Vocational and Technical Qualifications: Assessment Functioning of external assessments

An overview of the functioning of assessments in 27 qualifications and 49 units

November 2017

Ofqual/17/6319

Authors

This report was written by Beth Black, Qingping He and Stephen Holmes from Ofqual's Strategy Risk and Research directorate.

Contents

1 Executive summary

This report provides an overview of the technical functioning (i.e. test and question [item] functioning) of 49 external tests (or units) from 27 qualifications which are on or going onto Department for Education performance tables. These school-based qualifications were included in this project because they met three criteria – (i) contains external examined assessment (ii) inclusion in Department for Education's 16-19 performance tables for 2016 and (iii) volume of entry likely to be >500 on the basis of 2015 certification data. These 27 qualifications represented predominantly Level 1 and 2 qualifications but also some Level 3, across a range of subjects including health and social care, carpentry, hospitality, digital media, applied science and mathematics. The qualifications were from seven different Awarding Organisations (AOs).

This report provides an overview of test and item (question) functioning in these tests. It does not focus on other aspects of the qualifications such as the detailed consideration of the content in the specification, teaching time or delivery, or indeed the processes by which the assessments are constructed.

Technical functioning of tests and the items contained within a test is important because tests which categorise students (e.g. merit/pass/fail) need to function in a way which ensures the categorisation is based upon trustworthy items and that will lead to valid interpretation of individual students' marks and grades.

Overall, we found that the majority of tests (over 70%) functioned well or reasonably well.

However, there were some tests which had poor functioning either because too many items within the test had poor functioning and/or because the test design was suboptimal (e.g. too few items in the test).

AOs must have due regard to credible evidence which suggests that a change in its approach to the development, delivery and award of qualifications is required in order to ensure that the approach remains appropriate[1]. As such, each AO received an overall report summarising how their units performed in relation to the other tests; as well as a report for each test included in the research.

It is worth noting that some AOs (particularly those that have been established for some time and/or offer general qualifications) may themselves conduct such analyses routinely and have a well-established feedback loop to help evaluate the quality of assessments and how best to improve them. Other AOs in this study may

[1] Conditions of Recognition D3.2 (Ofqual, 2016)

not have been previously aware of these routine analyses and are now in a better position to conduct their own analyses.

This work, therefore, can be seen as an important tool in helping ensure AOs offer high quality external assessments for vocational and technical qualifications.

2 Introduction

In recent years, school-based qualifications in vocational and technical areas have been required to include external assessments (or examinations) in order to qualify for inclusion on school performance tables. For some qualifications and awarding organisations, external assessments may have been an established part of the qualification's assessment; but in other cases new assessments have been designed and sat and this may be a new endeavour for some AOs/subject experts. This piece of work aims to understand the quality of these examinations in respect of the technical functioning. The item and test analyses used are well-established ways of evaluating test functioning.

3 Data and analysis

For each unit, we asked for anonymised student (or 'candidate') level data for 2016. Data from all 49 units was analysed. In some units, AOs provided multiple versions of each unit relating to different test versions. In such cases, the analysis only utilised data from the largest entry version for each unit.

The test and item analyses are numerous and are outlined in Table 1 below.[2] Ideal values were based upon the professional judgement of the Research and Analysis and Standards teams in Ofqual guided by test construction literature (e.g. Ebel and Frisbie, 1991, Haladyna and Rodriguez, 2013, Opposs and He, 2013). These ideal values were meant to be a guide for test and item functioning, rather than absolute thresholds that rigidly define the difference between high quality and poor quality items or tests, without taking any other contextual information into account. In some contexts, items with values (slightly) outside these ideal values may be perfectly acceptable. Thus, use of such ideal values or benchmarks provide a useful shorthand to *begin* to evaluate the test functioning.

[2] The individual test reports included more detailed analysis – a table describing these analysis are included in Appendix A.

Table 1: Test and item analyses – a brief description of analyses in this report.

Analysis	What does it tell us	What are ideal values?
Item functioning		
Facility	This is a summary of the ease or difficulty of an individual item for the students taking the test. Facility = mean item score / maximum possible item score. Values range between 0 and 1. For a 10 mark item, 0 indicates that the average mark was 0 (0%) while 1 means that the average mark was 10 (100%).	Ideally, in a test which aims to differentiate between students, most facility values should be between around 0.3 and 0.8.
Discrimination indices	These tell us how well an item has contributed to the test in terms of spreading out students of different abilities. It reflects the extent of the relationship between the score on the item and the score on the overall test. R_Rest is the correlation between item mark and total test score minus the item score. Possible values vary between -1 and +1. The closer to 1, the greater the discrimination. A value of 0 indicates no discrimination as students of different abilities score the same. Items with negative values should be inspected closely because they may be measuring something different from the rest of the test.	Values should be positive. The higher the value, the more discriminating the item. Ideally, for tests which aim to differentiate between students of different abilities, values should be greater than +0.3 to indicate discrimination.
Test functioning		
Mean mark	On average, how well students have performed on this test	Around 50% of the maximum marks is generally considered appropriate for tests aiming to differentiate between students. For tests which are competency-based and 'enter when ready', it may be that a higher mean mark is appropriate.
Standard deviation (SD)	How well the test has spread out students in the available mark range.	Should be greater than ≈15% of the number of marks available.
Reliability coefficients	Reliability coefficients are measures of consistency of test results. The reliability measures reported here are derived based on the internal structure of the tests – internal reliability. Cronbach's Alpha – an estimate of reliability of a test derived based on the internal structure of the test (Cronbach, 1951). High values suggest the test is internally coherent – that the test is measuring a common construct.	Ideally greater than 0.8 to indicate acceptable levels of reliability.

4 Analysis outcomes

The figures below present all 49 units (tests) from the 27 qualifications in the sample for each item and test analysis.

For the majority of tests, the item facilities fell within the ideal range. Figure 1 shows for each of the 49 units the distribution of item facilities. Each unit's item facilities are displayed in a box and whisker chart where the box shows the middle 50% of the item facilities (interquartile range), and the whiskers represent the data outside of the interquartile range and they extend 1.5 times the interquartile range from the top and bottom of the box respectively. The larger the box and whiskers, the greater the variability in the item facilities within the test. Items with facilities that fall outside of the whiskers are shown as solid points and are considered to be outliers. The black line shows the median value – the midpoint of the item facilities within the unit (in other words, 50% of items fall above and 50% of items fall below). All the box plots have been ordered according to the median item facilities, from lowest to highest along the x-axis.

The majority of these tests have items where most facility values fall within the normally acceptable range. This is shown by most of the boxes on the box plots lying entirely or substantially within the green area.

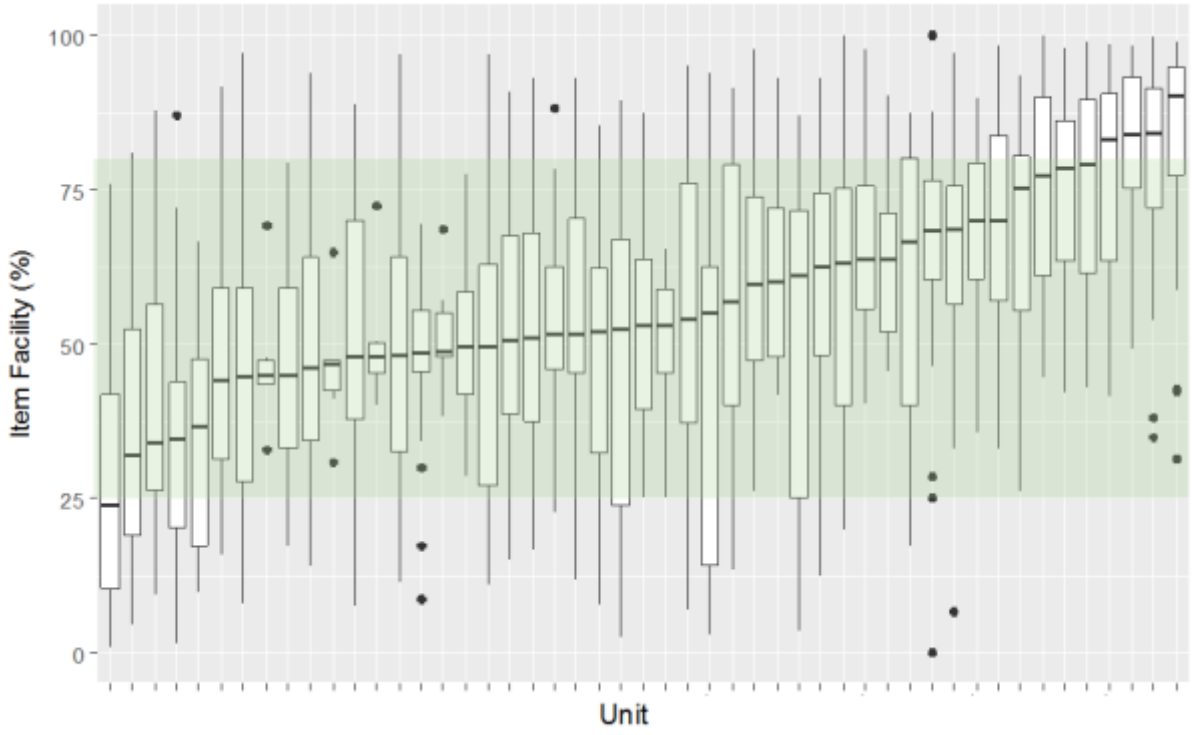


Figure 1: ***Box and whisker plots showing the distribution of item facilities for each of the 49 tests. The green area indicates the ideal range of item facilities. Tests arranged according to ascending order of median facility value (black horizontal line)***

The leftmost box plot in Figure 1 has predominantly 'difficult'[3] items, with over 50% of items below 25% (or facility value of .25) indicating that the students on these items scored on average 25% or less.

Four tests on the right of the figure represent tests with more than 50% of items being 'easy' – these had more than 50% of items with mean marks of 80% or more.

Many tests have quite a wide range of item facility values – shown by the size (height) of the boxes, length of whiskers and presence of outliers. One test has two very extreme outliers: one at 100 (or a facility of 1) and one at 0 (facility of 0) ie on one item all students received available marks and on another no student received any marks. Such extreme items may be valuable in terms of assessing important knowledge or skills, but they do nothing to contribute to the overall functioning of the test.

Where a test overall has either predominantly easy items or predominantly difficult items, it is likely that the tests will not have adequately discriminated between students of different levels of ability – they lead to narrow mark distributions and the awarded grades are close together.

The mean mark of a test is a direct function of the range of item facilities. Figure 2 below shows mean mark as percentage of the overall mark total for each test. One test has a very low mean mark at just 23% of the total marks available (on the left-most side of the graph) while three have very high mean marks of 80% or more of total marks available. As commented above, these tests with high scores are unlikely to lead to much differentiation between students; and for tests with more than one passing grade (true of all tests in this analysis), such high mean marks would be unlikely to provide adequate differentiation between students of different abilities.

[3] Item facilities tell us about the 'difficulty' or 'ease' of items only for those students taking the test; they do not tell us about 'inherent ease' or 'inherent difficulty'.

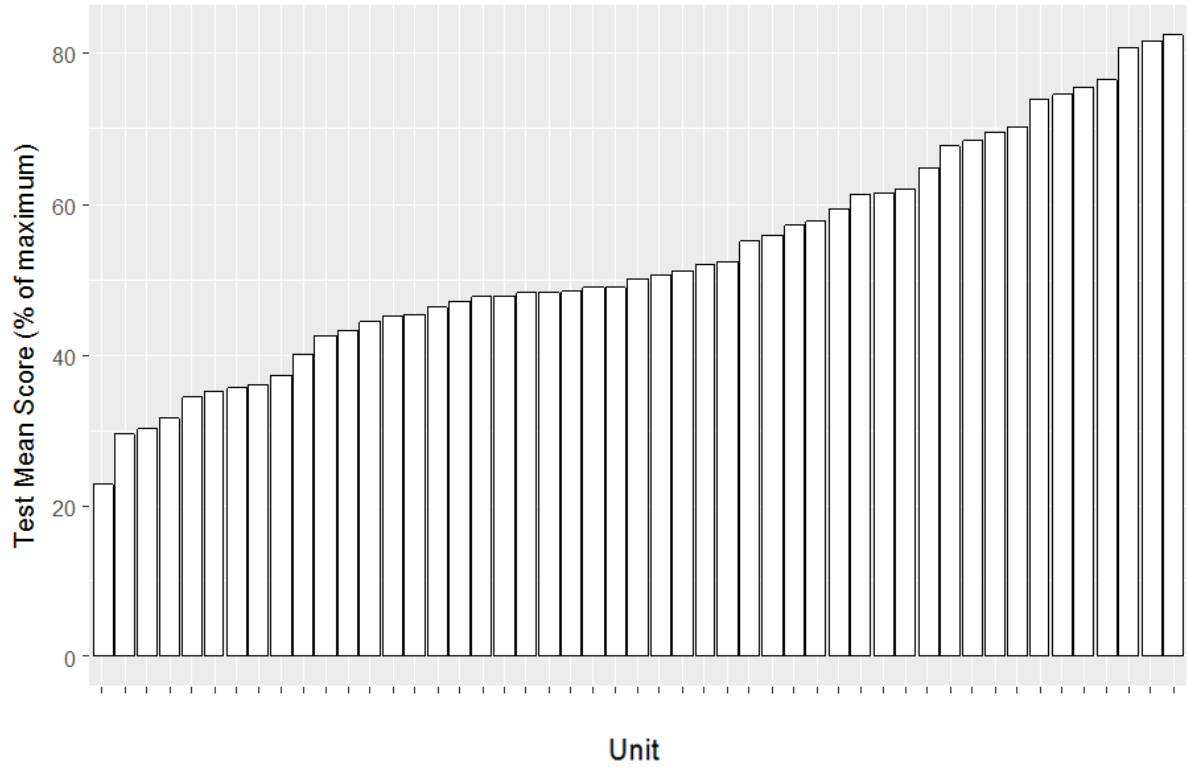


Figure 2: Bar chart showing the mean score for each of the tests (expressed as a percentage of overall available mark); arranged in ascending order of mean mark.

Tests which aim to differentiate between students of different ability (this might be indicated by having more than one passing grade) need items that in themselves discriminate (i.e. higher ability students do better than lower ability students on the individual items). If items do not discriminate, they are not contributing to the measurement properties of the test i.e. to spread out students according to their ability. Figure 3 displays, for each test, the distribution of item discrimination indices. As with figure 1, this uses box and whisker plots and is arranged in ascending order by median discrimination value.

Unlike facility indices, where there is an acceptable *range*, for discrimination indices the higher the better; and there is no advantage to the measurement properties of a test to have a range of discrimination values.[4] In short, the more discriminating the better. Ebel and Frisbie (1991) indicate that items with discrimination between 0.2 and 0.29 are 'marginal items, usually needing and being subject to improvement' before inclusion in a test; while items with discrimination indices less than 0.19 are

[4] NB It is normal (and acceptable) practice to have one or two items at the beginning of the test which most students get right to settle students into the test – items which have both high facilities and little discrimination.

poor items. Haladyna and Rodriguez (2013) indicate 0.15 or above as acceptable for test items which are objective (multiple choice or selected response) and which are contained within a test of reasonable length.

Again, the majority of tests (36 tests) have more than 50% of items with generally accepted levels of discrimination. Thirteen tests had items where more than half of the items did not have generally accepted levels of discrimination. These are on the far left side of the graph and the median line falls out of the green area (which indicates acceptable levels of discrimination). It might be possible for some items to be valuable in a test despite low levels of discrimination on the basis of testing important baseline knowledge or skills. But in a test which is not a competency test, but which aims to differentiate and grade students according to their ability, tests with high proportions of non-discriminating items are likely to be suboptimal.

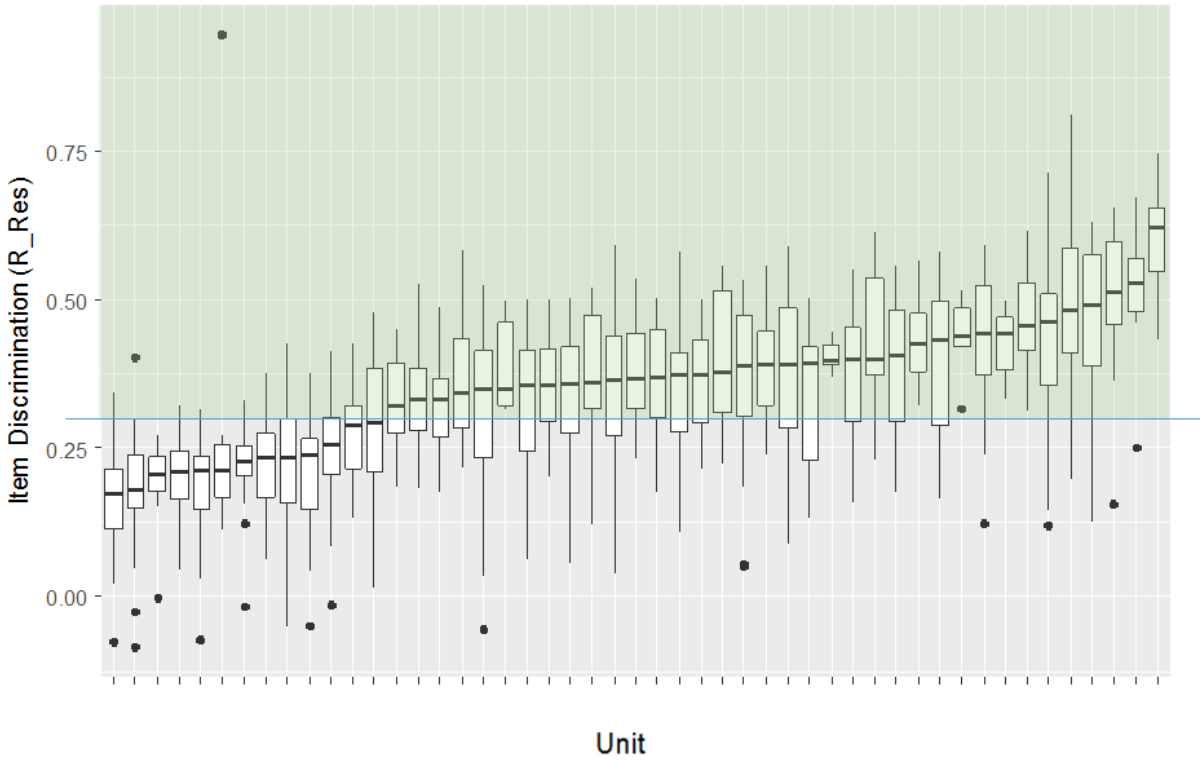


Figure 3: ***Box and whisker plots showing the distribution of discrimination indices (R_Rest) for each of the 49 tests. The green area indicates the ideal range of item discrimination indices. Tests arranged according to ascending order of median discrimination value (black horizontal line)***

The standard deviation of a test indicates the range of marks awarded to students on any test. As a general rule of thumb, ideally tests should have standard deviation on

or above 15% of the maximum mark; the larger the standard deviation, the better the test has spread students across the mark range.

Most of the tests had acceptable standard deviations while around twenty had suboptimal. Two tests had standard deviations less than 10% of the marks available. To help put this into context, a test with a standard deviation of 10% of the marks available would mean that for a 100 mark test, students would be so tightly clustered that 67% of students would fall within a range of 20 marks; and only 33% of students in the other 80 available marks. So, again, the tests on the extreme left side of the graph are concerning.

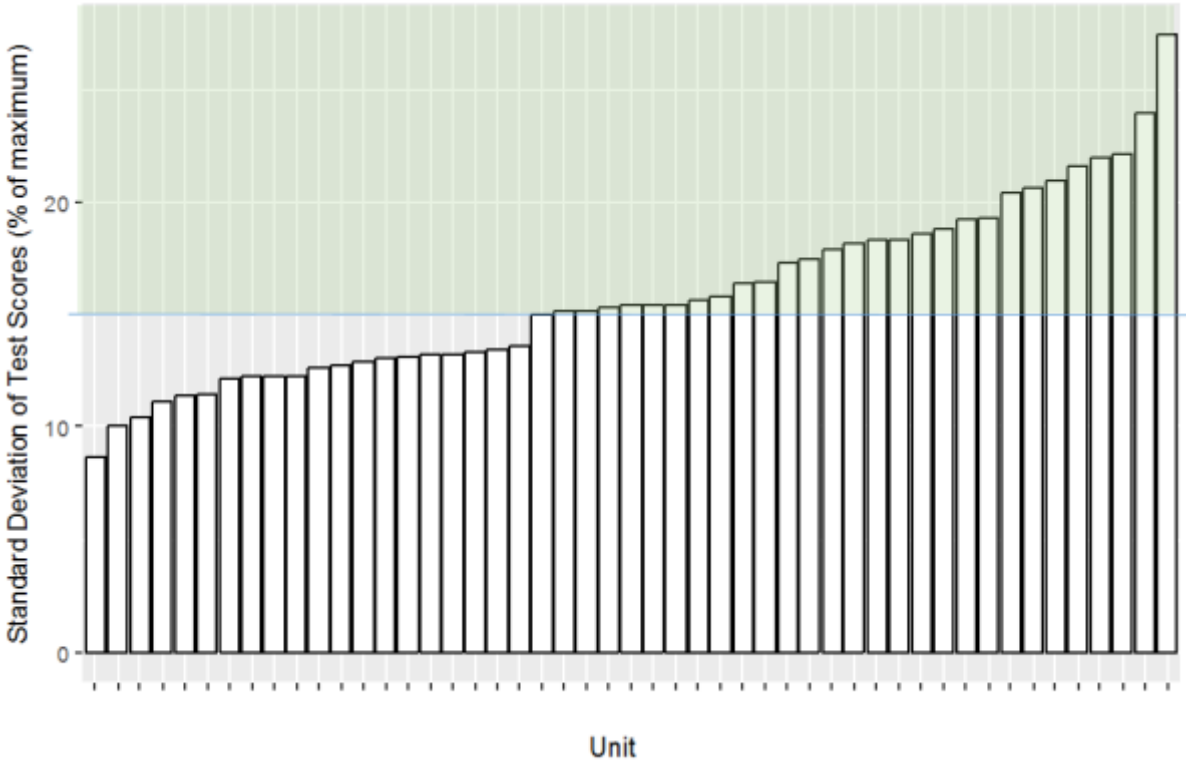


Figure 4: ***Bar chart showing the standard deviation for each of the tests (expressed as a percentage of maximum available mark); arranged in ascending order of standard deviation.***

The reliability of the tests is presented in Figure 5 below. This presents the values of Cronbach's alpha in ascending order. The green area indicates 'normally acceptable' values' of Cronbach's alpha, indicating tests which are internally consistent (and thus provide a consistent measure of the construct). Of the 49 tests in the research, 22 had values of Cronbach's Alpha over 0.8 (i.e. acceptable); while 27 had values below this range. Of these, 12 were relatively close to our 'arbitrary' benchmark with values between 0.7 and 0.8. As reliability is a function of both test length (tests with more

items and greater maximum marks generally have high values of reliability) as well as dimensionality (tests measuring a clear, single construct have high values of reliability), it is possible that one or both of these features explain why tests have lower levels of reliability. However, 3 tests had extremely low values (0.29, 0.33 and 0.33). Low reliability such as this is likely to undermine any value in the test and its ability to measure the stated construct.

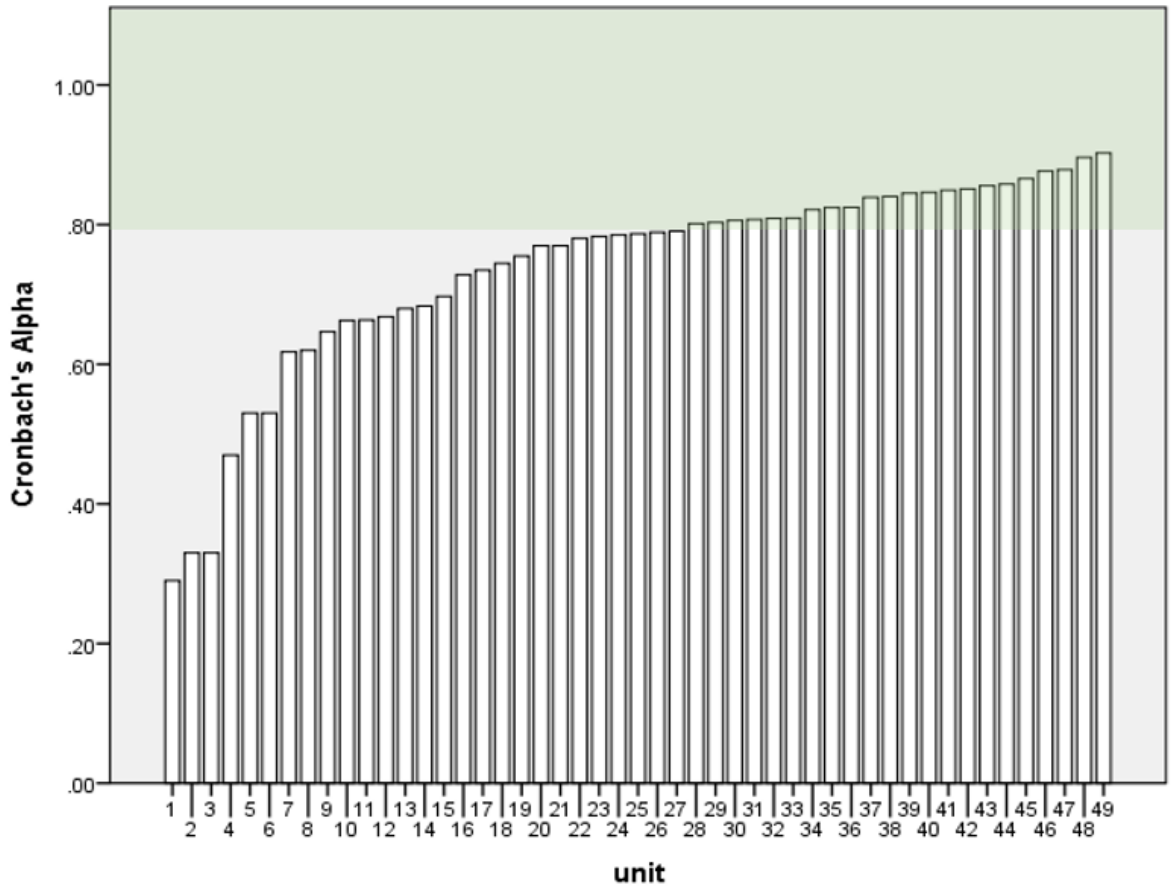


Figure 5: **Values of reliability - Cronbach's Alpha - for each test, arranged in ascending order of Cronbach's Alpha.**

Summary of test functioning

The analyses presented so far give a sense of the distribution of the tests for all the dimensions that indicate quality. However, they do not show the profile of any one particular test in respect of each dimension of functioning. Table 2 below summarises different types of test quality and functioning according to the combination of the different dimensions[5].

[5] Table 2 is based upon a concept of Haladyna and Rodriguez (2013, page 250) whose table evaluates item quality according to the clustering of individual item dimensions.

Given all the normal caveats around applying some arbitrary standards, this seems to indicate that 35 tests in the sample (71%) had good or reasonable functioning and 14 tests (29%) had some issues in terms of test functioning. It is possible (see discussion below) that some of these 14 tests did have adequate functioning for the contexts in which they operate, the constructs being measured, or in relation to the stated purpose of the tests. This is discussed further in section 4.

A short note on Table 2 and how it was derived:

- All four attributes of test functioning (profile of facilities, profile of discrimination indices, SD and Cronbach's Alpha[6]) have easy and clear to apply categories. For example, 'most items' means >50% of items.
- In exploring how to derive this table, we sub-divided each attribute into 2 or 3 sub-categories and looked at how many overall 'test types' this generated. The final table represents a balance between the sensitivity of the sub-categorisation of the attributes of test functioning and a manageable number of categories. With different underlying data it is possible that different decisions might have been made.
- Interestingly, once we settled upon these categorisations and the resulting types, it was observed there are particular patterns by AO such that some of the types are represented entirely by the tests from one AO. This may indicate that certain AO test construction and design principles do lead to certain patterns of test functioning. It also therefore gives some support the use of this typology in helping to identify and understand AO specific profiles.Table 1

[6] Mean mark was not included as it would not offer much additional help to the categorisation beyond the profile of item facilities).

Table 2: Some (arbitrary) standards for evaluation of test functioning – nine test types

Type	Description of facility indices	Description of discrimination indices	Description of SD	Reliability	Description of Type	Possible explanations or issues to explore?	Number of tests	Percentage of tests in study
1	Most items within acceptable range	Most items above 0.3	Above 15% of maximum marks	> 0.8	Test has good functioning.		15	30.6%
2	More than half of items have overly high facilities	Most items above 0.3	Above 15% of maximum marks	> 0.8	Easy but otherwise well functioning test	Test items too easy and poorly targeted at the cohort e.g. the cohort is more capable than the test. Could suggest that the test is not of the appropriate level.	2	4.1%
3	Most items within acceptable range	Most items above 0.3	Above 15% of maximum marks	< 0.8	Test has good functioning except for lower reliability	Possible explanations are too few items or a multidimensional construct being tested.	5	10.2%
4	Most items within acceptable range	Most items above 0.3	Below 15% of maximum marks	> 0.8	Test has reasonable functioning but some issues in differentiating between students	Possibly large tariff items have not used extremes of mark range.	3	6.1%
5	Most items within acceptable range	Most items above 0.3	Below 15% of maximum marks	< 0.8	Test has reasonable functioning but some issues in differentiating between students and lower than ideal reliability	See type 3 above; Possibly large tariff items have not used extremes of mark range.	10	20.4%
6	Most items within acceptable range	Fewer than half the items have acceptable discrimination values	Above 15% of maximum marks	< 0.6	Some issues in terms of the basis upon which students have been differentiated and low reliability	The test has problematic functioning in terms of the basis upon which students have been differentiated. Such low reliability indices can indicate tests with too few items. Items which have differentiated between students on an arbitrary basis (e.g. guessing).	6	12.2%

Type	Description of facility indices	Description of discrimination indices	Description of SD	Reliability	Description of Type	Possible explanations or issues to explore?	Number of tests	Percentage of tests in study
7	Most items within acceptable range	Fewer than half the items have acceptable discrimination values	Below 15% of maximum marks	< 0.8	Non-discriminating test.	Items are appropriately pitched. However, the test has problematic functioning in that it has not succeeded in differentiating between students and in terms of the basis upon which students have been differentiated.	5	10.2%
8a	More than half of items have overly high facilities	Most items above 0.3	Below 15% of maximum marks	> 0.8	Too easy and not differentiating	Test has reasonable functioning except poor item targeting has meant the test has not succeeded in differentiating between students.	1	2%
8b	More than half of items have overly low facilities	Most items above 0.3	Below 15% of maximum marks	> 0.8	Too difficult and not differentiating	Test has reasonable functioning except poor item targeting has meant the test has not succeeded in differentiating between students.	1	2%
9a	More than half of items have overly high facilities	Fewer than half the items have acceptable discrimination values	Below 15% of maximum marks	< 0.8	Too easy and poorly performing on all fronts		1	2%
9b	More than half of items have overly low facilities	Fewer than half the items have acceptable discrimination values	Below 15% of maximum marks	< 0.8	Too difficult and poorly performing on all fronts		0	0%

5 Discussion

The majority of tests had good or reasonable functioning overall according to the ideal values described in Table 1 and the categorisation and types outlined in Table 2.

Purpose of the tests

There are some potential issues regarding this type of analysis of tests and items and their categorisation. The key issue is the purpose of the test. Traditionally, these sorts of analyses have been conducted upon tests for which the main purpose is to *rank* students. Typically, in GCSEs and A levels, the purpose of tests is explicitly to rank students according to their knowledge, skills and understanding in relation to a broad proficiency domain (e.g. 'Chemistry' or 'History'). The tests should mean that a student with a higher mark or grade than another student has greater proficiency in respect of their knowledge, skills and understanding. Thus, for a test with the explicit purpose of ranking, the underlying test design principles should focus particularly heavily upon item discrimination. This means that such tests should avoid items which either most students would get right or most students would get wrong, as such items provide little information about how students differ in relation to the construct being tested.

For tests within the vocational and technical sphere, while some tests might have the purpose of ranking, others may have a different purpose – that of identifying those students who have mastered a particular proficiency (versus those who have not). These tests will have very different underlying principles for test design and construction. In particular, discrimination across the full mark range will not be the key principle; the focus will be on testing content that is deemed to be important or essential for mastery of the particular domain. This means that most students, if they have been entered appropriately, should get such items correct.

Thus, it might be the case that tests in the sample which have high facilities and low discrimination indices are functioning adequately – are of appropriate demand and discrimination – if the purpose of the test is mastery of a domain.

Therefore, an important question for this research is whether or not the tests' purposes are ranking (and hence amenable to the kinds of analysis with the ideal functioning described in Table 1) or whether they had another purpose and some other analyses or benchmarks of ideal functioning apply. The vital clue is in the grading scheme – the possible outcomes for students. A simple Pass/Fail grading scheme for a test would be consistent with a mastery test; while a grading scheme with two or more passing grades e.g. Merit/Pass/Fail or A to G grades would be consistent with a 'ranking' test. All of the tests in this report were of the latter sort and so the analyses and their benchmarks are likely to be relevant and be useful at least as a good starting point for an evaluation of the functioning of the test.

There may be tests in the sample which perhaps are more of a hybrid of purposes, or for which the purpose, design and grading scheme have not been logically aligned. If that is the case, this report is likely to at least flag up some of the risks associated with inconsistent design implementation.

Grade boundaries

It is also worth commenting upon grade boundaries. These have not been presented in the analyses due to different tests having very different cohorts and being of different levels and also because it would require an entirely different study to evaluate the overall *standard* of students with any particular grade in any particular test. For a ranking test, it is ideal for the grade boundaries to be well spread along the mark range and reasonably centred in the mark range and/or around the main part of the mark distribution. Many of the tests in the study had mark distributions and grade boundary locations which looked reasonable from a measurement perspective. However, for some tests, the passing grades were very close together and clustered in the top part of the mark range (with a risk of the measurement properties of the test being unable to adequately differentiate between students); or sometimes very low in the mark range (implying that relatively little proficiency in terms of knowledge, skills and understanding was required).

Reliability

How good or bad is the distribution of reliability values found in the tests in this study? Bramley and Dhawan (2010) looked at test reliability coefficients in 287 GCSE and AS/A level tests and so provides a useful comparison point. The average Cronbach's Alpha coefficient was around 0.8 in Bramley and Dhawan compared to 0.74 in our study. In Bramley and Dhawan, the lowest value was 0.421, with less than 3% having values less than 0.5. In this report on school-based vocational and technical tests (with a sample of 49 tests rather than 287), there is a greater proportion of tests with lower reliability – with 8% of tests having less than 0.5 and three of them having levels lower than any observed in the Bramley and Dhawan study. In general, it may be that the tests in our study have lower reliability values because they are shorter tests or because they have items which do not contribute to the reliability and/or because they are testing constructs which are naturally more multi-dimensional in nature compared to that of GCSEs and AS/A levels.

However, no matter the possibility of multi-dimensionality (or the purpose of the test) – tests with very low reliability coefficients as some seen here have little to offer in terms of measurement and render the value of the test questionable.

Context

In terms of the cohort of students taking a test, it is also important to note that the context of these tests and their cohorts were not examined as part of this research – and in some cases this might mean an interpretation of the adequacy of test functioning might change. For example, it might be that a test with a very low mean

mark value has items which are of appropriate demand but that the students were insufficiently prepared. Some of the tests included in the sample may have been newly introduced tests which had not been sat before. This may have meant that students and their teachers were not fully aware or prepared for the requirements of the test.

Good test functioning – a guarantee of a high quality test?

Finally, it is worth commenting upon how far reaching this kind of analysis is, and what conclusions about an overall test's quality can be drawn. Does it give the whole picture as to the quality or validity of a test? Specifically, is it the case that a test which has a good profile of test functioning – appropriately difficult items, good discrimination and so on, is a high quality test on all fronts? The short answer is no. While good test functioning is a necessary condition of a high quality test which aims to rank order students, it is, in itself, not sufficient. It is possible for items to produce credible looking statistics for reasons other than those related to the purpose of the test. There are many other aspects of test quality and validity which are important – not least ensuring that they are 'assessing the right thing' (Newton, 2017). This particular line of work does not address this; this work is not a full validation of the tests, and just one aspect of a review phase which AOs and their tests can benefit from.

6 Concluding comments

This work indicates that the majority of tests which were analysed had good or reasonable test functioning. This work has helped Ofqual engage with a number of AOs around the quality of external tests in school-based vocational and technical qualifications. We hope to conduct such analyses more routinely, as more and more vocational and technical qualifications include external assessments. AOs in turn will have greater regard to technical functioning and ensure this becomes part of their own quality assurance.

7 References

Bramley, T. and Dhawan, V. (2010). *Estimates of Reliability of Qualifications*. In: Ofqual *Reliability Compendium* (Chapter 7). Coventry: Ofqual [online]. Available: https://www.gov.uk/government/publications/reliability-of-assessment-compendium

Ebel, R. L. and Frisbie, D. A. (1991) Essentials of Educational Measurement, Prentice Hall, Engelwood Cliffs, New Jersey, US.

Haladyna, T.M. and Rodriguez, M.C. (2013) Developing and Validating Test Items. Routledge, Oxford, UK.

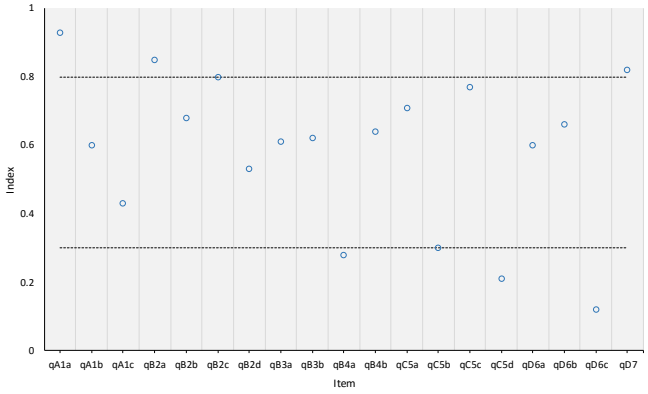
Newton, P. (2017) An approach to understanding validation arguments, Ofqual, Coventry. Available https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/653070/An_approach_to_understanding_validation_arguments.pdf (accessed 7 November 2017)

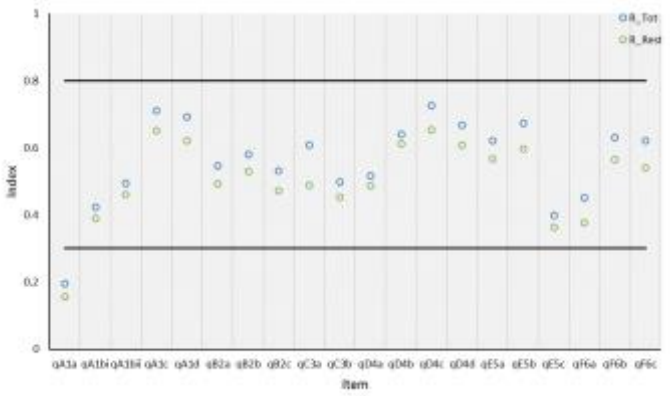
Opposs, D. and He, Q. (2013) Introduction to the Concept of Reliability. In: Ofqual *Reliability Compendium* (Chapter 7). Coventry: Ofqual [online]. Available: https://www.gov.uk/government/publications/reliability-of-assessment-compendium (accessed 10 October 2017)

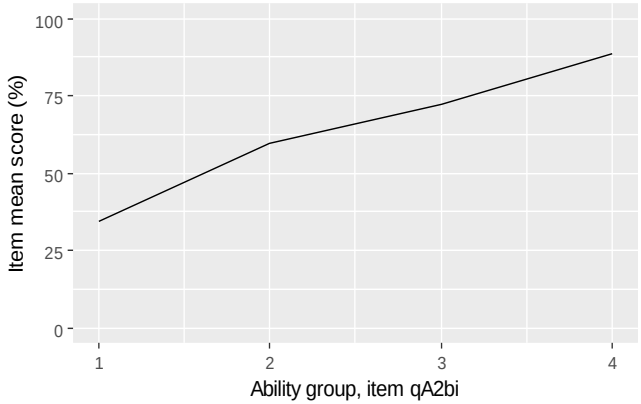
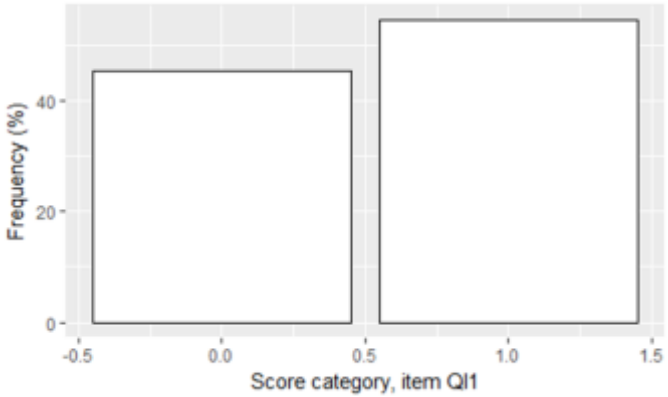
8 Appendix

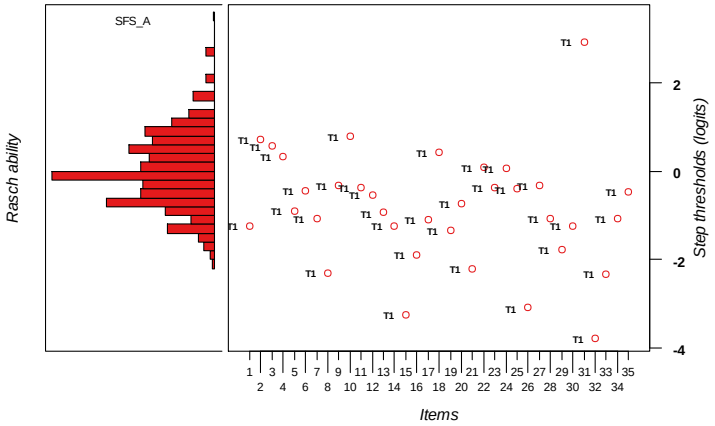
Table 3: Test and item analyses – a brief description of those provided to AOs for each test.

Analysis	Statistic or chart	What does it tell us	What are ideal values?
Test functioning			
Mean mark	Statistic	On average, how well candidates have performed on this test	Around 50% of the maximum marks is generally considered appropriate for tests aiming to differentiate between candidates. For tests which are competency-based and 'enter when ready', it may be that a higher mean mark is appropriate.
Standard deviation (SD)	Statistic	How well has the test spread out candidates in the available mark range	Should be greater than ≈15% of the number of marks available.
Reliability coefficients	Statistic	Reliability coefficients are measures of consistency of test results. The reliability measures reported here are derived based on the internal structure of the tests – internal reliability. Cronbach's Alpha – an estimate of reliability of a test derived based on the internal structure of the test. It may be interpreted under certain conditions as a measure of the internal consistency of the test – how closely related are a set of items as a group. Omega_H is based on factor analysis - tells us the percentage of the variance of test scores that can be explained by a general factor. It may be viewed as a measure of the unidimensionality of the test. Omega_T is a measure of the total test score reliability estimated based on factor analysis, involving the use of	Ideally greater than 0.8. to indicate acceptable levels of reliability.

Analysis	Statistic or chart	What does it tell us	What are ideal values?
		a general factor and a set of factors associated with items grouped together according to their relatedness. Reliability measures tell us something about the quality of the test in that if the test is repeated, high reliability measures indicate there should be high similarity in the test results.	
Mark distribution	Chart	This displays the distribution of marks for the whole tests. It is possible to see the extent to which there is skewness.	A good mark distribution should show a good distribution of marks, centrally located with little skewness and with most mark points used. For tests which are competency-based and 'enter when ready', it may be appropriate for the mark distribution to be more skewed towards the top end of the mark range.
Item functioning			
Facility	Statistic, tabulated	This is a summary of the ease or difficulty of an individual item. Facility = mean item score / maximum possible item score. Values range between 0 and 1; For a 10 mark item, 0 indicates that the average mark was 0(%) while 1 means that the average mark was 10 (100%).	Ideally, in a test which aims to differentiate between candidates, most facility values should be between 0.3 and 0.8.
Facility values plot.	Chart	This chart provides a visual summary of the range of facilities for all the items on the test.	Ideally, all or most should fall within the 0.3 to 0.8 range. The following is reasonable. 

Analysis	Statistic or chart	What does it tell us	What are ideal values?
			It might be the case that for tests which do not aim to discriminate between candidates of different ability – those assessments which have a mastery or competency model, that the profile may be different with a greater proportion of items having higher values.
Discrimination indices	Statistic, tabulated	These tell us how well an item has contributed to the test in terms of spreading out candidates of different abilities. It reflects the extent of the relationship between the score on the item and the score on the overall test. R_Tot – correlation between the item mark and whole test score; R_Rest – correlation between item mark and total test score minus the item score. Possible values vary between -1 and +1. The closer to 1, the greater the discrimination. A value of 0 indicates no discrimination. Negative values should be treated with caution.	Values should be positive. The higher the value, the more discriminating the item. Ideally, for tests which aim to differentiate between candidates of different abilities, values should be greater than +0.3 to indicate discrimination.
Discrimination indices plot	Chart	This plot provides a quick visual reference for the tabulated discrimination indices, both R_Rest and R_Tot. This helps to see the extent to which the items as a set have functioned.	In general, discrimination values should be above 0.3.  Most of the items have values above 0.3, and many above 0.5 – and so acceptable.

Analysis	Statistic or chart	What does it tell us	What are ideal values?
Item Characteristic Curves (ICCs)	Charts	ICCs depict both item facility with respect to ability and discrimination. ICCs plot facility (item mean score) by ability group split into ability quartiles. The slope of the graph indicates the overall discrimination such that an incline indicates that the item has successfully discriminated between candidates of differing ability, while a flatter line indicates that the item has failed to do so.	Ideally, ICCs should display an even slope ranging from approximately 20% for the least able quartile to approximately 80% for the most able quartile. 
Item mark distributions	Charts	These show frequency of marks awarded. While less useful for one mark questions, we have included these as they also represent the facility and, when presented alongside the ICCs, can help aid understanding.	
Achieved weighting versus intended weighting	Statistic, tabulated	This tells us whether individual items contribute their intended weight towards the assessment unit as a whole. Each item has an intended weighting represented by the assigned mark (e.g. a 5 mark item on a 50 mark test has an intended weighting of 10%). The achieved weighting takes into account the variability of the item marks in relation to the overall variability of the unit and how well they have discriminated candidates of different abilities.	Ideally, the ratio of achieved weighting to intended weighting should be as close to 1 as possible – indicating close alignment between the intended weighting and achieved weighting. Between 0.5 and 1.5 is broadly acceptable.

Analysis	Statistic or chart	What does it tell us	What are ideal values?
		Achieved weight = $\frac{R\_Tot \times SD\_item \times 100}{SD\_unit}$ - Where R_Tot is the correlation of item marks with total mark on the unit - SD_item = standard deviation of item marks - SD_unit = standard deviation of unit marks	
Wright map of item targeting	Chart	This chart shows how well the range of item difficulties matches the range of candidate ability. This is based upon a statistical model called Rasch, which takes into account item difficulty when estimating the ability of test takers. On the chart, the left hand side of the histogram shows the distribution of candidate ability. On the right hand side, each item is displayed along the x axis, and locations on the y axis indicate the difficulties of the marks assigned (the 'score categories') of the item ('step difficulty' or 'threshold'). The items with step difficulties at the bottom are easy in relation to the ability of the candidates, while the items with step difficulties at the top are difficult.	Ideally, the location of the majority of the item difficulty thresholds should sit within the band where the majority of candidate abilities are located. The Wright map below indicates that some of the items are a bit easy given the ability of the cohort, but they mostly look appropriate. 

We wish to make our publications widely accessible. Please contact us at publications@ofqual.gov.uk if you have any specific accessibility requirements.

This publication is available at www.gov.uk/ofqual.

Any enquiries regarding this publication should be sent to us at:

Office of Qualifications and Examinations Regulation

Spring Place
Coventry Business Park
Herald Avenue
Coventry CV5 6UB

Telephone 0300 303 3344
Textphone 0300 303 3345
Helpline 0300 303 3346