How long should a high stakes test be?

Tom Benton (Research Division)

Introduction

Educational assessment is used throughout the world for a range of different formative and summative purposes. Wherever an assessment is developed, whether by a teacher creating a quiz for their class, or by a testing company creating a high stakes assessment, it is necessary to decide how long the test should be. Specifically, how many questions should be included and how much time will be required to answer each of them.

The aim of this article is to review some of the most relevant psychometric literature on this topic and show the range of test lengths that would be implied in practice by the various recommendations.

As a counterbalance to this technical work, we also explore the lengths of high stakes assessments across different countries to see how much variation there is. Using international comparisons in this way acts as "a mirror, not as a blueprint" (White, 1987, as cited in Clarke, 2004). What is meant by this is that the lengths of assessments in other countries do not necessarily provide a pattern we should copy. However, by including comparisons to assessment practice in other nations, this research is prevented from becoming purely an exercise in self-justification and we are forced to reflect upon why different countries may come to different conclusions about how long high stakes tests should be.

Before beginning it is worth being clear that, obviously, the answer to the question of how long a test should be will depend upon a range of factors such as its purpose and the breadth of learning it is attempting to assess. Furthermore, the decision requires balancing the costs of long assessments and the impact on the experience of test takers against the likely benefits of increased accuracy. Ultimately such decisions are a matter of educational policy rather than something where a single recommendation can be derived mathematically. Nonetheless, this article attempts to provide practical advice from the perspective of psychometric reliability for considering how long a test should be.

The role of precedent

To start with, it is worth mentioning probably the most influential factor in setting test lengths – the role of precedent.

If a new qualification is intended to be comparable to an existing one, then it would be odd for them to require very different assessment lengths. For example, employers may be sceptical that a qualification requiring only half an hour of assessment will provide the same level of accuracy as one that needed four hours. Conversely, test takers may be upset to be told that the amount of time they need to spend taking exams has been doubled compared to previous years – that is, other test takers have been allowed to achieve the same level of benefit in less time. As such, decisions regarding test length are always likely to build upon what has been done for similar qualifications historically.

Following precedent can also be justified from a technical standpoint. If two qualifications are supposed to be used interchangeably, then it is reasonable to expect that they will measure performance equally accurately. Thus, unless one qualification can achieve high reliability in another way (e.g., adaptive testing), they should be of similar lengths. If reliability differs between two assessments this can have implications for equity. In very broad terms, a short and less reliable assessment will favour less able students as they have an increased chance of overperforming due to good luck. On the other hand, a longer and more reliable test will favour the most able students as it will give them the best chance to demonstrate their skills.

Recommended minimum levels of reliability

Aside from precedent, one way to determine test length is to say that a test should be long enough to meet certain minimum requirements in terms of reliability. Reliability refers to the extent to which we would expect test takers to get the same results were we to replicate the assessment process (Brennan, 2001). For example, the (hypothetical) replication we are interested in might consist of repeating the assessment using different test questions. We would hope that candidates' scores would not change too dramatically if this were done.

Table 1 provides a range of recommended minimum reliability levels for high stakes assessment that can be found in the academic literature. For each of the target reliability values, the second column provides details of at least one of the authors that have suggested it as a minimum. The final column provides some further notes on the language used in relation to this target.

Table 1: A range of minimum reliability levels for high stakes assessment suggested in the academic literature

Target reliability value	Source	Further notes
0.80	Evers (2001), Fry et al. (2012)	"Sufficient", "Typical target"
0.85	Cresswell and Winkley (2012), Frisbie (1988)	"Minimum"
0.90	Evers (2001), Fry et al. (2012), Nunnally (1978) (as cited in Drost, 2011)	"good" or "appropriate" for larger MCQ tests
0.92	Skurnik and Nuttal (1968) and others	Derived from aim that 95 per cent of pupils are accurately classified to within 1 grade. See later discussion in text.
0.95	Kubiszyn and Borich (1993) (as cited in Wright, 1996)	For an "acceptable standardized test"

In interpreting Table 1, it is crucial to note that every author providing these recommendations is clear that reliability will not simply depend upon the characteristics of the test (e.g., its length) but will also be influenced by other factors. To take one example, the quality of the administration conditions may affect the size of reliability coefficients (see Traub & Rowley, 1991, or Frisbie, 1988). Similarly, the authors do not pretend that their suggestions are underpinned by a fully logical argument such as balancing the costs of unreliability against the costs of longer tests. Rather, they simply represent benchmarks based upon the kind of values that have typically been achieved by test developers ever since the easy calculation of reliability indices has been possible.

The target reliability values in Table 1 assume that we are using classical reliability coefficients such as (but not necessarily limited to) Cronbach's alpha (Cronbach, 1951). Such indices of reliability use data on the correlations between scores on items within a test to infer the likely correlation between candidates' observed scores on the test and their scores on another (hypothetical) parallel test.[1] Note that reliability measures of this type are highly dependent upon the ability distribution of the candidates taking them. In particular, they will tend to yield low values in instances where all the students taking a test happen to have very similar levels of ability. To address this concern, the recommendations in Table 1 should be seen as assuming that the range of candidates entering an assessment are broadly representative of the wider population the exam is aimed at. For example, for recommendations to be applicable to a specific GCSE, it should be taken by a similar range of candidates as typically enter GCSEs.

1 A parallel test can be thought of as a test that measures the same constructs as the one being studied, and is equally hard and equally long as the test in question. For example, if two tests fit the Rasch model, they will be parallel if they have identical distributions of item difficulties.

One of the recommended minimum reliability values in Table 1 is 0.92. This is derived from recommendations in the literature relating to classification accuracy. Classification accuracy estimates the percentage of candidates whose grade matches the grade they should be awarded based on their (notional) true score. Their true score is the (hypothetical) score they would achieve on average across many tests parallel to the one they have taken. Classification accuracy is rarely used directly to determine minimum levels of reliability. The reason for this is that, as noted by Wheadon and Stockford (2010), "unless an examination is perfectly reliable, some of those who lie to just one side of a grade will have true scores that fall the other side of it. As a consequence, no examination system can have an accuracy of better than plus or minus one grade" (p. 5). With this in mind, several authors have turned their attention to ensuring that a high percentage of candidates are correctly classified to within plus or minus one grade. Skurnik and Nuttal (1968) suggested a target of ensuring that at least 95 per cent of pupils are accurately classified to within 1 grade. Wheadon and Stockford (2010) agreed that, while this target is essentially arbitrary, it seems a useful point of reference. A similar target (based upon classification consistency) was suggested by Mitchelmore (1981). To convert this suggested target into an equivalent value of classical reliability we have assumed that we are working with the current GCSE grade scale (see footnote for calculation steps[2]).

In summary, Table 1 suggests that, depending upon which author we rely on, the minimum reliability of a test is somewhere between 0.80 and 0.95. Notice that, based on the Spearman-Brown formula (given later) and all else being equal, a test with a reliability of 0.95 will be almost five times as long as one with a reliability of 0.8. Thus, while the exact choice of a target value for reliability may appear to be arguing over tiny details, when it comes to using this to determine test length, a small change can make a big difference.

Having identified a set of recommended minimum reliability levels from the literature, the next step is to estimate how long tests should be to meet these criteria. The steps for this calculation are the subject of the next section.

2 Specifically, from published statistics (GCSE (Full Course) Outcomes for main grade set for each jurisdiction) regarding GCSEs taken in England we know that in summer 2019, 4.5 per cent of candidates achieved grade 9. This implies, if scores were normally distributed, then the grade 9 boundary would be about 1.7 standard deviations above the mean. The same statistics reveal that 98.3 per cent of candidates achieved grade 1 or above meaning that the grade 1 boundary would be 2.1 standard deviations below the mean (if scores were normally distributed). Taken together this means that the eight grade bandwidths (between 1 and 9) would be spread out across 3.8 standard deviations, which in turn implies that the grade bandwidth will be 0.475 standard deviations. For a worst-case scenario of a candidate with a true score directly on a grade boundary, their observed grade will differ from their true grade by more than one if their observed score is too high by *two* grade bandwidths or if it is too low by a single grade bandwidth. This will happen at least 5 per cent of the time if the standard error of measurement is more than 0.28 standard deviations. This indicates a reliability of 0.92 (=1-0.28^2).

Calculating required test lengths

Psychometric formulae

One of the earliest suggested methods for predicting the reliability of a test from its length might be the Spearman-Brown formula (Spearman, 1910; Brown, 1910). This allows us to predict the impact on reliability of lengthening or shortening a test. The Spearman-Brown formula is:

$$\alpha_{comp} = \frac{k\alpha_0}{1 + (k-1)\alpha_0} \quad (1)$$

where α_{comp} is the predicted reliability of a new exam component, α_0 is the known reliability of a reference component, and k is the length of the new exam component relative to the reference one. For example, if we were interested in calculating the likely reliability after doubling the length of a test, k would be set equal to 2.

Similar formulae can be derived starting from an approach to measurement based upon the Rasch partial credit model (Linacre, 2000) so that, under reasonable assumptions, the formula can relate to the total available score in a test and not just the number of items. Other research provides methods to extend the calculations to more complex scenarios such as when combining scores from multiple different assessments potentially measuring different constructs (He, 2009; Wang & Stanley, 1970). In particular, to calculate the reliability of a qualification built from multiple components, all of equal length, and where the separate dimensions of ability they measure are all equally correlated with one another, we can use the following simplification of the Wang-Stanley formula (Wang & Stanley, 1970).

$$\alpha_{qual} = \frac{\alpha_{comp} + (n-1)\rho\alpha_{comp}}{1 + (n-1)\rho\alpha_{comp}} \quad (2)$$

where α_{qual} is the predicted reliability of a new qualification, n denotes the number of components comprising the qualification, and ρ the correlation between true scores in the separate dimensions of ability measured by different components. Note that the formula assumes that all components are equally weighted and that the overall qualification score is obtained simply by adding up all the scores on the components.

The two formulae above can be combined to give:

$$\alpha_{qual} = \frac{k\alpha_0 + (n-1)\rho k\alpha_0}{1 + (k-1)\alpha_0 + (n-1)\rho k\alpha_0} = \frac{k\alpha_0(1 + (n-1)\rho)}{k\alpha_0(1 + (n-1)\rho) + (1 - \alpha_0)} \quad (3)$$

If we want to find the required test length for each qualification component (relative to a known reference component) for a target level of reliability, equation 3 can be rearranged to:

$$k = \left(\frac{\alpha_{qual}}{1-\alpha_{qual}}\right)\left(\frac{1-\alpha_0}{\alpha_0}\right)\left(\frac{1}{1+(n-1)\rho}\right) \quad (4)$$

Finally, we note that our main interest is in the overall length of assessments across the qualification as a whole rather than the length of individual components. That is, we want our formula to suggest values for nk rather than simply k. Putting all this information together yields the following formula for the number of marks to include in a qualification (relative to a reference component with reliability α_0) to achieve a reliability of α_{qual}.

$$nk = \left(\frac{\alpha_{qual}}{1-\alpha_{qual}}\right)\left(\frac{1-\alpha_0}{\alpha_0}\right)\left(\frac{n}{n\rho+(1-\rho)}\right) \quad (5)$$

In order to make use of the above formula, we need values for α_0 and ρ. Ideally, we would like to discuss test length in units of time (e.g., minutes) rather than in terms of the number of available marks. For this reason, we also need to know how many minutes are typically allowed for each available mark in an exam. All of these matters are discussed next.

Reliability of reference component

First, we attempt to identify a suitable value for α_0. This can be done by looking at empirical data on test reliability historically.

By far the largest amount of published data on test reliability is in the form of Cronbach's alpha. This type of data provides a natural starting point for calculations. For example, Bramley and Dhawan (2010) published a wealth of information on the reliability of OCR examinations such as a chart showing how Cronbach's alpha increases along with the number of marks in a test (see their Figure 1.4). A similar chart, based on all OCR GCSE and AS/A Level components (that is, individual examination papers) taken by at least 500 candidates across the five years from the start of 2015 until the end of 2019, is shown in Figure 1.[3] This chart summarises the reliability coefficients associated with almost 1600 assessments. Assessments are grouped by rounding the number of available marks to the nearest 10, and the distribution of reliabilities within each group is shown in the form of a boxplot. The largest number of assessments (more than 300) had a maximum mark of 60. As can be seen, for this maximum mark band, the reliability coefficients were just above 0.8 on average. Slightly fewer assessments (but still more than 200) had a maximum mark of 50. The average reliability for these assessments was very close to 0.8. Very few assessments had maximum marks below 50 and so these elements of the chart can be ignored.

3 Figure 1 also includes reliability estimates for papers with optional questions. In these cases, Backhouse's formula P (Backhouse, 1972) is used as a substitute for Cronbach's alpha.

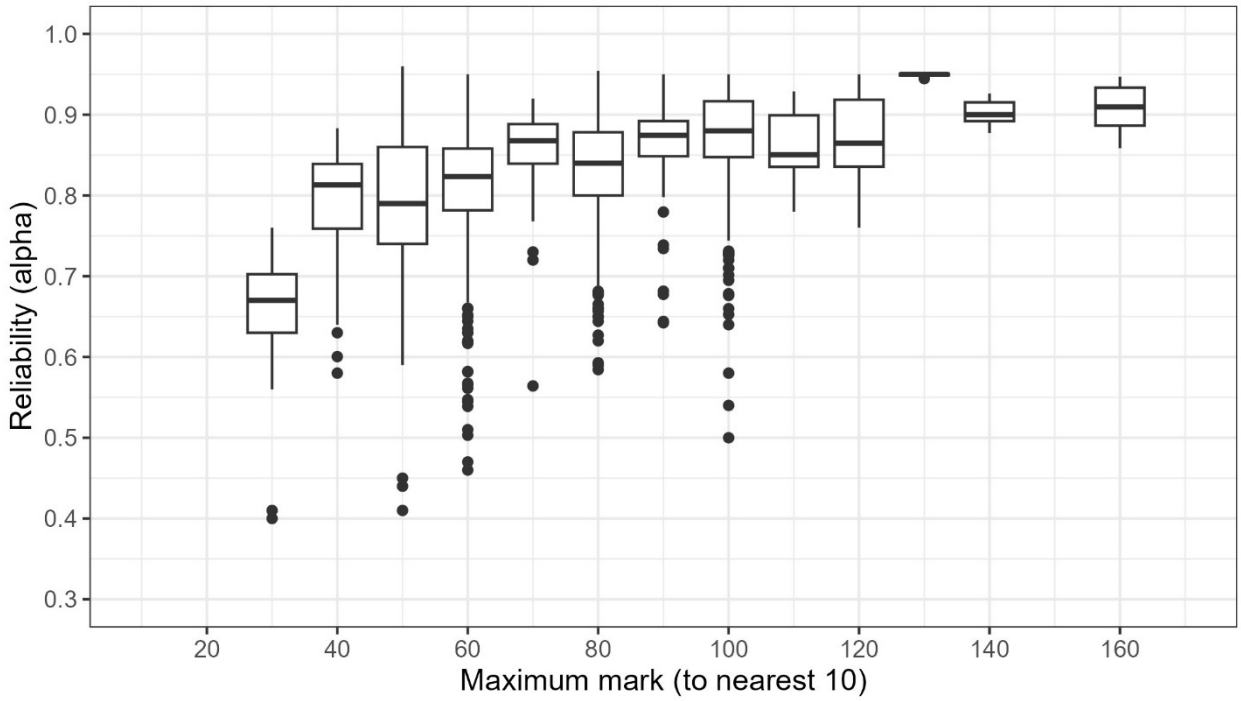

Figure 1: Relationship between test reliability and total test mark for all OCR GCSE and AS/A Level examinations entered by at least 500 candidates between 2015 and 2019

If we assume that an average test with 50 marks has a reliability of 0.8, then according to the Spearman-Brown formula, tests with maxima of 60, 70, 80 and 90 should have reliabilities of 0.83, 0.85, 0.86 and 0.88. Broadly speaking, this fits with the average reliabilities we can see in Figure 1.

However, if we continue to use the Spearman-Brown formula, we would expect tests with maxima of 100, 120 and 140 to have reliabilities of 0.89, 0.91 and 0.92 respectively. These expectations are not reflected by the data in Figure 1. This is likely to be because, as mentioned above, reliability coefficients depend upon a range of factors that may be associated with test length and not just test length itself. In particular, in our data, longer tests are more likely to be part of an A Level, and shorter ones more likely to be part of a GCSE. A Levels tend to have slightly lower reliability coefficients for the same test length (perhaps due to the more restricted range of candidates involved). For example, among 60 mark tests the median reliability of a GCSE component is 0.83 whereas for an A Level it is 0.81.

Despite the differences between qualifications, it seems reasonable to use the starting point of 0.8 for a 50-mark test because calculations of test length should evaluate how reliability changes with test length within a fixed group of candidates.

Compared to some published statistics of test reliability, a starting point of 0.8 for a 50-mark test may seem disappointingly low. For example, recent published statistics for Key Stage 1 national curriculum tests in English suggest that these 40-mark reading tests for 7-year-olds achieve a reliability of about 0.95.[4] However,

4 See Tab 28 of National Curriculum Test Handbook: 2016 and 2017 technical appendix.

the apparently high reliability in that context may be because of the very large diversity in reading ability among children of that age. As such, it may not be something we would expect to repeat at GCSE. This is partially confirmed by the fact that the same set of published statistics suggest that the reliability of 50-mark national curriculum reading tests for 11-year-olds (Key Stage 2) is lower at 0.89.

Note that, in using this starting point we are not assuming that every 50-mark test will always yield an alpha coefficient of 0.8. The exact values of reliability coefficients are dependent upon numerous factors. In particular, the range of abilities of the candidates taking the test will have a big influence. However, this factor is largely beyond the control of the test developer. What we can do is try to create a test with sufficient length such that, assuming it were taken by a set of candidates with a range of abilities typical of those entering a GCSE, we would have a good chance of alpha exceeding some target value. The starting assumption that a 50-mark test will typically have a reliability of 0.8 (under these circumstances) allows us to do exactly that. To put it another way, for the purposes of using our formula we will set α_0 to be 0.8 and assume that this refers to a reference test form with a maximum mark of 50.

Correlation between true scores on different components

In order to apply equation 5, we also need a value for the correlation between true scores on different components (ρ). Such a value can be obtained using information in Benton (2021a) which indicates that the correlation between observed scores on separate components within an A Level is typically 0.64 while the median reliability of the same components is 0.83. Combining this formula with Charles Spearman's 1904 correction for attenuation formula (Spearman, 1987) yields a value of just below 0.8 ($\approx 0.64/\sqrt{0.83*0.83}$) for the estimated correlation of true scores on separate components. We will use this value in our calculations of required test lengths.

Note that performing the same calculations based on GCSEs taken in summer 2019 leads to a somewhat higher value for the correlation (approximately 0.9). However, as we will see, even with a value of 0.8, accounting for qualifications consisting of multiple components (presumably measuring slightly different skills on different occasions) has a fairly limited impact on the amount of assessment time required in total.

Time per mark

Finally, we require a clear understanding of the usual relationship between the maximum available mark on a test and its (usual[5]) duration in minutes.

5 In England, exam candidates with special educational needs, disabilities or temporary injuries can be allowed extra time to complete an examination if they need it. For the purposes of this paper, we focus on the amount of time that is allowed to students without these access arrangements.

The relationship between the number of marks in an examination and duration is shown in Figure 2. Each point in the chart represents an OCR exam component taken between 2015 and 2019. Separate chart panels have been used for GCSEs and AS/A Levels. A small amount of jitter has been added to the points in the chart to allow the distribution of times and total marks to be seen more clearly. The dashed diagonal lines represent lines of equality. A blue regression line, based upon regression through the origin, is also included. Regression through the origin was used as it is consistent with the (sensible) idea that an exam with no marks would be expected to take no time.

Across both qualification types, the number of minutes allowed for an exam is rarely less than the total number of marks and is usually slightly higher. This fits with the idea in assessment folklore of "a mark a minute" – although internet searches suggest this phrase is used far more often as a guide for students about how long they should spend on exam questions rather than for test developers deciding upon exam duration. The gap between test length in marks and duration in minutes is slightly larger for A Levels than for GCSEs.

Based upon the regression lines, in broad terms, the number of minutes allowed for an exam has tended to exceed the number of available marks by about 20 per cent. We will use this figure as a basis to identify the likely duration of tests needed to meet the reliability thresholds listed earlier.

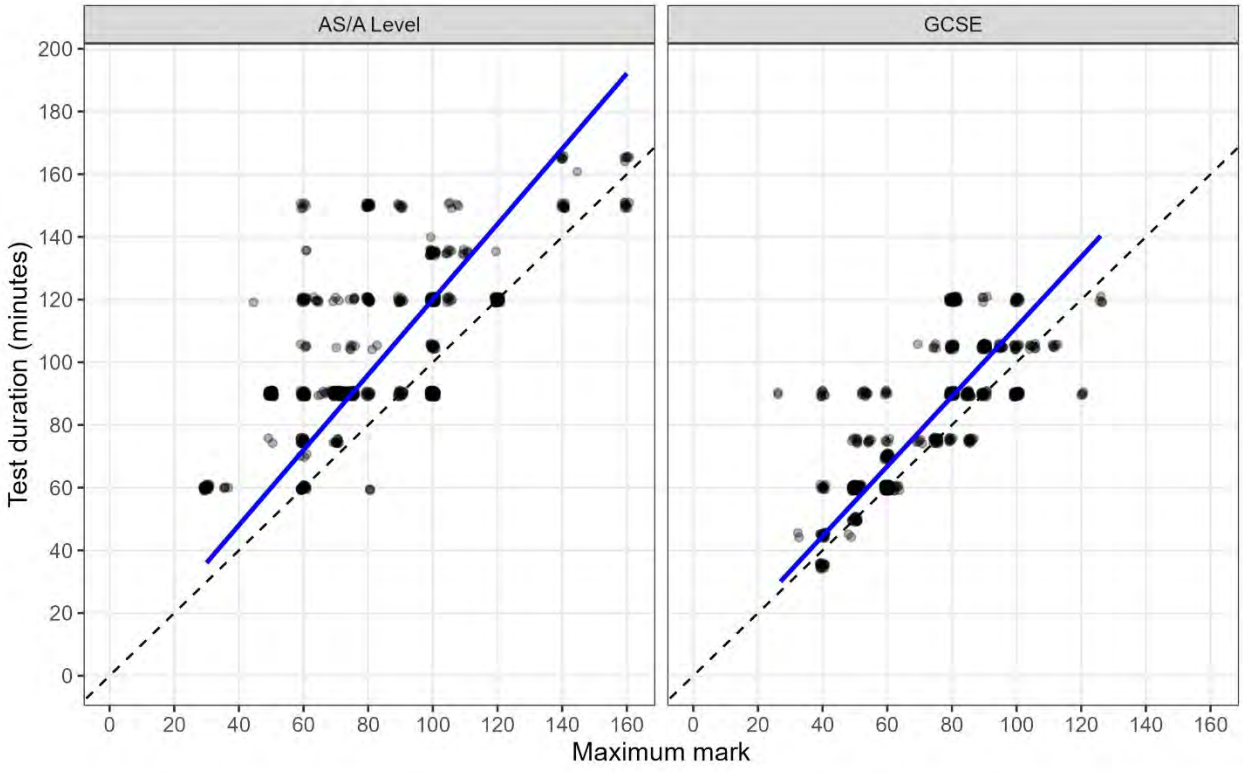

Figure 2: Relationship between test length and test duration for all OCR GCSE and AS/A Level examinations entered by at least 500 candidates between 2015 and 2019. A solid blue regression line (regression through the origin) is included within each chart. The dashed line represents a line of equality between test length in marks and duration in minutes.

Suggested test durations

By combining all of the above assumptions with the formula in equation 5, we can derive the following formula for the total amount of assessment time that is required for a qualification with n components to likely achieve a reliability α_{qual}.

$$Total\ required\ assessment\ time = 15\left(\frac{\alpha_{qual}}{1-\alpha_{qual}}\right)\left(\frac{n}{0.8n+0.2}\right) \quad (6)$$

This formula is derived from equation 5 but replacing both α_0 and ρ with 0.8, multiplying by 50 as our reference reliability comes from a test with 50 marks, and multiplying by 1.2 as we (currently) typically allow 1.2 minutes for each available mark in a test.

Using equation 6, Table 2 shows how the recommended length of tests varies according to the target for reliability we use to determine test length, and the number of components of which the qualification is comprised. The lowest target for reliability considered in this table is the value of 0.80 from Evers (2001). To hit this benchmark our analysis suggests that a high stakes qualification should comprise of at least 50 marks and require about an hour of exam time at a minimum. If the qualification comprises of several components, presumably measuring different skills on different occasions, then the total exam time should increase by perhaps 10 minutes. In other words, spreading measurement across different components has only a minor impact on the total amount of assessment time required to meet reliability requirements.

As expected, as reliability requirements become more stringent, the suggested test lengths increase. Aiming for a reliability coefficient of 0.9 requires a total exam time of between 2 and 3 hours. Aiming for Wheadon and Stockford's (2010) point of reference that qualifications should classify students into the correct grade plus or minus one at least 95 per cent of the time (i.e., a reliability of 0.92) generally requires total examination times in excess of 3 hours. Finally, to achieve the most stringent reliability target we have considered (0.95) will typically require between 5 and 6 hours of assessment.

Table 2: Estimated required total minutes of assessment depending upon target reliability level and the number of components in the assessment

Target reliability	Number of assessment components				
	1	2	3	4	10
0.8	60	67	69	71	73
0.85	85	94	98	100	104
0.9	135	150	156	159	165
0.92	173	192	199	203	210
0.95	285	317	329	335	348

At this point, those involved in the creation and regulation of GCSEs in England may be tempted to congratulate themselves. As it happens, a typical GCSE in England consists of two components (i.e., two separate exam papers) and requires roughly 3 and a half hours of exam time in total. Based on Table 2, this is only slightly higher than the recommended amount of assessment (192 minutes) needed to achieve a qualification reliability of 0.92. From our discussion earlier, this is also roughly the test length required to ensure that, 95 per cent of the time, the grades awarded to candidates are equal to their true grades plus or minus one. However, to avoid this article descending into self-congratulation, and to force us to reflect more deeply on the length of assessment that is actually needed at different stages of education, we next compare the amount of time spent in high stakes examinations in England to that in other countries.

Test lengths in high-performing jurisdictions

Table 3 provides a summary of test durations for qualifications taken in England as well as qualifications/assessments taken in 10 high-performing jurisdictions. The 10 comparator jurisdictions in this article have been chosen from those identified in Elliott et al. (2015) and Suto and Oates (2021). Only assessments that are high stakes for the pupil (leading to a recognised qualification) or are compulsory within their region are included. In addition, the focus is on assessments taken at similar ages to GCSEs and A Levels. For example, although the NAPLAN tests are taken in grade 6 (age 11/12) and grade 9 (age 14/15) in Victoria, the details in the table are based on the grade 9 tests. Note that not all countries identified in Elliott et al. (2015) and Suto and Oates (2021) are included here. This is due to not finding detailed information on the duration of examinations in some countries at the time of undertaking the review for this research in early 2021. Nonetheless, although Table 3 is far from being a comprehensive review of the durations of compulsory and high stakes examinations in high-performing jurisdictions, it hopefully provides a sufficiently wide variety of decisions to facilitate further discussion about test lengths. Links to the websites that were used as a source of information are provided at the end of the article.

As shown in Table 3, and based on qualifications awarded in summer 2019, GCSEs in England require an average of 3.5 hours of exams[6] (typically two exams of an hour and 45 minutes each), whereas A Levels require 6 hours on average[7] (typically three exams of 2 hours). As such, both qualifications are long enough to generally meet some of the highest benchmarks for reliability displayed earlier in Table 2.

Exams at ages 14 to 17

Table 3 allows us to compare the duration of GCSEs to the duration of other exams taken by students of similar ages in education systems around the world.

6 Excluding double science (which counts as two qualifications) and restricting to GCSEs that currently use exams only for assessment.

7 Also restricted to A Levels assessed using exams only. The A Levels requiring the longest exam time are Latin and Classical Greek (7 hours each). All others with these criteria require 6 hours of exam time.

As can be seen, the majority of such assessments require considerably less time per subject than GCSEs. The shortest such assessments are the NAPLAN tests in Australia (Victoria) where the longest assessments (reading and numeracy) take only slightly over an hour each. The relatively short duration of the NAPLAN assessments might be justified by the fact that, although compulsory, the exams are relatively low stakes with the central purpose being to monitor student progress. Similar comments might be made about the assessments used within the Provincial Achievement Testing Program in Canada (Alberta), many of which only require a little over an hour each.

Junior Leaving Certificate exams in the Republic of Ireland have slightly higher stakes as they form part of graduation from secondary school. This may be reflected in the slightly longer required amount of exam time per subject (2 hours). Note that, for these qualifications, marks from exams are supplemented by an additional 10 per cent of marks that are available via school-based assessments. Required exam times for exams taken at ages 14 to 17 in New Zealand (3 hours), Singapore (3.5 hours), and Massachusetts (4 hours) are more similar to those required for GCSEs in England. However, to set this comparison in context we need to consider how many subjects students enter on average. In England, students take nine GCSEs on average (Carroll & Gill, 2018). As such, we expect the average GCSE student in England to spend almost 32 hours taking exams. In contrast, in Singapore the maximum (not the average) number of O Levels a student can take is nine (in the Special and Express stream), and most students will take fewer than this. The maximum number that can be taken in the Normal (Academic) stream is seven. Similarly, according to UCAS, in New Zealand students are typically required to study between five and six subjects for each level of NCEA. In Massachusetts, graduation only requires that students pass exams in three subjects. As such, the total amount of time spent in exam rooms will be substantially lower in these jurisdictions than for students taking GCSEs in England.

The Comprehensive Assessment Programme (CAP) in Chinese Taipei provides an interesting alternative set of arrangements to the GCSE. It is taken at a similar age to GCSEs and is high stakes in that it is a required part of progression to the next stage of education. It relies entirely on external assessment in the form of examinations. However, rather than requiring lengthy separate examinations for different subjects, all subjects are assessed in 7 hours of assessments split across two days. This represents an intense assessment procedure for the student but one that requires far less time than is needed for a student in England to complete all of their GCSEs. In fact, considered as a whole, the CAP actually represents one of the shortest total assessment times of any of the high stakes exams at age 14–17 shown in Table 3. The reasons why shorter assessment time is possible for CAP are not clear. However, it would seem likely that a focus on overall achievement across all subjects rather than a need to have highly reliable assessment for each individual subject may partially explain the difference.

End of secondary and university entrance

From Table 3, the total amount of examination time required for A Levels in England does not seem unusual compared to other countries that focus on individual subjects. For example, both New Zealand (3 hours) and Poland (4 hours) tend to require slightly less examination time per subject but will also typically require students to study larger numbers of subjects (five or six in New Zealand, at least four in Poland). Note that the NCEA in New Zealand also incorporates a substantial amount of internal assessment. The amount of exam time per subject in Canada (Alberta) is the same as A Levels in England. However, it is worth noting that, in contrast to England, despite their length, Diploma Examinations in Alberta only provide 30 per cent of each student's final qualification mark with the remainder dependent upon schools' own assessments.

An interesting contrast to the amount of time required for A Levels is provided by university entrance exams in Japan and South Korea. These exams are extremely high stakes for students as they are the primary means of determining university entrance. However, as a whole they require considerably less exam time than in A Levels in England. Whereas students in England are typically required to spend between 18 and 24 hours taking exams (depending upon the number of A Levels they study), in Japan all assessment is completed in 12 hours (spread over two days) and in South Korea it is completed in 6.5 hours (all on one day). The reduced total assessment time may be because of the very clear single purpose of the exams (university entrance) and the resulting possibility of focusing on results across all subjects combined rather than needing highly reliable results in each individual subject. Of course, the highly compressed timescale for assessment in these countries (one or two days) also has some disadvantages such as the amount of pressure it places on students.

Table 3: Times required for various examinations in England and other high-performing jurisdictions

Country	Assessment name	Target group	Typical exam time required	Additional internal assessment	Number of subjects taken
Australia (Victoria)	National Assessment Program – Literacy and Numeracy Testing (NAPLAN)	Year 9 (Age 14/15)	40–65 minutes per subject	No	4
Canada (Alberta)	Diploma Examination	End of senior high school (university entrance)	Up to 6 hours per subject[8]	Yes. 70% of the final course-mark is derived from internal assessment.	Unclear
Canada (Alberta)	Provincial Achievement Testing Program	Grade 9 (Age 14/15)	1.25 to 3.25 hours per subject[8]	No	4
Chinese Taipei	Comprehensive Assessment Programme for Junior High School Students (CAP)	Year 9 (Age 14/15)	7 hours in total	No	5
England	GCSE	Year 11 (Age 15/16)	3.5 hours per subject	Only in a minority of subjects	9 on average
England	A Level	End of secondary education	6 hours per subject	Only in a minority of subjects	3 or 4
Japan	National Center Test for University Admissions	University entrance	12 hours (approx.) in total	No	6 (if separate sciences counted as one subject each)
New Zealand	National Certificate of Educational Achievement (NCEA)	Year 11 (Age 15/16) through to end of secondary education	3 hours per subject (all levels)	Yes. Internal and external assessments both feed into a credits system.	Typically 5 or 6
Poland	egzamin maturalny ("Matura")	End of secondary education	3 hours per subject	No	At least 4
Republic of Ireland	Junior Certificate	Third year of Junior Cycle (Age 15/16)	2 hours per subject	Yes. 10% of qualification marks from internal assessment.	Possibly 7 or 8 per pupil on average[9]
Singapore	O Levels	Secondary years 4 or 5 (Age 14–17)	3.5 hours per subject	No	Between 4 and 9
South Korea	College Scholastic Ability Test (CSAT)	University entrance	6.5 hours in total	No	7
USA (Massachusetts)	Massachusetts Comprehensive Assessment System (MCAS)	Grade 10 (Age 15/16)	Untimed but recommended time is 2 hours per subject	No	At least 3

8 Intended time for students to complete the test. Since 2017 all students are allowed double this amount if they desire it.
9 Based on dividing the total number of entries to Junior Certificate exams by an estimate of the number of eligible pupils.

Summary and discussion

Reflecting on the results in this paper we can see that, although psychometrics can help us think about how long exams need to be to achieve acceptable reliability, ultimately, the decision is one of policy. The costs of increasing the length of exams in terms of the burden on students, schools and assessors, need to be balanced against the likely benefits of reliable assessment such as public confidence and ensuring that the correct decisions are made about students' futures.

The need for judgement in making this decision can be seen in several ways throughout this article. Firstly, while psychometrics has supplied formulae relating test length to reliability, different authors have made different recommendations regarding what level of reliability is acceptable. Secondly, a brief review of test lengths from different countries around the world reveals a fairly wide variety of approaches in practice.

It is clear that GCSEs and A Levels in England are of sufficient length to likely meet the levels of reliability that are recommended in the academic literature. However, some of the (less stringent) recommendations might also be met by somewhat shorter examinations. Furthermore, comparison with decisions in other countries make it clear that different decisions are possible. This is particularly evident for examinations taken at ages 14–17 where the total exam time for GCSEs in England appears relatively high compared to other countries.

To some extent, differences in length can be explained by differences in purpose. In particular, some of the shortest examination lengths were seen for assessments that are primarily formative in their purpose such as NAPLAN in Australia (Victoria) or the Provincial Achievement Testing Program in Canada (Alberta). Nonetheless, there are also examples of countries such as the Republic of Ireland where high stakes qualifications are awarded based on substantially shorter exams than in England. Furthermore, although O Level exams in Singapore are of similar length to GCSEs in England, students tend to take fewer such exams.

From the analysis of the length of exams in other jurisdictions (e.g., CAP tests in Chinese Taipei; university entrance tests in South Korea and Japan), it seems possible to reduce the total exam length by focusing on overall achievement across all subjects, rather than attempting to provide highly reliable assessment for each one individually.

Decisions about test length require a clear understanding of the purposes of assessment. This would certainly include considering whether an assessment is primarily formative or summative as well as how it may be combined with other information to impact on decisions about students' futures. It might also include a consideration of comparability and ensuring that any new qualification meets broadly the same requirements as existing ones to which it will be compared.

Limitations

The calculations in this report have been derived from looking at average reliabilities across lots of assessments. As such, while they are intended to provide a reasonable guideline to help in determining test lengths, they may be less appropriate in situations where we have good reasons to expect reliability to differ from the typical situation. One example might be where we expect a greater amount of variation between markers. Since, all else being equal, marking error will tend to have some impact on reliability coefficients such as alpha, we may reasonably expect exams consisting of a few essays to have lower reliabilities than suggested by the formulae in this article. As such, we may wish to compensate by increasing the number or length of such exams in a qualification. A more detailed consideration of the relationship between marking error, reliability, validity and recommended test lengths could be the subject of further research.

Recommendations

As is clear from the above discussion, there is no single correct answer to the question of how long a test should be. However, there are perhaps a few general principles that are always worthy of consideration in making this decision. Based on the research described in this article, some potential principles are:

- If the purpose of a test is primarily to provide formative feedback to a student on their progress, a test length of about one hour would be fairly typical of what is required in different countries.
- If an assessment is expected to have a direct impact, on its own, on decisions made about individual students then, for consistency with all but the most permissive psychometric criteria, the test should be at least 90 minutes long. Having said this, there are a few possible justifications for shorter assessments:
  - If they are measuring a very narrow construct. For example, a test of whether primary school children know their times tables, or whether they can read words using phonics, could not reasonably be expected to take longer than half an hour for each student.
  - If computer adaptive testing is used to achieve reliable assessment in a shorter amount of time (but see Benton, 2021b, for a wider discussion of this).
- If the primary focus of assessment is on overall performance across subjects (rather than within each individual subject), as little as one hour per subject may be sufficient to achieve reasonable reliability.
- If an assessment is for students' final qualifications before university, at least 3 hours of exam time per subject is not unusual internationally. Given the high stakes of qualifications taken at this age, this would appear to be a sensible lower bound for test length.
- If a new qualification needs to be directly comparable to an existing one (e.g., for use in school performance tables), it is sensible to ensure that elements of assessment design such as test length are kept reasonably similar.

References

Backhouse, J. K. (1972). Appendix two: Mathematical Derivations of Formulae P, Q, Q', and S. In Nuttall, D. L., & Willmott, A. S. (1972). *British examinations: techniques of analysis*. National Foundation for Educational Research in England and Wales.

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

Benton, T. (2021b). Item response theory, computer adaptive testing and the risk of self-deception. *Research Matters: A Cambridge University Press & Assessment publication*, *32*, 82–100.

Bramley, T., & Dhawan. V. (2010). *Estimates of reliability of qualifications*. Ofqual/11/4826.

Brennan, R. L. (2001). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement*, *38*, 295–317.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, *3,* 296–322.

Carroll, M., & Gill, T. (2018). *Uptake of GCSE subjects 2017*. Statistics Report Series No. 120. Cambridge Assessment.

Clarke, D. (2004). Researching classroom learning and learning classroom research. *The Mathematics Educator*, *14*(2).

Cresswell, M., & Winkley, J. (2012). *Introduction to the concept of reliability*. Chap. 1 in *Ofqual's Reliability Compendium*, edited by D. Opposs and Q. He. Ofqual/12/5117.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.

Drost, E. A. (2011). Validity and reliability in social science research. *Education Research and Perspectives, 38*(1), 105–123.

Elliott, G., Rushton, N., Darlington, E., & Child, S. (2015). Are claims that the GCSE is a white elephant red herrings? Cambridge Assessment Research Report.

Evers, A. (2001). Improving test quality in the Netherlands: Results of 18 years of test ratings. *International Journal of Testing*, *1*(2), 137–153.

Frisbie, D. A. (1988). NCME Instructional Module on reliability of scores from teacher-made tests. *Educational Measurement: Issues and Practice*, *7*(1), 25–35.

Fry, E., Crewe, J., & Wakeford, R. (2012). The Qualified Lawyers Transfer Scheme: innovative assessment methodology and practice in a high-stakes professional exam. *The Law Teacher*, *46*(2), 132–145.

He, Q. (2009). *Estimating the reliability of composite scores*. Ofqual/10/4703.

Kubiszyn, T., & Borich, G. (1993). *Educational testing and measurement*. Harper Collins.

Linacre, J. M. (2000). Predicting reliabilities and separations of different length tests. *Rasch Measurement Transactions*, *14*(3), 767.

Mitchelmore, M. C. (1981). Reporting student achievement: How many grades? *British Journal of Educational Psychology, 51*(2), 218–227.

Nunnally, J. C. (1978). *Psychometric theory*. McGraw-Hill Book Company, pp. 86–113, 190–255.

Skurnik, L. S., & Nuttal, D. L. (1968). Describing the reliability of examinations. *Journal of the Royal Statistical Society. Series D (The Statistician), 18*(2), 119–128.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, *3,* 271–295.

Spearman, C. (1987). The proof and measurement of association between two things. *The American Journal of Psychology, 100*, 441–471.

Suto, I., & Oates, T. (2021). *High-stakes testing after basic secondary education: How and why is it done in high-performing education systems?* Cambridge Assessment Research Report.

Traub, R. E., & Rowley, G. L. (1991). An NCME instructional module on understanding reliability. *Educational Measurement: Issues and Practice*, *10*(1), 37–45.

Wang, M., & Stanley, J. (1970). Differential weighting: A review of methods and empirical studies, *Review of Educational Research*, *40*, 663–705.

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.

White, M. (1987). *The Japanese educational challenge: A commitment to children*. The Free Press.

Wright, B. D. (1996). Reliability and separation. *Rasch Measurement Transactions*, 9(4), 472.