances. This is the focus in the next part of this article.

Reliability of various assessment processes

Before we can consider this issue, we need some idea of the reliability of assessments under normal circumstances. Our focus will be on A Levels and will be on the reliability of entire qualifications rather than the individual examinations (components) which comprise these. Although several existing pieces of research have published data on the reliability of individual examination components (e.g., Bramley & Dhawan, 2010; Hayes & Pritchard, 2013; Wheadon & Stockford, 2010) very few of these have attempted to calculate reliability at whole qualification level. In order to estimate the reliability of a typical A Level, we use the following pieces of information:

- At present, a typical A Level will require students to complete at least three separate examination components which have roughly equal weight.
- The median reliability (Cronbach's alpha) of individual A Level examination components is 0.83 (Bramley & Dhawan, 2010, Table 2.5). Note that since Cronbach's alpha is often considered a lower bound for the true reliability of a component (see Hayes & Pritchard, 2013), this means that the above estimate of the reliability of a full A Level may be slightly conservative.
- Analysis of A Level components taken with OCR in June 2019 reveals the median correlation between examination components taken as part of the same A Level qualification (subject), is 0.64 (based on 173 pairs of components).[3]
- Combining the above facts with a simplified version of the Wang and Stanley formula for calculating composite reliability (He, 2009) gives us:

$$\textit{Typical A Level reliability} = \frac{3\alpha + 6\rho}{3 + 6\rho} = \frac{3 * 0.83 + 6 * 0.64}{3 + 6 * 0.64} = 0.925$$

  Where α is the expected reliability of each individual component, ρ is the expected correlation between components, "3" is the number of components, and "6" is the number of ways in which three components can be paired (respecting the direction of pairings).

As such, we will use a reliability of 0.925 as representing the reliability of a typical A Level. This estimate is a little lower than the estimate of 0.963 for a full-length Mathematics A Level provided by Benton (2014). This may reflect the reduced amount of assessment in reformed A Levels,[4] the use of the (more conservative) Cronbach's alpha in the present calculations, or the focus on a typical A Level subject (using medians) rather than exclusively on Mathematics.

To provide a richer context for the calculations in subsequent sections, we will also estimate the "reliability" of an alternative, and less reliable, assessment method—forecast scores. The word "reliability" is placed in inverted commas in the previous sentence to represent the fact that, as discussed earlier, we are defining reliability as the squared

3 Restricted to components taken by at least 200 students.

4 Since recent reform, A Levels no longer combine results from assessments taken over a period of two years and, instead, focus on (fewer) examinations taken in a single session.

correlation between any given measure and true scores from examinations. This is not quite the same as the usual definition of reliability (i.e., how much would forecast scores change if we replicated the forecasting procedure) but it is the most relevant definition in the context of this article.

The word "forecast" is used above as the estimates of reliability (below) are derived from an analysis of A Level forecast grades submitted to OCR by teachers in May 2014 ahead of students completing their exams. They may differ from the predicted grades that were sent to UCAS (Universities and Colleges Admissions Service) earlier in the academic year as part of the university application process (see Gill & Benton, 2015, for further details). Our interest in forecast grades stems from the fact they are the best information we have available to say anything about the reliability of teacher estimated grades such as CAGs. Although we have collected data on CAGs, data on exam achievement of the same students in the same subjects is not available (because exams were cancelled). As such, comparing forecast grades and actual exam grades is the best source of information we have. Further discussion of the relationships between different types of teacher estimated grades is provided in McManus et al. (2020).

With the above discussion in mind, we estimate the reliability of forecast scores as follows:

- For OCR A Levels in summer 2014 the median polychoric correlation[5] between forecast and actual A Level grades was 0.82 (across 57 A Level specifications).
- Given the calculations earlier and adjusting for the unreliability of actual A Level results, we use a standard correction formula (Spearman, 1904) to estimate that the correlation between forecast scores and true qualification scores is 0.85 (=0.82/√0.925).
- Thus, the estimated "reliability" of forecast scores is 0.72 (=0.85^2).

Note that forecast scores, as such, have never existed—only forecast grades. However, we use the concept here to help us think about the likely reliability of procedures based upon teachers estimating the likely future achievement of their students (such as was done for the production of CAGs in summer 2020).[6] That is, the estimate gives an idea of the confidence with which teachers can estimate the future performance of their students. The idea of a continuous forecast score is useful as it allows us to apply concepts relating to the positioning of grade boundaries within this context.

The idea that results based on this type of teacher assessment are less reliable than those based on formal examinations has some support in the existing research literature. For example, McManus et al. (2020) showed that forecast grades tend to have considerably lower correlations with future undergraduate and postgraduate achievement at university than actual A Level grades.

5 We use a polychoric correlation to adjust for the fact that coarse scales (such as grades) are likely to be less highly correlated than fine scales (such as marks). The median Pearson correlation between forecast and actual grades was slightly lower at 0.77. Only A Levels with at least 200 entrants were included in calculations. See Gill and Benton (2015) for more details about the accuracy of forecast grades.

6 See Ofqual (2020).

It should be noted that the reliability estimate here only concerns a particular type of teacher assessment. The estimate does not necessarily apply to other types of internal school assessment such as where teachers mark pieces of standardised coursework or controlled assessment and these marks are then moderated. Previous research (Benton, 2016) has shown that, in at least some circumstances, such approaches to assessment are likely to be just as reliable as formal examinations.

Possible choices regarding standards in the face of added unreliability

To recapitulate, we wish to explore different approaches to setting grade boundaries in the situation where we have been forced to switch from a form of assessment with a reliability of 0.925 (approximate reliability of a full A Level qualification) to a form of assessment with a reliability of 0.72 (approximate reliability of forecast scores). We will consider three different approaches to setting grade boundaries:

- **Retain the grade distribution.** That is, we decide that despite the change to the reliability of qualifications we should retain the grade distribution we have always had. For the purposes of calculations in this section we will assume the grade distribution should match the overall grade distribution for A Levels from 2019 (JCQ, 2019).
- **Ensure no increase in the percentage of students awarded a grade below their true grade.** That is, we aim that the number of students awarded a grade below their true grade is no higher for the less reliable assessment method than for the normal assessment method. We abbreviate this method as the No Added Disadvantage strategy.
- **Maximise accuracy.** That is, we choose grade boundaries to maximise the percentage of students awarded a grade equal to their true grade. This leads to slightly different results from method 1. This will be explored more as we display the results.

To ease the calculations in this section, we will assume that all scores (both true and observed) follow normal distributions. Given that any set of scores could be transformed to follow a normal distribution (i.e., converted to normalised scores) this is not such a huge assumption. Since, after transformation, scores could be expressed on any scale, grade boundaries will be expressed in terms of the proportion of students at each given grade or above. Note that, based on these assumptions, all of the calculations in this section were completed entirely by applying mathematical formulae (i.e., there was no need to simulate data). Examples of the R code used to complete calculations are provided in the appendix.

To begin with, we focus upon setting an A* boundary. Results for other grades will be presented later but this provides a good starting point to illustrate the concepts in this research. Note that we define an A* as referring to the top 7.8 per cent of students.

Figure 3 shows the estimated misclassification rate at A* for a full-length A Level depending upon the percentage of students we award the grade to. The green curve shows how the proportion of students awarded a grade that is too high (i.e., an A* when their true grade is below A*) changes with the proportion awarded A* overall. If very few people

are awarded A* (the left of the chart) then hardly any will be over-graded. As the number awarded A* increases so does the proportion of students that are awarded a grade that is too high. The blue line is of more interest. This shows the percentage of students under-graded as the grade boundary moves. Harsh grading (to the left of the chart) results in a greater percentage under-graded than lenient grading (to the right of the chart). The red curve denotes the total level of misclassification—the sum of under-grading and over-grading.

The vertical purple line represents where we would position grade boundaries for the full-length qualification. That is, we award A* to exactly 7.8% of students. Because this assessment is not perfectly reliable (no assessment is) some students are misclassified. Specifically, given an assessment reliability of 0.925, 1.6% of students would be awarded an A* when their true grade is lower than this and, similarly, 1.6% of students would not be awarded A* when they deserve it. These values match as, by the definition we have used, the distribution of true grades equals the grade distribution on the full A Level. The dashed horizontal line demarks this level of over/under-grading and will be carried forward to the next chart.

For interest, we note that even for a full-length A Level, we can slightly improve classification accuracy with a slightly different grade boundary (the orange vertical line). Specifically, if we are interested in identifying the top 7.8% of students by true score (not observed score) then we should use a slightly harsher boundary (i.e., award fewer A*s). This is because, given that A*s are fairly rare, we need strong evidence to convince us that a student is most likely in this category rather than combining a slightly weaker true ability with a little luck with measurement error. Using this slightly harsher boundary would very slightly decrease the overall misclassification rate from 3.2% to 3.1%.

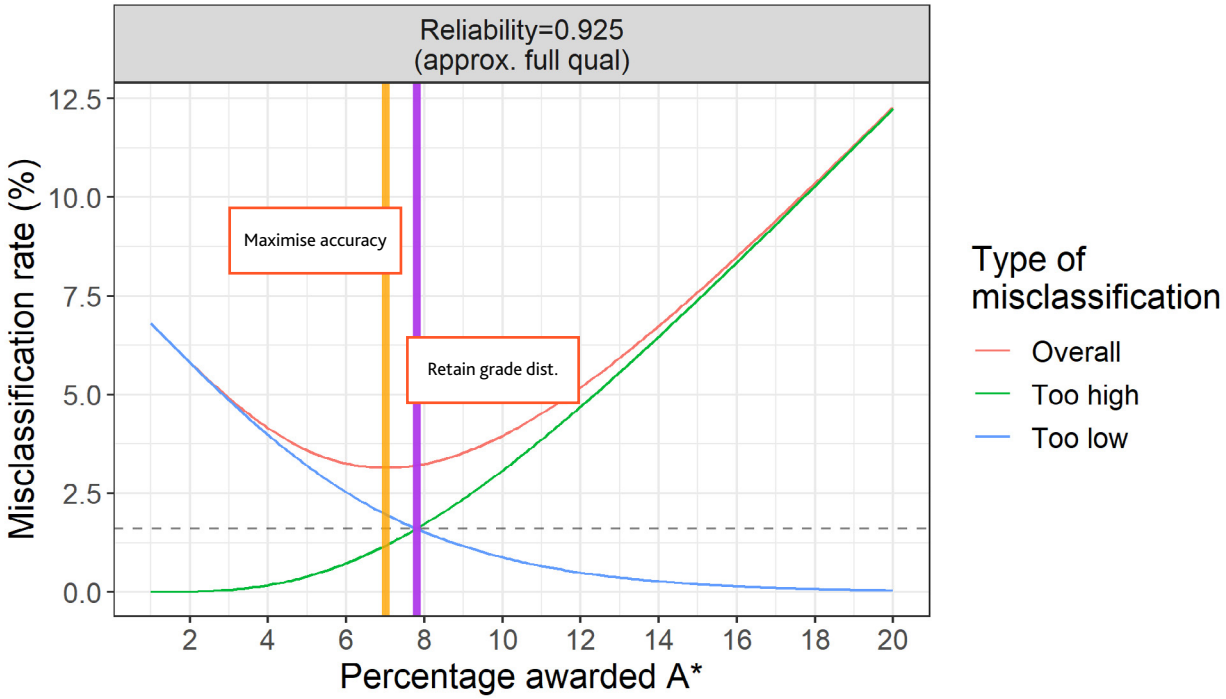

Figure 3: The relationship between percentage of students awarded A* and the misclassification rate at this grade for an overall assessment reliability of 0.925. The vertical lines denote pass rates associated with different awarding strategies.

Now we turn to our main topic of interest—the impact of our different strategies to setting grade boundaries as the reliability of the assessment drops. Figure 4 shows the relationship between the percentage of students awarded A* and the misclassification rate if the reliability of the assessment dropped to 0.72. The first thing to note is that, as we would expect, the overall misclassification rates rises. The pink vertical line indicates the results if we decide to continue to award an A* to the top 7.8% of candidates by their observed score despite the fact that we now have a less reliable assessment. Doing this results in the overall misclassification almost doubling from 3.2% to 6.3% with half of these students awarded a grade that is too low.

An alternative approach is to try to ensure that the proportion of students who are not awarded an A* when they deserve one remains at the target level (the dashed horizontal line, at 1.6%) that would be achieved by the full-length qualification. This goal is achieved by awarding A*s to 14.2% of candidates (the purple vertical line). In other words, in order to ensure that the reduction in reliability does not result in more students being under-graded than usual, we need to allow this many students to be awarded an A* in total. This point represents a logical maximum for the level of benefit of the doubt we might apply at A*. Going any further would mean even fewer students being under-graded than we would achieve in a normal year—not a rational response to a change in circumstances. Naturally, increasing the percentage of A*s we award to this extent will result in a big increase in the percentage of students who are over-graded. Specifically, this would mean that 8% of all students would be awarded an A* when they do not deserve one.

Finally, we consider the percentage of students we would award A* to if our goal was to maximise accuracy (i.e., minimise misclassification). This point is shown by the orange vertical line in Figure 4—reducing the percentage awarded A* to 4.7 per cent. This reduction in A* pass rate occurs because, given the unreliability of the assessment, we

need a lot of convincing that a student is genuinely among the top 7.8 per cent of true scores. In other words, we would need to see students even further towards the top of the observed score distribution.

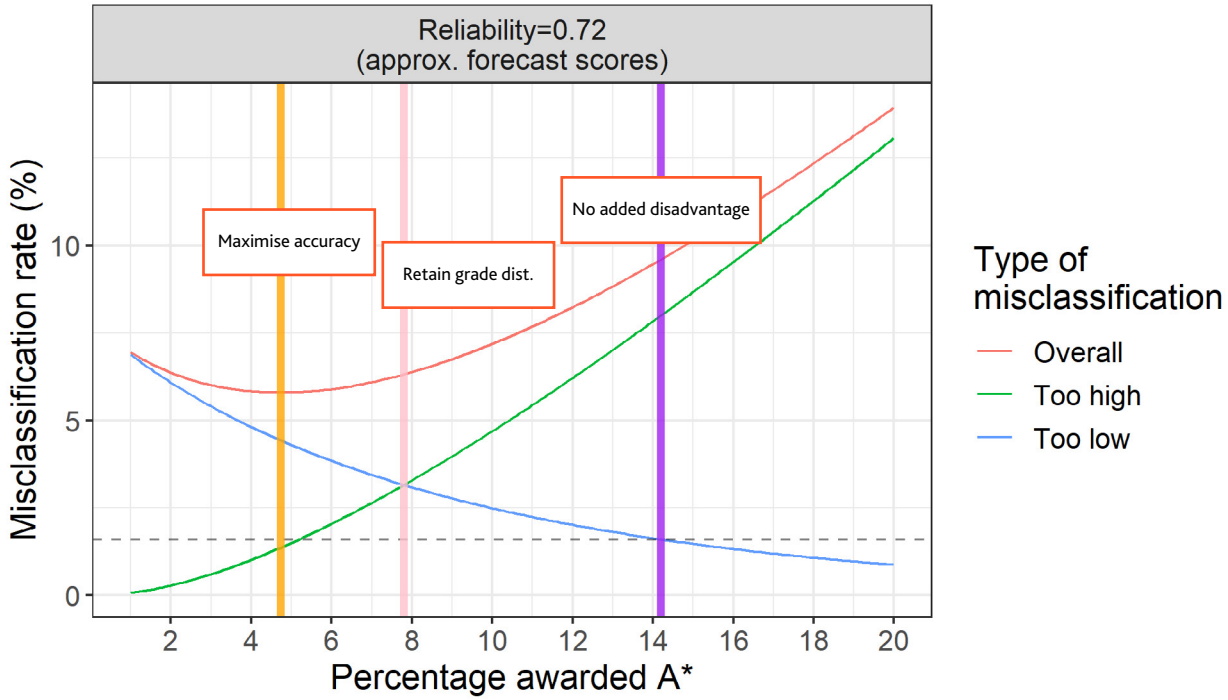

Figure 4: The relationship between the percentage of students awarded A* and the misclassification rate at this grade for an overall assessment reliability of 0.72. The vertical lines denote pass rates associated with different awarding strategies.

One way to look at Figure 4 is in terms of the balance between aggregate outcomes and individual risk. Assessment organisations are frequently accused of being too concerned with grade distributions at a national level and not enough on outcomes for individuals. With this in mind, it is worth noting that the horizontal axis in Figure 4 represents overall aggregate outcomes. Against this, the vertical axis represents the risk of individuals[7] being awarded a grade different from their true grade. The different strategies for choosing boundaries represent different choices regarding the balance of controlling aggregate outcomes versus managing individual risk. Retaining grade distributions places all of the emphasis on aggregate outcomes. The No Added Disadvantage strategy places all of the emphasis on ensuring the risk to individuals does not increase. It is worth noting that if our focus is on managing individual risk, while we would increase aggregate outcomes in a year where reliability is lower, we would subsequently reduce aggregate outcomes once reliability had risen back to normal levels (i.e., the situation in Figure 3).

Figure 5 shows the results of the same kinds of analysis focusing on grade E. Note that, based on the national A Level grade distribution in 2019,[8] for the purposes of calculations,

7 For example see https://www.theguardian.com/education/2020/aug/17/gavin-williamson-seeks-blame-ofqual-exams-debacle-a-level-gcse

8 See https://www.jcq.org.uk/wp-content/uploads/2019/08/A-Level-and-AS-Results-Summer-2019.pdf

97.6 per cent of students are defined to be at grade E or above. The bottom panel shows the results for a full-length qualification. Note that, since failing to achieve grade E is so rare, the overall misclassification rate at this grade is very low (just 1.2 per cent). Note also that, in contrast to grade A*, since true scores below grade E are rare, we need a lot of evidence to convince us that a candidate is not worth a grade E. As such, at this grade, classification accuracy is maximised by setting grade boundaries more leniently (the vertical orange line).

The top panel of Figure 5 shows the impact on misclassification if the assessment reliability drops to 0.72. As before, in order to ensure the number of students under-graded is the same as normal, the percentage of students awarded grade E and above should increase (the vertical purple line). However, what is different here from the analysis at grade A* is that, to maximise accuracy, the percentage awarded E or above should rise even further (the orange vertical line). This reflects the fact that, with an unreliable assessment, we would need a lot of convincing that a student was not worthy of at least a grade E.

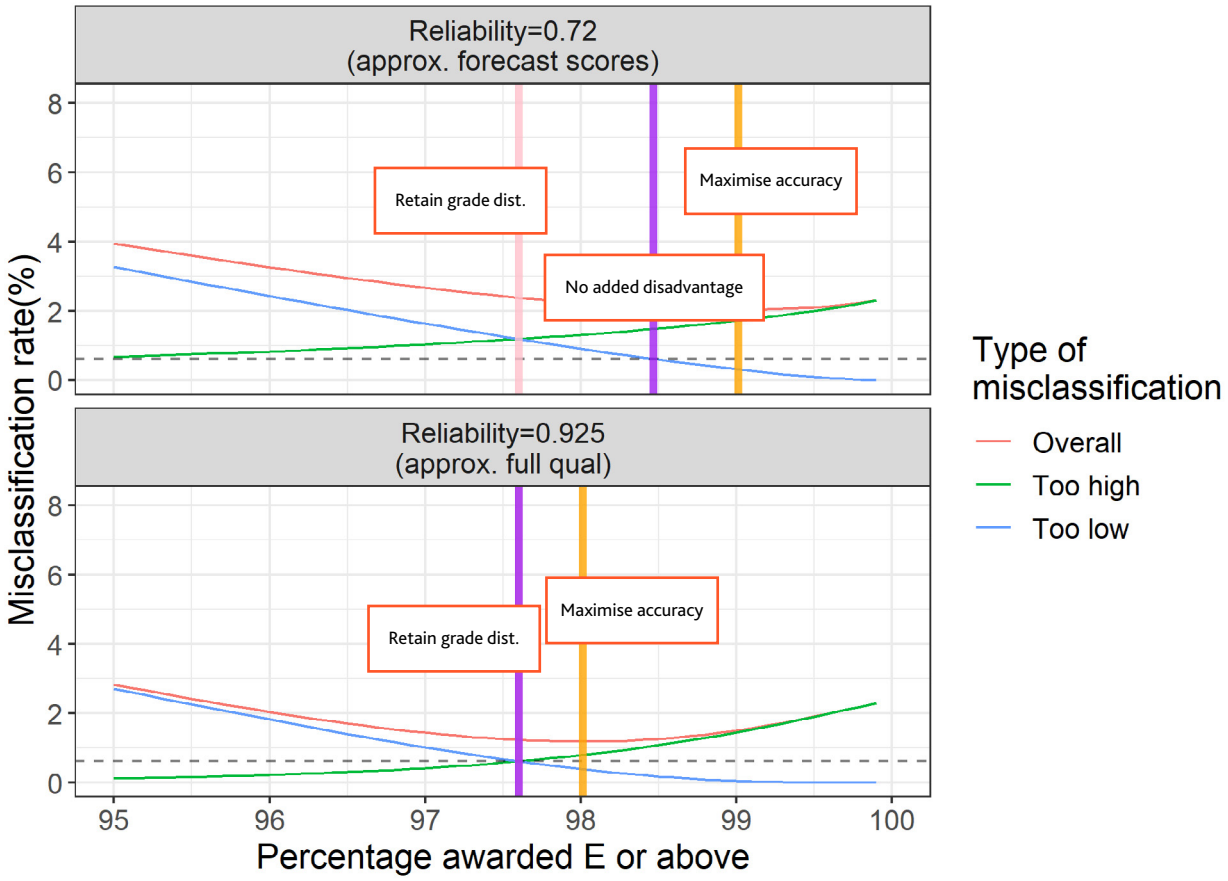

Figure 5: The relationship between percentage of students awarded E or above and the misclassification rate at this grade for overall assessment reliabilities of 0.72 and 0.925. The vertical lines denote pass rates associated with different awarding strategies.

The impact of different choices on grade distributions and misclassifications

Having illustrated our approach to thinking about the issue of using generosity to combat unreliability, this section presents a more complete set of results across all grades and across a greater range of assessment reliabilities. To begin with, we will present results relating to the strategy of ensuring that there is no increase in the percentage of students that are under-graded. After this, results relating to the strategy of maximising classification accuracy will be presented. In both cases we will compare results to the strategy of retaining the grade distribution regardless of the reliability of the test

Ensuring No Added Disadvantage

Table 1 shows how the No Added Disadvantage strategy affects the cumulative percentage awarded each grade as assessment reliability drops. The top row of the table (reliability=0.925) represents the assumed target grade distribution in a normal year. As can be seen, as the reliability of the assessment drops, so the required pass rate at each grade to avoid any increase in disadvantage goes up. For a small reduction in reliability (from 0.925 to 0.90), increases of around 1 percentage point at each grade are enough to mitigate the risks of decreased reliability. However, as reliabilities get lower, more drastic changes to the grade distribution are required.

Table 1: Percentage of students awarded each grade or above for different levels of reliability under the No Added Disadvantage strategy to setting grade boundaries.

Reliability	Cumulative % at each grade					
	A*	A	B	C	D	E
0.925	7.8%	25.5%	51.6%	75.8%	91.0%	97.6%
0.900	8.5%	26.8%	53.1%	76.8%	91.5%	97.8%
0.850	9.9%	29.4%	55.8%	78.7%	92.3%	98.0%
0.800	11.4%	32.0%	58.4%	80.4%	93.0%	98.2%
0.750	13.1%	34.6%	60.8%	81.8%	93.6%	98.4%
0.720	14.2%	36.2%	62.2%	82.6%	94.0%	98.5%

As described earlier, we estimate that the "reliability" of a teacher forecast approach to assessment is about 0.72. With this in mind, and taking this as our best guess of the reliability of CAGs, it is of interest to compare the grade distribution from the No Added Disadvantage approach with this level of reliability in Table 1 to the actual distribution of A Level CAGs awarded in summer 2020 (Table 2). What is interesting is that at grades A*, A and B there is a lot of similarity between a logical grade distribution based on the No Added Disadvantage strategy and the actual distribution of CAGs from teachers. For example, under the No Added Disadvantage strategy for a reliability of 0.72, we would increase the proportion of students awarded A and above from 25.5% to 36.2%. In reality, CAGs from teachers suggested that the proportion awarded A and above should rise

from 25.5% overall in 2019 to 37.6% in 2020. This may indicate that many teachers had a natural intuition for how confident they could be in their own estimates and applied a logical level of benefit of the doubt to help ensure that students were not disadvantaged relative to a normal year. Of course, the similarity in results may be purely coincidental. However, it does illustrate how major changes in grade distributions between years need not necessarily in themselves indicate inappropriate decisions. If our aim is to protect students from any adverse effects of added unreliability, a change in grade distributions is a logically justifiable result.

Table 2: Distribution of Centre Assessed Grades (CAGs) in summer 2020 (reproduced from Ofqual, 2020, Table 9.6, p.134)

	Cumulative % at each grade					
	A*	A	B	C	D	E
Centre Assessed Grades	13.9%	37.7%	64.9%	87.0%	96.4%	99.7%

Table 3 shows the overall misclassification rate of the No Added Disadvantage strategy across different levels of assessment reliability. Note that these misclassification rates are estimated for each grade separately. For example, the 7.1 per cent misclassification rate for grade A at a reliability of 0.925 means that we would expect this percentage of people to either have a true grade of A or above and be awarded a lower grade, or to have a true grade below A but to be awarded A or above. As one would expect, the lower the reliability of the assessment, the greater the percentage of students that are misclassified. The comparison of these misclassification rates to those we would get from a strategy of retaining the same grade distribution regardless of reliability (Table 4) is of more interest. As can be seen, for small reductions in reliability there is little difference between the overall accuracies of the two approaches. However, at grade A*, for the lowest reliabilities, the No Added Disadvantage approach does lead to substantially more misclassifications overall than the alternative strategy. Having said this, for grades C to E it actually leads to slightly fewer misclassifications than retaining the grade distribution (for an explanation see the earlier discussion surrounding Figure 5).

Table 3: Overall misclassification rates at each grade for the No Added Disadvantage strategy to setting grade boundaries for different levels of reliability.

Reliability	Misclassification rate at each grade					
	A*	A	B	C	D	E
0.925	3.2%	7.1%	8.8%	6.9%	3.6%	1.2%
0.900	3.9%	8.4%	10.3%	8.0%	4.1%	1.4%
0.850	5.3%	11.0%	13.1%	9.8%	4.9%	1.6%
0.800	6.8%	13.6%	15.6%	11.4%	5.6%	1.9%
0.750	8.5%	16.2%	18.0%	12.9%	6.2%	2.0%
0.720	9.6%	17.8%	19.4%	13.7%	6.5%	2.1%
0.700	10.4%	18.9%	20.4%	14.2%	6.7%	2.2%

Table 4: Overall misclassification rates at each grade if the same grade distributions are retained across different levels of reliability.

Reliability	Misclassification rate at each grade					
	A*	A	B	C	D	E
0.925	3.2%	7.1%	8.8%	6.9%	3.6%	1.2%
0.900	3.7%	8.2%	10.2%	8.0%	4.1%	1.4%
0.850	4.6%	10.2%	12.6%	9.9%	5.1%	1.7%
0.800	5.3%	11.8%	14.7%	11.5%	5.9%	2.0%
0.750	6.0%	13.3%	16.7%	13.0%	6.6%	2.3%
0.720	6.3%	14.2%	17.7%	13.8%	7.1%	2.4%
0.700	6.6%	14.8%	18.4%	14.3%	7.3%	2.5%

Maximising accuracy

Table 5 shows the cumulative percentage of students that would be awarded each grade if we adopted the strategy of maximising classification accuracy at each level of reliability. As can be seen, as the reliability of the assessment lowers, this strategy results in awarding fewer top grades (A and A*) but being more generous at lower grades (C, D and E). The strategy has little effect on the cumulative percentage awarded grade B or above. At the lowest grades (D and E) this strategy is even more generous than the No Added Disadvantage strategy (Table 1). However, comparison with Table 2 shows that the strategy would still not be as generous as CAGs awarded in summer 2020 at these grades.

Table 5: Percentage of students awarded each grade or above for different levels of reliability under the Maximise Accuracy strategy to setting grade boundaries.

Reliability	Cumulative % at each grade					
	A*	A	B	C	D	E
0.925	7.0%	24.7%	51.7%	76.7%	91.8%	98.0%
0.900	6.7%	24.4%	51.7%	77.0%	92.1%	98.1%
0.850	6.2%	23.7%	51.7%	77.6%	92.7%	98.4%
0.800	5.6%	23.1%	51.8%	78.3%	93.3%	98.6%
0.750	5.1%	22.3%	51.8%	79.0%	93.9%	98.9%
0.720	4.7%	21.9%	51.9%	79.5%	94.3%	99.0%
0.700	4.5%	21.6%	51.9%	79.9%	94.5%	99.1%

Table 6 shows the misclassification rates that would result from this strategy. Although (by design) these values are always lower than for the other two strategies we have explored (Tables 3 and 4), the difference with the strategy of retaining the same grade distribution is always very small. Specifically, for the range of reliabilities displayed here, the difference in misclassification rate is always within 1 percentage point.

Table 6: Overall misclassification rates at each grade for the Maximise Accuracy strategy to setting grade boundaries for different levels of reliability.

Reliability	Misclassification rate at each grade					
	A*	A	B	C	D	E
0.925	3.1%	7.1%	8.8%	6.9%	3.5%	1.2%
0.900	3.6%	8.2%	10.2%	8.0%	4.0%	1.4%
0.850	4.4%	10.1%	12.6%	9.8%	4.9%	1.6%
0.800	5.0%	11.7%	14.7%	11.3%	5.6%	1.8%
0.750	5.5%	13.1%	16.7%	12.7%	6.2%	2.0%
0.720	5.8%	13.9%	17.7%	13.5%	6.5%	2.0%
0.700	6.0%	14.4%	18.4%	14.0%	6.7%	2.1%

Conclusion

This article has considered how we might logically go about applying generosity to grade boundaries to address a situation where, for whatever reason, an assessment is less reliable than normal. We have seen that a rigid adherence to the idea that no-one should end up with a lower grade than they deserve is not practical. In fact, with any level of unreliability, and a sufficiently large number of candidates, such a goal could only be achieved by awarding the highest possible grade to every candidate. However, once we acknowledge that, even under normal circumstances, no assessment is perfectly reliable, we can make progress with this issue. In particular, we can determine the most appropriate grade distribution given an idea about how much reliability has decreased and how we wish to balance the competing desires to control aggregate outcomes and to manage the level of risk to individuals.

One strategy we might adopt is to maximise the overall accuracy of awarded grades. This will lead to different results than simply retaining the same grade distribution regardless of how far reliability has fallen. Specifically, it will make us more reluctant to award the very highest grades as it is difficult for an unreliable assessment method to verify that these are warranted. Conversely, under this same strategy, we would be more generous with awarding grades at the lower end as it is difficult for an unreliable assessment to verify that a candidate *does not* deserve to be awarded a grade of at least this level.

Another option is the strategy we have termed No Added Disadvantage. Under this option, which places the most emphasis on managing the risk to individuals, we set grade boundaries to ensure that the overall proportion of students awarded a lower grade than merited by their true ability does not increase as the reliability of assessment falls. This option requires generous awarding across all grades with the required level of generosity increasing to compensate for lower assessment reliabilities.

One interesting finding from this research comes from comparing the results from this strategy to the actual distribution of CAGs supplied by schools in summer 2020. By using data from forecast grades in the past, and by noting that CAGs were a form of forecast, we generated a data-driven estimate of the likely reliability of CAGs. Based on this, we found that, in part, the grade distribution in summer 2020 represented a perfectly reasonable application of benefit of the doubt by teachers to ensure that students were not negatively affected by the circumstances they found themselves in. The exception to this was at the lowest A Level grades where, based on our approach to calculations, teachers appeared even more generous than would be recommended by the principle of No Added Disadvantage.

The above paragraph has some implications for awarding once assessment returns to normal. Some would argue that the generous distribution of grades from 2020 should be carried forward into the future as not doing this is unfair to subsequent cohorts of students. However, this fails to recognise the possible role of benefit of the doubt in teachers assigning grades in 2020. As we have shown, at least at the higher grades, the 2020 grade distribution represents a perfectly sensible operation of benefit of doubt given how hard it is to know how students will perform in a future set of exams. However, if the same generous grade distribution is applied in future assessment series where there is far less doubt over performance, any intended benefit of the doubt by 2020 teachers is overridden. One way to think of this issue is as follows. It is possible that some teachers in

2020 awarded grades to students based on the highest grade they think they might have reasonably achieved. This helped protect individual students from being under-rewarded. However, had these same teachers known that the performance standards required for each grade were going to be lowered, they could have extended this same principle of benefit of the doubt further to other students—but did not. Lowering the required performance standards could re-introduce the possibility of students missing out on the grade they deserve relative to future cohorts of candidates due to the unreliability of the assessment procedure they were subject to in 2020.

To put this another way, where there is more doubt there is more benefit of the doubt. Applying this principle in a logical way means that more top grades should be awarded in a year of unusually unreliable assessment than in a normal year. Arguing that the grade distribution from 2020 should simply be carried forwards on grounds of fairness places too much emphasis on national statistics and not enough on individual students—something that assessment organisations are often accused of!

More generally, this article has discussed issues relating to awarding in the knowledge that assessments are not perfectly reliable. Any discussion of classification accuracy in qualifications is uncomfortable as it opens the door to sound and fury about "students being awarded the wrong grade". However, in reality, classification accuracy is just putting a quantitative value on an experience most exam takers are probably familiar with. Many people can probably remember hoping that certain topics or types of problem would be included in their exams and then being either pleased when they were or disappointed when they were not. Reliability research just attempts to put a number on the extent to which these types of familiar experiences have an impact upon grades. Similarly, the attempt in this article to quantify the reliability of teacher forecasts is not an attack on teachers—simply an effort to quantify how hard it is to know what grade a student will achieve months before they take an exam.

More importantly, it is only by acknowledging and quantifying likely levels of reliability that we can make specific recommendations around issues such as benefit of the doubt. The research presented in this article would not have been possible if we had tried to start from a point of view that in a normal year assessment is perfectly reliable.

Within any crisis, grade boundaries are an obvious element of assessment for people to focus on as, from a practical perspective, altering them is relatively easy. Nonetheless, it would clearly be naïve to think that all the problems created by a pandemic (or another unforeseen circumstance) can be solved simply by changing grade boundaries. After all, altering grade boundaries does not change anything about the fundamental measurements that have been made about students—it simply alters the way these measurements are labelled. With this in mind, although this article has focused upon grade boundaries, this is not intended to diminish the role of other elements of the assessment system in dealing with a crisis. Arrangements for appeals, additional assessment opportunities, advice to users of results (such as universities) and various other interventions can all be important parts of the ways in which we can alleviate the risk of individual students being disadvantaged.

References

Benton, T. (2014). Calculating the reliability of complex qualifications. *Research Matters: A Cambridge Assessment publication, 18,* 48–52. https://www.cambridgeassessment.org.uk/Images/465814-calculating-the-reliability-of-complex-qualifications.pdf

Benton, T. (2016). *Evidence for the reliability of coursework.* Paper presented at the 17th annual AEA Europe conference, Limassol, Cyprus, 3–5 November 2016. https://www.cambridgeassessment.org.uk/Images/346753-evidence-for-the-reliability-of-coursework.pdf

Bramley, T., & Dhawan, V. (2010). *Estimates of reliability of qualifications.* Ofqual/11/4826. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/578868/2011-03-16-estimates-of-reliability-of-qualifications.pdf

Gill, T., & Benton, T. (2015). *The accuracy of forecast grades for OCR A levels in June 2014.* Cambridge Assessment Statistics Report Series No.90. https://www.cambridgeassessment.org.uk/Images/241261-the-accuracy-of-forecast-grades-for-ocr-a-levels-in-june-2014.pdf

Hayes, M., & Pritchard, J. (2013). *Estimation of internal reliability.* Ofqual/13/5258. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/378056/2013-01-17-estimation-of-internal-reliability.pdf

He, Q. (2009). *Estimating the reliability of composite scores.* Ofqual/10/4703. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/578878/2010-02-01-composite-reliability.pdf.

JCQ. (2019). *GCE A Level & GCE AS Level Results Summer 2019*. https://www.jcq.org.uk/wp-content/uploads/2019/08/A-Level-and-AS-Results-Summer-2019.pdf

Mc Manus, I. C., Woolf, K., Harrison, D., Tiffin, P.A., Paton, L. W., Cheung, K. Y. F., & Smith, D. (2020). *Calculated grades, predicted grades, forecasted grades and actual A-level grades: Reliability, correlations and predictive validity in medical school applicants, undergraduates, and postgraduates in a time of COVID-19.* https://www.medrxiv.org/content/10.1101/2020.06.02.20116830v1.full.pdf

Ofqual. (2020). *Awarding GCSE, AS, A level, advanced extension awards and extended project qualifications in summer 2020: Interim report.* Ofqual/20/6656/1. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/909368/6656-1_Awarding_GCSE__AS__A_level__advanced_extension_awards_and_extended_project_qualifications_in_summer_2020_-_interim_report.pdf

Shaw, S. D., & Nisbet, I. (2021). Attitudes to fair assessment in the light of COVID-19. *Research Matters: A Cambridge Assessment Publication, 31,* 6–21.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15,* 72–101.

Stewart, H., Proctor, K., & Weale, S. (2020, August 17). Gavin Williamson seeks to blame Ofqual for exams debacle. *The Guardian.* https://www.theguardian.com/education/2020/aug/17/gavin-williamson-seeks-blame-ofqual-exams-debacle-a-level-gcse

Wheadon, C., & Stockford, I. (2010). *Classification accuracy and consistency in GCSE and A Level examinations offered by the Assessment and Qualifications Alliance (AQA) November 2008 To June 2009.* Ofqual/11/4823. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/578862/2011-03-16-aqa-classification-accuracy-and-consistency-in-gcse-and-a-levels.pdf.

Appendix: R code to create Figures 3 and 4

The code below shows how to create charts like those shown in Figures 3 and 4. The calculations in the remainder of the report were done in a similar way.

```
#code to create Figures 3 and 4
library(mvtnorm)
library(ggplot2)
#set main pass rate of interest (based on national A* percentage 2019)
pass_rate=7.8
#convert to boundary on standard normal distribution
boundary_main=qnorm(1-pass_rate/100)
#FIGURE 3
#set reliability of interest
reliability=0.925
#calculate error rates for a variety of boundaries
trial_pass_rates=seq(1,20,0.01)
trial_boundaries=qnorm(1-trial_pass_rates/100)
#correlation matrix between true and observed scores
sigma <- cbind(c(1,sqrt(reliability)),c(sqrt(reliability),1))
#overall percentage awarded a grade too low given their true score
error_rate_low=sapply(trial_boundaries
                      ,function(i) 100*pmvnorm(mean=c(0,0)
                                  ,sigma
                                  ,lower=c(boundary_main,-Inf)
                                  ,upper=c(Inf,i)))
#overall percentage awarded a grade too low given their true score
error_rate_high=sapply(trial_boundaries
                      ,function(i) 100*pmvnorm(mean=c(0,0)
                                  ,sigma
                                  ,lower=c(-Inf,i)
                                  ,upper=c(boundary_main,Inf)))
error_rate_tot=error_rate_low+error_rate_high
#identify target under-graded for Figure 4 (later)
targ_low=error_rate_low[trial_boundaries==boundary_main]
#make chart
min_error_x=trial_pass_rates[which.min(error_rate_tot)]
chartdat=data.frame(x=rep(trial_pass_rates,3)
                      ,y=c(error_rate_low,error_rate_high,error_rate_tot)
                      ,type=c(rep("Too low",1901)
                                  ,rep("Too high",1901)
```

```
                                        ,rep("Overall",1901)))
ggplot(data=chartdat,aes(x=x,y=y,col=type))+geom_line()+
        geom_hline(yintercept=targ_low,lty=2,alpha=0.5)+
        theme_bw()+scale_x_continuous(breaks=seq(0,20,2))+
        labs(x="Percentage awarded A*",y="Misclassification rate (%)"
                        ,col="Type of\nmisclassification")+
        theme(text=element_text(size=14))+
        geom_vline(xintercept=min_error_x,col="orange",size=1.5
,alpha=0.8)+
        geom_vline(xintercept=pass_rate,col="purple",size=1.5,alpha=0.8)+
        ggtitle("Reliability=0.925")
#FIGURE 4
#set reliability of interest
reliability=0.72
#correlation matrix between true and observed scores
sigma <- cbind(c(1,sqrt(reliability)),c(sqrt(reliability),1))
#overall percentage awarded a grade too low given their true score
error_rate_low=sapply(trial_boundaries
                        ,function(i) 100*pmvnorm(mean=c(0,0)
                                ,sigma
                                ,lower=c(boundary_main,-Inf)
                                ,upper=c(Inf,i)))
#overall percentage awarded a grade too low given their true score
error_rate_high=sapply(trial_boundaries
                        ,function(i) 100*pmvnorm(mean=c(0,0)
                                ,sigma
                                ,lower=c(-Inf,i)
                                ,upper=c(boundary_main,Inf)))
error_rate_tot=error_rate_low+error_rate_high
#make chart
min_error_x=trial_pass_rates[which.min(error_rate_tot)]
#find pass rate that gives closest to previous under-grading rate
NAD_x=trial_pass_rates[which.min(abs(error_rate_low-targ_low))]
chartdat=data.frame(x=rep(trial_pass_rates,3)
                        ,y=c(error_rate_low,error_rate_high,error_rate_tot)
                        ,type=c(rep("Too low",1901)
                                ,rep("Too high",1901)
                                ,rep("Overall",1901)))
ggplot(data=chartdat,aes(x=x,y=y,col=type))+geom_line()+
        geom_hline(yintercept=targ_low,lty=2,alpha=0.5)+
        theme_bw()+scale_x_continuous(breaks=seq(0,20,2))+
        labs(x="Percentage awarded A*",y="Misclassification rate (%)"
                        ,col="Type of\nmisclassification")+
        theme(text=element_text(size=14))+
        geom_vline(xintercept=min_error_x,col="orange"
,size=1.5,alpha=0.8)+
        geom_vline(xintercept=pass_rate,col="pink",size=1.5,alpha=0.8)+
        geom_vline(xintercept=NAD_x,col="purple",size=1.5,alpha=0.8)+
        ggtitle("Reliability=0.72"
```