The internal reliability of some City & Guilds tests

A project report prepared by City & Guilds and its research partners for the Office of Qualifications and Examinations Regulation

Andrew Boyle The City and Guilds of London Institute

Zeeshan Rahman The City and Guilds of London Institute

Ofqual/13/5257 January 2013

This report has been commissioned by the Office of Qualifications and Examinations Regulation.

Contents

Tables

Figures

Acknowledgments

We would like to thank Sandra and Rod Johnson at Assessment Europe for providing technical expertise and reviewing the report.

Thanks to Linda Miller at the Institute for Employment Studies who also reviewed the report.

Finally, thanks are due to all City & Guilds colleagues involved in providing relevant information, data and support, particularly Joe Francis, Gareth Hopkins, Michelle Piper, Tom May and Milja Curcin.

Executive summary

This document reports a research project conducted by the City and Guilds of London Institute, supported by its research partners Assessment Europe and the Institute for Employment Studies between October 2011 and March 2012 under tender number 126 let by the Office of Qualifications and Examinations Regulation (Ofqual). Ofqual's project title was 'the estimation of internal reliability'.

This report starts with a review section, which begins by outlining the history and current status of vocational qualifications (VQs) and noting that VQs have always been on a 'parallel but separate track' to academic examinations. VQs are supported by their particular conceptions of quality, which can be judged on their own merits. The conceptualisations of quality underpinning VQs pre-date recent government interventions, and can be evaluated notwithstanding those interventions. Most commentators on the qualifications scene agree that government intervention in qualifications has increased over the past 30 or so years. However, there was never a 'golden age' of no intervention, and the merits of VQs have to be argued substantively; it is not enough to simply blame botched interventions for perceived weaknesses.

The review goes on to describe key concepts underlying VQ assessment. The two main concepts are competence-based assessment and criterion-referenced testing (CRT). The former's importance is that English VQs have been said to address this concept since their inception. However, the elaboration of competence-based assessment gives few direct hints about which reliability techniques would be the most appropriate for analysis. In contrast, there is a substantial psychometric literature suggesting suitable reliability indices for CRTs.

Procedures are then described for selecting a sample of tests upon which to conduct reliability analysis along with appropriate indices to express the reliability of the tests.

The findings were as follows:

- 68 tests were analysed. These were balanced across (QCF) levels 1, 2 and 3. The sample also included qualifications of varying risk status, and qualifications that had licence to practise functions and which formed parts of apprenticeship programmes.
- 47 of the analysed tests were fixed form (all candidates received the same questions) and 21 were randomly generated (different candidates received different questions – making up tests to tight content specifications). These two types of test design required differently designed reliability analyses.

- Cronbach's alpha coefficient and Guttman's lambda-2 index were calculated for the 47 fixed form tests. The mean values and standard deviations of the coefficients were almost identical in the two cases: mean of 0.79 for alpha and 0.80 for lambda-2, with standard deviations of 0.11 and 0.10 respectively.
- The phi dependability coefficient for absolute measurement and the phi (lambda) coefficient (a dependability coefficient adjusted to account for the presence of a cut score – lambda) were calculated for all 68 tests in the sample. Mean values were 0.72 and 0.88, respectively, with standard deviations of 0.14 and 0.12.
- The reliabilities for different forms of tests were compared. Paper fixed form tests appeared to provide the highest reliability values, whereas online randomly-generated tests appeared to provide the lowest indices amongst the assessment types. However, no test was conducted to establish whether this difference was statistically significantly different.
- The reliability of data sets with different types of kurtosis were compared. There did not appear to be major differences between leptokurtic (steeply peaked) data sets and mesokurtic (those data sets whose peakedness was within the range of the normal distribution).
- Other reliability techniques were implemented on small numbers of tests for exemplification. These techniques were: Livingston's index and the Standard Error of Measurement (SEM) (along with 95% confidence intervals – CIs).

Several observations are offered in the discussion of findings, plus a number of recommendations. These are:

- The report, following that of Harth & Hemker (2011), and in parallel to the City & Guilds partner project in the current round of tenders (Johnson et al, 2012), adds significantly to the emerging body of reliability research on VQs.
- The analyses indicate that VQs can provide reliable outcomes. The mean values for Cronbach's alpha in this project were very similar to those returned in reliability analysis of GCSEs and A levels.
- Selecting a sample of tests for analysis was inhibited by inconsistencies between regulatory and awarding body databases. City & Guilds and Ofqual should review and update their respective databases in light of the issues raised in this report and work collaboratively to do so where this is mutually advantageous.
- The classical test theory index Cronbach's alpha is a starting point for reliability analysis. It has limitations – particularly when applied to CRTs – but its ubiquity and ease of calculation mean that all credible awarding organisations should be able to calculate and interpret it.
- We did not find Guttman's lambda-2 index to add much to alpha, and would not recommend its operational use by awarding organisations.
- The phi coefficient seemed intuitive to us, practical to calculate and to have a basis in principle that was suitable for CRT applications. We recommend that vocational awarding organisations investigate its use operationally.

- Phi (lambda) has similar philosophical bases to phi and thus appears to be useful. It also has the potential to 'bundle' absolute measurement and an adjustment to account for a cut score in a single index. This may make it a very efficient and parsimonious index for VQ purposes. However, awarding organisations may consider that this 'bundling' creates confusion and may prefer to keep the classification and dependability aspects of reliability estimation distinct.
- Livingston's index is a less defensible squared error loss index than phi (lambda), but it could be used (with suitable health warnings) by an awarding body that was not able to calculate phi (lambda).
- The derivation of the SEM and associated CIs is a reputable technique to show the reliability of test outcomes as defined in score precision. The interpretation of the amount of score precision necessary depends upon context. High score precision in other assessed elements within a qualification might mitigate low precision within the measurement from a knowledge test.
- We noted two concerns for estimating the reliability of CRT outcomes which need to be thought through. When scores were 'bunched' in only a section of the scoring scale reliability indices should – in principle – be low. Also, a mean score that was very close to a cut score ought to be a concern – suggesting high levels of misclassification. However, both these issues need to be considered carefully. They may suggest limits of reliability analysis, which we suggest is best conducted in tandem with informed validity research.

Introduction

Basis for this study

This research report is the final output of research undertaken in response to an invitation to tender (ITT) issued by the Office of Qualifications and Examinations Regulation (Ofqual) on 17th June 2011 (Ofqual, 2011a). In its specification of requirements, Ofqual stated the research aims and objectives as follows:

> While previous research on the Reliability Programme explored mainly the use of Cronbach's alpha as a measure of internal reliability, research funded under this specification is intended to generate estimates for a range of internal reliability indices for a selection of assessments. The research may involve the following activities:
>
> - To select a range of assessments where it can reasonably be expected that test-related unreliability represents the major source of measurement error. For example, tests and examinations that are composed of multiple choice questions (MCQs) and/or short-answer constructed responses that can be objectively marked may be investigated.
> - To produce estimates for a range of internal reliability indices for the selected assessments.
> - To analyse, interpret, compare and report on the reliability evidence generated.
> - To assess the practical applications of the specific estimation techniques used in the research. (Ofqual, 2011a, p. 1)

In its successful response to Ofqual's ITT, City & Guilds claimed that their approach would afford Ofqual the following benefits:

- An opportunity to rectify the previous under-representation of [vocational qualification] VQ assessments in reliability research.
- An opportunity to sponsor research based on large sets of suitable tests.
- An opportunity to derive meaningful and coherent descriptions of tests – rectifying inconsistencies in current awarding body and regulatory databases.
- The application of a range of reliability indices to test score data, including the use of diverse indices beyond the commonly used Cronbach's alpha.
- An understanding of fundamental issues in vocational and – indeed – all forms of assessment; relationship between [Criterion Referenced Testing[1]] CRT and [Norm Referenced Testing] NRT, role of knowledge testing in VQs, empirical

[1] See pp. 10ff, below for more on criterion referencing.

> basis for statements about competence-based assessment, etc. (City & Guilds, 2011a, p. 11)

As the final output from this project, this report seeks to deliver on the promise stated above.

Structure of this report

This report begins with a review of the history and current status of technical education and VQs; key concepts in the assessment of VQs; and the concept of reliability and reliability indices. The method section follows with an explanation of the approaches to taxonomy development, test selection, data collection and preparation, and analysis. The third section presents findings from the reliability analysis, the implications for which are then explored in the discussion section. The report ends with conclusions and recommendations.

Review

A brief history of technical education and VQs

Technical education and training systems in England have histories running back to the industrial revolution (Lang, 1978) and beyond (Evans, 2007). Institutions recognisable to modern-day readers had commenced the provision of examinations in technical subjects by the 1870s, for example the Society of Arts, subsequently the Royal Society of Arts (RSA) and now ‘the R’ in the OCR awarding organisation (Watts, 2008) and the City and Guilds of London Institute (Lang, 1978; City & Guilds, 1993).

Vocational education and training (VET) and associated VQs have always seemed to be on a ‘parallel but separate track’ to academic education and examinations. Advocates of VQs enumerate their advantages, both in their own right and in contrast to academic qualifications:

> Despite its slow and at times haphazard development, the technical examination system in England has possessed a number of positive features. It offered real opportunities to students for entry and subsequent promotion in their chosen occupations. The examinations offered were more flexible than their school and university counterparts, matching the wide range of crafts, trades, vocations and occupational sectors involved. In addition to written examinations, assessments of practical activity were undertaken in special workshops or science laboratories. Teachers, employers and other key players were more closely

> involved, with some examinations set by teachers themselves and externally moderated. (Evans, 2008, p. 13)

Almost all historians of qualifications (eg Tattersall, 2007) agree that qualifications have been subject to increased centralisation and government regulation over the past 30 or so years. In the mid-1980s, a suite of National Vocational Qualifications (NVQs) was instituted following a review of VQs in England and Wales, and the National Council for Vocational Qualifications (NCVQ) was established to police the new national system (Evans, 2008, p. 12: Bees & Swords, 1990). In recent years, the Qualifications & Credit Framework (QCF – see Ofqual, 2008a) has largely superseded the former National Qualifications Framework (NQF) as a national repository for NVQs and other VQs (Ofqual, 2011b, p. 3).

Some influential commentators have been highly critical of increased centralisation; for instance, Wolf writes of ‘the sclerotic, expensive, centralised and over-detailed approach that has been the hallmark of the last two decades’ (Wolf, 2011, p. 21), whilst Oates has challenged the thesis that the QCF can be (or indeed needs to be) a bringer of ‘coherence’ (Oates, 2007, p. 13).

The increased centralisation of recent years is more or less universally accepted as a matter of fact. Equally, the influence of critiques of such centralisation is worthy of note. However, one must also hold in mind that government has always had a role in the provision of technical examinations; the ‘Department of Science and Art’ (a precursor of modern education Ministries[2]) provided technical exams from the 1850s (Evans, 2008, p. 16) and a history of City & Guilds has chapter headings such as ‘government constraints, 1919 – 1933’ (City & Guilds, 1993, p. 63) and ‘growing government interest, 1944 – 1964’ (*ibid.* at p. 88).

Thus, at the end of this section, two complementary thoughts are proposed:

- The practice and traditions of VQs are long established and consonant with high quality provision. These traditions pre-date recent government initiatives such as NVQ and QCF.

[2] With their various nomenclatures.

- Government control of VQs is widely acknowledged, and criticisms of this have influenced policy makers. However, it would be wrong to imagine that there was a pre-lapsarian era in which government intervention was not an issue and everything in the vocational garden was rosy.

The current status of VQs

Millions of people achieve VQs every year. Ofqual's latest 'qualifications market report' shows total qualification achievements by type for the year 2009/10 as follows:

• NVQ	979,000
• Vocationally-related qualifications (VRQs)	2,607,300
• QCF	771,300

(Ofqual, 2011b, pp. 60 – 63)

Very high stakes can attach to VQs; both for the qualification holders as individuals and for society more broadly. For example, City & Guilds offers a level 2 NVQ in Domestic Natural Gas Installation and Maintenance. Operatives who hold this NVQ will be eligible to apply to become a member of one of the United Kingdom's Gas Registration Bodies without the need to undertake further independent assessments in the areas covered by the NVQ. City & Guilds NPTC Level 2 Award in the Safe Use of Pesticides (QCF) is a legal requirement by the Chemical Regulatory Directorate for anyone applying pesticides on a commercial basis. City & Guilds keeps the National Register of Sprayer Operators, which facilitates Continuing Professional Development to ensure ongoing training.

VQs can use different assessment methods to academic qualifications. As noted above, VQ assessment has long involved practical tasks and teacher/tutor assessment, and features such as 'on-demand testing' (Wheadon et al, 2009) are commonplace. Some VQ awarding organisations have also preceded unitary awarding bodies in adopting innovative assessment approaches. For example, a senior Qualifications & Curriculum Authority officer said the following in 2004:

> City & Guilds … are a leading light in terms of showing us and showing other awarding bodies the ways in which technology objectives should be set and met at a corporate level. So City & Guilds have corporate objectives to hit in terms of migrating their business into electronic learning, e-learning and e-assessment by 2007. One of the very few organisations that at corporate level has those sort of objectives set already. (Ripley, 2004, p. 2)

When Wheadon and colleagues surveyed awarding organisations in 2009 on their use of on-demand e-assessment, they found that City & Guilds had approximately 900 on-demand e-assessments, the largest number of any surveyed organisation (Wheadon et al, 2009, p. 41). City & Guilds' on-demand e-assessments are generally facilitated by item banks – databases holding test questions. There are generally several questions addressing each Assessment Criterion (AC) within the unit that is being tested. Whilst every test taker receives a test of the same (sometimes very prescriptive) specification, (which typically assesses all ACs) each test will be assembled from different items selected from the bank.

The VQ system is complex. As noted above, this complexity has been criticised. But it can be countered that a certain amount of complexity is inevitable in a system that describes the competences and knowledge required for the very broad range of occupations and professions that exist in the modern economy. It is unsurprising, for instance, that a layperson might find some of the intricacies of gas fitting or commercial pesticide spraying qualifications inaccessible.

The intended benefits of the QCF have been stated as follows:

> The aim of the QCF is to support the establishment, maintenance and continuing development of a qualifications system that is:
>
> - inclusive – able to recognise the achievements of all learners at any level and in any area of learning
> - responsive – enabling individuals and employers to establish routes to achievement that are appropriate to their needs, and recognised organisations to develop units and qualifications in response to demand
> - accessible – building a system based on clear design features that are easy for all users to understand
> - non-bureaucratic – based on mutual trust and confidence, supported by a robust and proportionate approach to regulation and quality assurance. (Ofqual, 2008a, p. 5)

However, as noted above, the alleged benefits of the QCF have been challenged. Specifically, although NVQs can exist within the QCF (Ofqual, 2008b), the QCF has tended to muddy the distinction between workplace-based (and assessed) qualifications (NVQs) and mainly-college-based (and assessed) VRQs (Ofqual, 2011b, p. 26). In addition, while work-based qualifications are assessed primarily on the basis of performance evidence (observations of performance and/or artefacts produced during work), typically on multiple occasions stressing different environmental aspects, VRQs are assessed primarily on the basis of knowledge.

These different contexts and approaches to assessment have a negative impact for reliability investigation, by obscuring the (likely differing) psychometric assumptions that underlie different qualification types.

Other features of the QCF further inhibit reliability research. The QCF aim of providing flexible qualifications that can be built up unit-by-unit has required every unit to be freestanding and capable of being combined with other units so as to form a qualification. This poses problems for determining what the proper value of a reliability coefficient should be; if a unit can be part of an award (worth 1 to 12 credits), a certificate (worth 13 to 36 credits) or a diploma (worth 37 credits or more) (Ofqual, 2008a, p. 6), how can a researcher interpret whether the derived value of a unit's reliability index is sufficient? Further, Harth & Hemker (2011, p. 13) describe the complex manner in which QCF rules of combination implement conjunctive, compensatory and complementary approaches. It seems likely that the interaction of such varying combination routines with the (already complex) techniques for estimating the reliability of composite scores (He, 2010) was not considered by QCF designers.

VQ units and qualifications are housed in databases. Some of these are owned by Ofqual, whereas others are private systems within awarding organisations. The existence of several databases and the lack of commonality between them make it difficult to build taxonomies and estimate reliability. These databases, the information they hold and the associated issues for reliability will be discussed in more detail in the 'Taxonomies' section of this report (below at p. 32).

Key concepts in the assessment of VQs

Competence

A government review of VQs published in 1986 alleged significant faults with existing provision (Jessup, 1991, p. 10), including confusion amongst employers regarding the meaning and value of the then plethora of different vocational awards, and the failure of vocational awards to 'guarantee' occupational competence (Barnes, 1992). The NVQ system was instigated to remedy such perceived faults. The following was an early definition of an NVQ:

> [An NVQ is …] a statement of competence clearly relevant to work and intended to facilitate entry into, or progression in, employment and further learning, issued to an individual by a recognised awarding body.
>
> The statement of competence should incorporate specified standards in:
>
> - the ability to perform in a range of work-related activities, and
> - the underpinning skills, knowledge and understanding required for performance in employment (Jessup, 1991, p. 15)

There is a large literature on meanings and implications of the term ‘competence’, a thorough review of which is beyond the scope of the current research. Rather, in the following section, salient points that have implications for reliability analysis are highlighted.

The term ‘competence’ is ‘compelling in its common sense and rhetorical force’ (Norris, 1991, p. 331). However, it is also a notion of considerable complexity. Jessup’s definition (cited above) makes plain that two elements are relevant to competence: performance in work, and underpinning skills, knowledge and understanding. Some who write about these two facets of competence tend to assert the benefits of one by emphasising the dis-benefits of the other. For example, the benefits of training for work-place performance is sometimes argued by characterising de-contextualised teaching and learning (learning outside the workplace) as lacking practical applicability, and the ability to motivate disengaged learners (eg Jessup, 1991, p. 10). Conversely, some critics regard competence-based training that is highly targeted on work tasks as reductive, narrow and not assisting learners to develop skills that can be transferred beyond the current role (Norris, 1991, p. 335) nor to develop a coherent body of knowledge. Both types of arguments are somewhat stereotypical; a more sensible position is that competence-based education and training needs to be **both** practical, engaging and performance-related **and** have sufficient knowledge elements so that learners can understand the wider context and transfer what they have learned.

There are substantial tracts discussing the philosophical, psychological and educational implications of the concept of competence (eg Gonczi, 2001; Koeppen et al, 2008; QCDA, 2010). However, the psycho*metric* implications of the definition of competence as used in English VQ systems have not been explicitly and fully developed. Some researchers, for example Meretoja et al (2004), reporting research from Finland, have applied conventional psychometric approaches to measures of professional competence (in their case nursing).

However, the definitions of competence operationalised in VQs in England are likely to inhibit the straightforward transfer of psychometric models used in other contexts.

Firstly, competence in English VQs is established by assessor judgement. The assessor judges the candidates as either 'competent' or 'not yet competent' (Harth & van Rijn, 2010, p. 8). In theory, such judgements are absolute; the candidate either is or is not (yet) competent, there is no place for 'partial credit' (*ibid.*, and Harth, personal communication). Further, the judgements are categorical rather than scores, and it flows from this both that they do not have distribution or spread and that there are no 'wrong answers'. Thus, it is more principled to consider competence to be either 'present' or 'absent' and hence for a data file to contain 1 and – (or some other indicator of 'null' or 'not present') rather than 1 and 0 (corresponding to 'right' and 'wrong', respectively).

For knowledge assessment in NVQs, oral or written questions may be used (including multiple-choice questions)[3]. Harth & Hemker (2011, p. 14) state that, in principle, any such knowledge assessments ought to have a pass mark of 100 per cent. In practice, knowledge-test pass marks are generally lower than 100 per cent, but they are still higher than they would be for many tests in academic qualifications. Also, the purpose of such knowledge tests is not to array candidates relative to each other along the whole scoring scale, but rather to ascertain whether candidates have mastered the relevant knowledge. Often scores will be restricted to a limited part of the scoring scale for a particular test (typically the higher end). These features suggest that the spread of scores in VQ knowledge tests is likely to be relatively constrained when compared to other forms of knowledge assessment, such as academic examinations. This will have implications for reliability estimation.

Competence has been used in various European VET systems; and various nuanced interpretations of the concept have been elucidated by Brockmann, Clarke & Winch (2011). However, Tight (2002, p. 132) has suggested that the term 'competence' as understood in English VQs has roots in US research from the 1970s. As such, competence and its assessment have close conceptual links to the notion of criterion-referenced testing (CRT) that was widely discussed in America in the 70s. CRT is the subject of the next section.

[3] Chapter five of the companion to the current report (Johnson et al, 2012) gives a brief history of knowledge testing within VQs in England.

Criterion referencing

As well as having an American literature, CRT has been discussed in the UK, especially in the context of grading and comparability (Robinson, 2007, pp. 115 – 118). But – perhaps because of its US origins – a crucial advantage of CRT as opposed to competence-based assessment is the extensive discussion of psychometrics that has been conducted, leading to the resolution of many issues. The major weakness of CRT as a framework within which to understand English VQs is that not all English VQs are explicitly identified or described as CRTs. The mitigation of that weakness – which applies particularly to knowledge tests in VQs – is that the assessments exhibit properties which, as a matter of fact, resemble features of CRTs, even if their designers have declined or neglected to name them thus.

A recent and authoritative work defines criterion referencing as follows:

> Criterion-referenced interpretations characterize an examinee on the basis of a test performance without reference to the performance of other individuals. In theory, the interpretation is determined by the absolute level of the examinee's score, without reference to the performance of anyone else. (Haertel, 2006, pp. 66 – 67)

Osterlind suggests that CRTs should have the following features:

a) clearly defined performance standards for measurement

b) test items constructed specifically to address the intended performance standards

c) scores that can be interpreted in terms of an individual's achievement of the specified performance standards. (Osterlind, 1988, pp. 87 – 87)

As argued above, although tests in VQs are not often explicitly referred to as CRTs, they do exhibit all the features in the definitions above. VQ units contain Learning Outcomes (LOs) and Assessment Criteria (ACs) which are based on National Occupational Standards (NOS). In turn, the VQ tests to assess the knowledge and/or performance for those units are based on LOs and ACs (Harth & Hemker, 2011, pp. 10 – 12). Assessment results can be interpreted in terms of qualifications holders' competence in respect of the NOS; this is especially the case where VQs are high-stakes and function as licences or quasi-licences to practise.

The data sets generated from CRTs have particular features that are different to those that pertain to data sets generated from NRTs. The distribution of total scores in a CRT will tend to be narrower than the distribution of total scores from an NRT (Hambleton et al, 1975, p. 9). It is also possible that CRT score distributions might only occupy a small part of the scoring scale for a test (typically the top end, if the pass mark is high). However, this may be a design feature of a CRT, and not necessarily a matter of concern to the test developers, even though it might be a factor that challenges reliability estimators (cf. Johnson & Johnson, 2011).

The fact that CRT applications have to sample all areas of a given domain may be a challenge for measurement models that make strong assumptions that test items address a undimensional latent trait (Wainer et al, 1990; Sands, Waters & McBride, 1997).

Reliability

Definition of the concept and state of the art

Ofqual has defined reliability as follows:

> Reliability refers to the consistency of outcomes that would be observed from an assessment process were it to be repeated. High reliability means that broadly the same outcomes would arise. A range of factors that exist in the assessment process can introduce unreliability into assessment results. Given the general parameters and controls that have been established for an assessment process – including test specification, administration conditions, approach to marking, linking design and so on – (un)reliability concerns the impact of the particular details that do happen to vary from one assessment to the next for whatever reason.
>
> Sources of error include:
>
> - occasion-related (e.g. if assessed on another day, the student might have been less tired)
> - test-related (e.g. if a different test had been set, the student might not have been disadvantaged by the unfamiliar wording of one essay title or advantaged by the topic of another)
> - marking-related (e.g. if a different marker had been assigned, the student might have been marked down for using unusual stylistic constructions)
> - grading-related (e.g. if a different team of people had been involved in the process of level setting, different thresholds might have been set). (Ofqual, 2011a, p. 4; see also, Newton, 2009, p. 183)

There is an extensive literature elucidating the concept of reliability, methods for estimating its extent, and necessary considerations in interpreting such estimations. Chapters in successive editions of the American textbook *Educational Measurement* provide canonical treatments of the state of the art in reliability. Successive chapters include: Stanley (1971), Feldt & Brennan (1989) and Haertel (2006). The outputs of Ofqual's reliability programme define the state of the art in reliability conceptualisation and empirical estimation in England in 2011. Reports can be found at: www.ofqual.gov.uk/standards/reliability.

Previous studies of VQ reliability

From over 20 reports on the Ofqual reliability micro-site, only one is based on VQs. A recent review of VQ research took the view that 'there are significant gaps in recent reliability research in the UK' (Johnson, 2006, p. 37).

Murphy and colleagues (1995) wrote a report for the NCVQ. They reviewed then existing reliability literature, and brought together 31 assessors to review portfolios and discuss the extent of agreement in judgements of competence. They also observed 15 assessed tasks and investigated differences of opinion between assessors of such tasks. They described (in words) the extent of agreement and causes of disagreement for the different portfolio assessments but did not derive any reliability indices (Murphy et al, 1995, pp. 19 – 20). They also suggested further work that could be carried out to investigate VQ reliability (*ibid.* at p. 28).

Greatorex (2005) reported analysis of simulated candidate evidence (*n* = 15) on two NVQ units. This analysis returned Cronbach's alpha internal consistency values in excess of 0.9 in both NVQ units (*ibid.* at p. 155). Greatorex suggested that the NVQ system was capable of sustaining reliable outcomes, but also proposed that further work was necessary on the issue of inter-rater agreement (*ibid.* at p. 161).

Harth & Hemker (2011) conducted research in Ofqual's reliability programme. Their detailed discussion of the context and assumptions of VQs (Harth & Hemker, 2011, pp. 10 – 16) enumerated rationales for VQ assessment and pointed out the conceptions of quality that underpin them. They went on to devise an approach to reliability estimation that was appropriate for the various described approaches to VQ assessment (*ibid.* at pp. 17 – 22). In particular, they proposed the use of: Cronbach's α (alpha) and Guttman's λ_{-2} (lambda-2) coefficients to estimate decision consistency (*ibid.* at pp. 19 – 20, and see below at p. 19); and Gower's coefficient and Cohen's kappa ($_C\kappa$) as estimates of assessor agreement (*ibid.* at pp. 20 – 22, and see below at p. 18).

They summarised their findings with respect to the three qualifications that they analysed as follows:

> … inter-assessor (rater) agreement is ‘high’ (Gower coefficient ranging from .90 to .99) and inter-rater (assessor/IV) reliability (Cohen’s kappa) is ‘substantial’ for the Electrotechnical pathway (kappa > .75) and ‘almost perfect’ for the hairdressing qualifications (kappa > .95).
>
> …
>
> The data available from the Electrotechnical Services pathway allowed us to estimate the internal consistency of decisions by estimating a coefficient similar to Cronbach’s alpha by means of considering units as items. In this case, reliability estimates had values larger than .95, considered to be very high, especially in the context presented by vocational assessment. (Harth & Hemker, 2011, p. 37)

Harth & Hemker (*ibid*.) make various suggestions for further work, and in general demonstrate that it is feasible to analyse VQ reliability and that such analyses can demonstrate highly reliable assessment. The current work seeks to build on the foundation established by Harth & Hemker; we address a different part of the VQ portfolio (knowledge tests rather than judgements of competence) and draw out some implications about the nature of data sets and suitable coefficients to use given the inferences that certain VQs sustain.

Factors that affect values on reliability indices

Now, we move on to consider the factors that can impact upon indices of reliability. Reliability can be expressed as any of:

- the squared correlation between true score and observed score
- the correlation between observed scores on two equivalent tests
- the ratio of true score variance to total variance
- the ratio of true score variance to the sum of true score variance and error variance.

Ofqual’s definition (as cited above at p. 13) puts replication at the heart of the conception of reliability while the focus in the cited extract from the regulator’s specification of requirements for the current project (above at p. 4) speaks of ‘internal reliability’. This latter focus suggests strongly that features of the ‘internal structure of tests’ (to use the sub-title of Cronbach’s famous 1951 article) – or at least the structure of the data sets generated in response to tests – is the crucial factor in understanding reliability indices. This includes some features of data sets that might not perhaps spring to mind immediately, if one was thinking only in terms of the proportion of total variance that is true score variance. Thus, in the following section, we review some of the structures of test-generated data sets which can impact on the values of reliability coefficients.

The first aspect of data set structure that can impact on reliability is the *spread* of scores. When a reliability index seeks to convey the extent to which test takers' scores can be distinguished from each other (a classic NRT aim, see above), then a test will appear more reliable if the scores of candidates have a larger spread, i.e. variance. This will be true even if the amount of error variance is constant (Traub & Rowley, 1991, pp. 177 – 178).

The extent to which test items address *the same ability* can affect reliability. If all items in a test address the same or similar abilities, then the reliability index is likely to be higher. This is true for those indices in which covariance between items constitutes a term in the formula, and also for those indices that are based on a measurement model that requires measures to be arrayed on an underlying, latent trait (see below at p. 21).

Any researcher using statistical techniques needs to be sensitive to differences between populations and samples (Gravetter & Wallnau, 1996, p. 201). Further, the reader will note that the reliability formulae set out below are mostly expressed in terms of population parameters. This is a challenge for the research in that observed test scores upon which calculations are performed are sample statistics drawn from a larger population. As such, they are potentially biased estimates of population parameters.

Having said that, Cronbach & Shavelson (2004, p. 414)'s remark seems relevant:

> … a naïve response would be to say that if [random sampling] assumptions are violated, the Alpha calculations[4] cannot be used. No statistical work would be possible, however, without making assumptions and so long as the assumptions are not obviously grossly inappropriate to the data, the statistics calculated are used, if only because they can provide a definite result that replaces a hand-waving interpretation.

Further, Bramley & Dhawan have commented that indices such as Cronbach's alpha are – for all their weaknesses – 'a *de facto* standard' (2010, p. 10). As such they are worth calculating in the current project. However, the impact of failing to deal fully with sample bias may be to limit the generalisability of findings. This issue is particularly acute in the

[4] See below, equation 2.

case of CRT, because while some indices may provide perfectly adequate results to convey the extent to which tests reliably perform NRT tasks (essentially rank ordering candidates), their use for more absolute interpretations may be less defensible. For that reason, reliability approaches that contain explicit corrections for sampling bias are to be preferred in the case of CRT.

The final point in this section echoes the observations of Hutchison & Benton (2009, p. 4) regarding different measurement philosophies. Reliability indices derived from different traditions can give different results. Also, the complexity of the formulae used to convey indices can obscure important matters of principle. Finally, researchers differ in the extent to which they would 'mix and match' reliability approaches from different traditions; some would be prepared to take aspects from across the range of reliability techniques, whereas others would stick firmly within their own paradigm. At the very least, researchers – whichever and however many tradition(s) they take their reliability indices from – should state the assumptions underlying the reliability analyses that they use and any limitations of the inferences that can therefore be made.

Types of reliability measure

We have reached a somewhat paradoxical juncture; the specification of requirements for this project refers to 'internal reliability' – a term that implies strongly that reliability indices should be generated to describe the structure of data sets generated in response to tests. Yet we also know that data sets from CRTs tend to have different features from those from NRTs and that what is 'good' in a data set distribution for an NRT is not necessarily good for a CRT. We also know that in CRT the categorisation of a candidate as 'competent' or 'not yet competent' (master or non-master of the topic) is crucial.

For this reason it is important to note that different sources have categorised reliability approaches for CRTs into three basic types.

Internal consistency and dependability: within classical test theory, internal consistency reliability measures use a range of approaches in order to model the administration of a parallel test, and to assess the extent to which results would differ if such a test were administered. Internal consistency indices typically draw conclusions from analyses of total test scores, and/or sub-divisions thereof, including item scores (Hutchison & Benton, 2009, pp. 22 – 23) or other sub-divisions of total test scores such as scores on latent trait scales (*ibid.* at pp. 28 – 29). Internal consistency analyses do not, of themselves, say anything about the extent of misclassification (people truly competent who are classified as not competent or vice versa) (Newton, 2009, p. 185).

Generalizability theory provides co-efficients of 'relative' and 'absolute' measurement (Johnson & Johnson, 2010, p. 23). The coefficient of absolute measurement 'provide[s] a handy way of estimating the overall dependability of the scores on a CRT without reference to a cut-score' (Brown, 1990, p. 90).

So-called '*threshold-loss indices'* are an alternative to internal consistency and/or dependability coefficients that provide information on the extent of misclassification (Brown, 1990, p. 81; Traub & Rowley, 1980, pp. 526 – 529). Gower's coefficient and Cohen's kappa as used by Harth & Hemker (see above at p. 14) are examples of such indices. An indication of the extent of misclassification is of course crucial in an analysis of the reliability of CRTs. However, whilst threshold-loss indices do give information on the extent of misclassification, they do not distinguish between misclassifications; any misclassification – however gross – is treated equally. Also, threshold-loss indices give no indication of the internal structure of the data set generated by the studied test, and hence do not comply with the specification of requirements for this project.

In contrast, *squared-error loss agreement* approaches 'take into account the distances of students' scores from the cut-point, i.e., the degree of mastery or non-mastery rather than just the simple categorization' (Brown, 1990, p. 88). Thus, they afford the possibility of modelling the extent of misclassification **and** they could be conceived as fitting the 'internal reliability' definition set out in the specification of requirements.

Examples of indices and the replications that they quantify

The main features of English VQ assessment systems and the considerations affecting reliability estimation have been set out in the previous sections. In the section that follows, we define formally[5] the reliability indices that we intend to employ and comment briefly on the considerations that affect the implementation of each index.

[5] In writing the indices we attempt to balance between amending indices to provide consistency, and keeping as closely as possible to source notation. In all cases, we cite the sources for our indices, so that an interested reader could follow up. We state the meaning of symbols under every formula – even when this is repetitious.

Internal consistency and dependability indices

Kuder & Richardson (1937) derived several formulae to estimate the internal consistency of test data sets. Their equation 21 (known as KR-21) is relevant for the current project:

$$ {}_{21}\rho_{xx'} = \left(\frac{n}{n-1}\right)\left(1 - \frac{\mu_x(n-\mu_x)}{n\sigma^2{}_x}\right) \qquad (1) $$

where:
${}_{21}\rho_{xx'}$ is the Kuder-Richardson 21 reliability co-efficient for test *x*
n is the number of items in the test
μ_x is the raw (population) mean score for test *x*
and $\sigma^2{}_x$ is the (population) variance of scores on test *x*

KR-21 is relatively little used in modern-day reliability research, but its benefit for our purposes is that it is possible to calculate it using total test scores, rather than item scores (Hutchison & Benton, 2009, p. 24). Thus it has possibilities in cases where different candidates receive different items. However, any such use of KR-21 would come with the 'health warning' that the derived index does not model any unreliability deriving from the candidates receiving different items.

KR-21 has largely been superseded in reliability estimation by Cronbach's alpha (Cronbach, 1951, p. 299). The formula for alpha is:

$$ {}_{\alpha}\rho_{xx'} = \left(\frac{n}{n-1}\right)\left(1 - \frac{\sum_{i\neq j}\sigma_{ij}^2}{\sigma^2{}_x}\right) \qquad (2) $$

where:
${}_{\alpha}\rho_{xx'}$ is the Cronbach's alpha reliability co-efficient for test *x*
n is the number of items in the test
σ_{ij}^2 is the (population) covariance of scores for items *i* and *j*
and $\sigma^2{}_x$ is the (population) variance of scores for test *x*

Cronbach's α is an extremely widely used index (Sijtsma, 2009, p. 108). However, that use is controversial. Cronbach himself (Cronbach & Shavelson, 2004) disavowed the α coefficient towards the end of his life in favour of the standard error of measurement[6]. Sijtsma (2009) has stated that α is often a 'gross underestimate' of the true reliability,

[6] Cronbach (1972) was also one of the 'founding fathers' of generalizability theory.

and Harth & Hemker (2011, p. 20) simplifying a formula from Sijtsma's paper, suggested the following relationship:

$$_{\alpha}\rho_{xx'} \leq {}_{\lambda_{-2}}\rho_{xx'} \leq \rho_{xx'} \qquad (3)^{7}$$

where:

$_{\alpha}\rho_{xx'}$ is the Cronbach's alpha reliability co-efficient for test *x*

$_{\lambda_{-2}}\rho_{xx'}$ is the Guttman's lambda-2 alpha reliability co-efficient for test *x, and*

$\rho_{xx'}$ is the 'true' reliability for test *x*

As an internal consistency coefficient, α gives no information on the extent of misclassifications. Secondly, as an index from the NRT tradition, α will tend to be higher in those instances where scores are spread along the length of a scoring scale, and may give misleading results if scores are 'bunched' in one part of the scale. Thirdly, as an index from classical test theory (CTT), α is subject to the assumption of strictly parallel forms. Strictly parallel forms have four properties:

- Identical test specifications
- Identical observed score distributions when administered to any (indefinitely large) population of examinees
- Tests which covary equally with one another, and
- Covary equally with some other measure, where that other measure is a measure of the same construct

(Haertel, 2006, p. 69)

It is likely that these assumptions do not always hold in CRT.

Finally, as α depends in part on the variance of item scores, it is only possible to calculate its value in cases where all candidates receive the same items.

According to Sijtsma (2009), λ_{-2} gives a better estimate of the true reliability than α , and so it may be appropriate to use it for the current project. Guttman (1945) derived six reliability indices and Callender & Osburn (1979, p. 89) give the following interpretation of λ_{-2} by showing its relationship to λ_{-1}:

[7] However, Raykov & Marcoulides (2011) suggest that this relationship does not hold.

$$ {}_{\lambda_{-1}}\rho_{xx'} = 1 - \frac{\sum_{i=1}^{n} \sigma_i^2}{\sigma_x^2} $$

and (4)

$$ {}_{\lambda_{-2}}\rho_{xx'} = {}_{\lambda_{-1}}\rho_{xx'} + \frac{\sqrt{\frac{n}{n-1}\Gamma_2}}{\sigma_x^2} $$

where:
n is the number of items in the test
σ_x^2 is the variance of scores on test *x*
σ_i^2 is the variance of a single item *i*
Γ_2 is the sum of the squares of the covariances between items, a sum which includes *n(n-1)* terms
${}_{\lambda_{-1}}\rho_{xx'}$ is the Guttman's lambda-1 alpha reliability co-efficient for test *x*
${}_{\lambda_{-2}}\rho_{xx'}$ is the Guttman's lambda-2 alpha reliability co-efficient for test *x*

Item response theory (IRT) is a family of measurement models that has many applications in the evaluation (Embretson & Reise, 2000) and scoring (Thissen & Wainer, 2001) of tests. The IRT model containing the fewest parameters is the Rasch model. Rasch model adherents emphasise its supposed elegance, parsimony and closeness to the 'essence' of measurement (eg Andrich, 2004). However, Rasch modelling/theory[8] has been controversial in the United Kingdom[9] (Panayides, Robinson, & Tymms, 2010).

IRT (including Rasch) practitioners tend to think of reliability in different terms to those discussed so far. The Test Information Function (TIF) is a frequently-used indicator of measurement quality (Hutchison & Benton, 2009, p. 28). This has the benefit of showing the magnitude of measurement information, and of showing its location on an ability/difficulty (trait) scale (*ibid*.). However, TIFs arguably do not constitute a reliability

[8] The nomenclature that commentators use tends to hint at their attitude to Rasch.
[9] Controversies around the alleged mis-use of Rasch have not typically involved disputes about Rasch practitioners' conceptualisation of reliability. Critics of Rasch tend to allege that results of Rasch modelling have been used even when data mis-fit the model. There has also been concern about some of the strong claims made by Rasch practitioners – for instance, that it can provide 'sample independent measurement'. The different approach to reliability is a curiosity rather than a cause of dispute.

measure within Ofqual's definition, since they make no direct statement about replication.

However, Rasch practitioners have developed several analogues to conventional internal consistency coefficients which may be useful in this project. Wright (1996) defines a 'separation index' such that

> … separation is the number of statistically different performance strata that the test can identify in the sample. This can be pictured by placing an error distribution in each stratum. A separation of '2' implies that only two levels of performance can be consistently identified by the test for samples like the one tested. (Wright, 1996, p. 472)

Linacre has used the separation index to develop an analogue to an internal consistency index within the Winsteps program (Linacre, 2009). This index is notated as follows by Hutchison & Benton (2009, p. 63):

$$SEP = \frac{SDT}{SE}$$

and (5)

$$REL = \frac{SEP^2}{1 + SEP^2}$$

where
SDT is the expected true SD of the target sample
SE is the mean test standard error of measurement
SEP is the separation index
REL is the person separation reliability index.

SEP (and hence *REL*) are reliability indices calculated from Rasch trait estimates. Some Rasch theorists (Andrich, 1982; Linacre, 1997) contrast Rasch reliability indices with 'raw score reliability indices' such as α. Trait estimates essentially amount to 'adjusted scores' for the case where candidates receive items of differing difficulties. Also, IRT has always been able to handle missing data matrices (witness IRT's prominent role in Computer Adaptive Testing). Further, the *REL* index functions even in the situation where different candidates receive different items. Thus, this appears to have potential for the analysis of randomly-generated tests (where different candidates receive different questions from an item bank) in the current project.

Rasch models make strong assumptions about data; indeed they have been said to be 'rigid' (Johnson & Johnson, 2010, p. 19). Further, it is often the case that researchers will write tests with the aim of fitting the Rasch model, rather than the other way around. Also, as we have seen above (p. 14), VQs are written to assess all of a qualification's LOs and AC and thus are not designed to array candidates in respect of a single latent trait. Thus, a necessary condition for the use of a Rasch-based reliability index would be the calculation of a statistic to show the extent to which scores fit the Rasch model. The Rasch Measurement Software 'Winsteps' provides the following guide to interpreting the Winsteps misfit statistic:

Misfit statistics	**Interpretation**
>2.0	Distorts or degrades the measurement system.
1.5 – 2.0	Unproductive for construction of measurement, but not degrading.
0.5 – 1.5	Productive for measurement.
<0.5	Less productive for measurement, but not degrading. May produce misleadingly good reliabilities and separations.

Table 1: Interpretation of Rasch model parameter-level mean-square fit statistics

(Linacre, 2009, p. 444)

Generalizability theory (Cronbach et al, 1972; Brennan, 2001; Cardinet, Johnson & Pini, 2010) affords several possibilities. The (relative) generalizability co-efficient can be noted as follows:

$$E\rho^2 = \frac{\sigma^2(p)}{\sigma^2(p) + \sigma^2(\delta)} \qquad (6)$$

where

$E\rho^2$ is the (relative) generalizability coefficient

$\sigma^2(p)$ is the universe score[10] variance, and

$\sigma^2(\delta)$ is the relative error variance

(Brennan, 2003, p. 11)

When data are arranged in a two-factor 'crossed design' (see below), $E\rho^2$ gives equivalent results to Cronbach's α (*ibid.* at p. 12). Whilst $E\rho^2$ is computed using relative error variance (a concept most suitable for NRT applications), generalizability

[10] Where 'universe score' is the concept in G-theory most analogous to 'true score' in CTT.

theory also provides coefficients that are based on absolute error variance. The dependability coefficient can be noted as follows:

$$\Phi = \frac{\sigma^2(p)}{\sigma^2(p) + \sigma^2(\Delta)} \qquad (7)$$

where
Φ is the (absolute) dependability coefficient
$\sigma^2(p)$ is the universe score variance, and
$\sigma^2(\Delta)$ is the absolute error variance
(Brennan, 2003, p. 12)

In practice, for the simplest assessment model, in which candidates take the same test, the absolute error variance is the sum of the item variance and the residual variance (which is confounded with the candidate-item interaction variance), both terms divided by the number of items. The benefit of this 'absolute' coefficient in principle is that it is not dependent on the CTT assumption of strictly parallel forms, and thus is more appropriate for the case of CRT measurement.

For those indices that take into account item variances (α , $E\rho^2$ and Φ) it has been implicit thus far that data will be set up in a matrix or 'crossed' format (noted as *i* x *c* – items crossed with candidates). This scenario is the most efficient for extracting the maximum amount of measurement information about both candidates and items. Generalizability theory also addresses situations in which facets are 'nested' within each other (Cardinet, Johnson & Pini, 2010, pp. 13 – 14). This is the case, for example, when candidates receive different items from each other (see above, p. 8),which can be conceptualised as items 'nested within' candidates (noted as *i* : *c*)[11].

Squared error loss indices

Since internal consistency, generalizability or dependability indices do not give information on the extent of misclassification, one or more squared-error loss agreement

[11] Some other approaches may also be suitable to derive indices of internal consistency or dependability for the case where different candidates receive different items. Stanley (1971, pp. 425 – 428) has developed an ANOVA-based reliability co-efficient for the case where different candidates receive different items. That index is essentially an internal consistency coefficient; however, it is plausible to imagine an extension of Stanley's (1971) coefficient to include (mean – cut score) terms (following the logics of Livingston (1972) and Brennan & Kane (1977)). Also, Haertel (2006, p. 97) discusses how to estimate variance components in the face of unbalanced or missing data.

indices are necessary. Livingston's index (${}_{L}K^{2}$) (Livingston, 1972) was the earliest squared-error loss agreement index to be developed. The following notation is based on that of Traub & Rowley (1980, pp. 536 – 537)[12]:

$$ {}_{L}K^{2}(x,\tau)=\frac{\rho_{xx'}\sigma_{x}{}^{2}+(\mu_{x}-c)^{2}}{\sigma_{x}{}^{2}+(\mu_{x}-c)^{2}} \qquad (8) $$

where
${}_{L}K^{2}(x,\tau)$ is Livingston's *K* (squared) index
x is 'the realization of a true or latent measurement' plus measurement error (Traub & Rowley, 1980, p. 524) – i.e. observed scores.
τ is the 'true or latent measurement' (*ibid.*) – i.e. 'true score'
$\rho_{xx'}$ is some NRT internal consistency reliability estimate
c is a pre-determined cut score, and
μ_{x} is the mean observed score on the test
$\sigma_{x}{}^{2}$ is the variance of observed scores on the test

The following is a logical consequence of the definition of ${}_{L}K^{2}$:

$$ \rho_{xx'} \leq {}_{L}K^{2} \qquad (9) $$

where
${}_{L}K^{2}$ is Livingston's index, and
$\rho_{xx'}$ is some NRT internal consistency reliability estimate

Further, if the mean score is equal to the cut score, then the squared error loss index reduces to the internal consistency index (Traub & Rowley, 1980, p. 537).

In ${}_{L}K^{2}$, $\rho_{xx'}$ is **some** NRT internal consistency reliability estimate. Most often α has been the internal consistency that has been 'plugged into' ${}_{L}K^{2}$. However, ${}_{L}K^{2}$ is written sufficiently generally that some other internal consistency/NRT-derived index could be substituted. This may be useful for cases where different candidates receive different items, and this precludes the use of certain indices (typically those using a

[12] See also Haertel (2006, p. 99) for an alternative version of the ${}_{L}K^{2}$ formula.

‘crossed design’ – see above). Values $\rho_{xx'}$ from KR-21 or the Rasch person separation index could be used as the value for $\rho_{xx'}$ in the equation above.

Given that the Rasch person separation index is based on trait measures rather than raw scores, whereas the KR-21 index is not, and that ${}_{L}K^{2}$ containing any internal consistency index is greater than or equal to an internal consistency coefficient, we may tentatively posit the following relationship:

$$\rho_{xx'} \leq {}_{L(REL)}K^{2} \leq {}_{L(KR-21)}K^{2} \qquad (10)$$

where
$\rho_{xx'}$ is some NRT internal consistency reliability estimate
${}_{L(REL)}K^{2}$ is Livingston’s index calculated using the Rasch person separation index
${}_{L(KR-21)}K^{2}$ is Livingston’s index calculated using the KR-21 internal consistency index

This approach of ‘plugging in’ NRT internal reliability co-efficients into Livingston’s index may be useful to researchers who do not – for some reason – have access to the phi (lambda) approach that is about to be described. It is acknowledged to be an imperfect solution, however. Firstly, KR-21 tends to be rarely used nowadays. Also, there are several reasons to challenge the use of the Rasch person separation index ‘inside’ Livingston’s index. These are:

- The Rasch person separation index is not the main way in which that measurement tradition conceives of reliability, and therefore the properties of that index have been less rigorously interrogated in research.
- We know of no published attempts to ‘plug’ the Rasch person separation index into Livingston’s index, and hence this would be a novel or idiosyncratic procedure.
- If the variance quantification within the Rasch person separation index were conducted using different assumptions to those used to quantify variance in other parts of Livingston’s formula (particularly the denominator), this would limit the interpretations that a user of this index could make.

Thus, the above approach is possible but less desirable than the phi (lambda) approach, which we describe now.

Brennan & Kane (1977) developed the phi (lambda) ($\Phi(\lambda)$) index. This is a squared-error loss agreement index with philosophical underpinnings from generalizability theory. This index has been said to be preferable to ${}_{L}K^{2}$ in the case of CRT since it requires fewer classical test theory assumptions (Brown, 1980, pp. 92 – 93). The formula for $\Phi(\lambda)$ is noted as follows by Haertel (2006, p. 99)[13]:

$$\Phi(\lambda) = \frac{\sigma^2(p) + (\mu - \lambda)^2}{\sigma^2(p) + (\mu - \lambda)^2 + \sigma^2(\Delta)} \qquad (11)$$

where
$\Phi(\lambda)$ is the (absolute) dependability coefficient for a given cut score, λ
λ is the cut score (expressed as a proportion of the total test score[14])
μ is the grand mean score
$\sigma^2(p)$ is the universe score variance, and
$\sigma^2(\Delta)$ is the absolute error variance[15]

$\Phi(\lambda)$ shares some features with ${}_{L}K^{2}$, but its calculation also includes some steps that make it more appropriate for CRT applications. When the mean score approaches the cut score, $\Phi(\lambda)$ approaches Φ. However, in estimating $(\mu - \lambda)^2$ from sample data, we need to make an adjustment, because $(\bar{X} - \lambda)^2$ is a biased estimate of $(\mu - \lambda)^2$, the amount of the bias being equal to $\sigma_{\bar{X}}^2$ (where $\bar{X}$ is the sample mean). Whilst this adjustment involves an extra calculation when compared to the ${}_{L}K^{2}$, it also means that $\Phi(\lambda)$ is a more principled treatment of sample-to-population generalisation.

The Standard Error of Measurement

Haertel defines the Standard Error of Measurement (SEM) as:

> … a number expressed in the same units as the corresponding test score [which] indicates the accuracy with which a single score approximates the expected value of possible scores for the same examinee. (Haertel, 2006, p. 65)

[13] See also Brennan (2001, p. 13).

[14] For binary data tests. If μ is the mean score then λ should be on the same metric as μ.

[15] We have used this notation for phi (lambda) in preference to that of the original article (Brennan & Kane, 1977, p. 281) on the grounds of parsimony and consistency with the earlier G-theory equations. The three indices are noted in the formulae as: $E\rho^2$, Φ and $\Phi(\lambda)$. But in the findings tables that follow, we will refer to them, respectively as: 'coeff_G', 'phi' and 'phi (lambda)'.

SEM is the square root of the measurement error variance (Haertel, 2006, p. 69). SEM is related to the reliability, and is often expressed in terms of ρ, as in this widely-used formula:

$$SEM = \sqrt{(1-\rho)\sigma_X^2} \qquad (12)$$

where
SEM is the Standard Error of Measurement for a total test score
ρ is the reliability coefficient
σ_X^2 is the variance of a set of test scores *X*
(Hutchison & Benton, 2009, p. 62)

However, SEM is logically distinct from the reliability co-efficient; it is the value of the reliability coefficient that depends on the value of measurement error and not the other way around. It is not necessary to calculate the reliability coefficient to get the SEM.

Many modern measurement software applications provide SEM estimates as a default output. Both 'Winsteps' and 'EduG'[16] – the software applications used in the current project – do so (see page 46 for description of these applications).

An SEM value can be used to calculate a confidence interval around an observed score. For example, under the central limit theorem (Gravetter & Wallnau, 1996, p. 204), one can infer that there is a 95% probability that the true score lies within 1.96 x SEM either side of the observed score (Harvill, 1991, p. 33). This is a useful property of the SEM, and it is not liable to provide misleading results in the case of widely or narrowly spread data (Hutchison & Benton, 2009, p. 26). However, in the current project, it would be a stretch to describe SEM as an 'internal reliability co-efficient' and thus it would be strictly speaking outside the scope of the project as specified.

[16] EduG provides an 'absolute SE', and one needs to divide that quantity by the number of items to get the SEM.

Index	Type of index	Strengths	Weaknesses/limitations
Kuder-Richardson 21 (KR-21)	Internal consistency	• Simple to compute using off-the shelf statistical packages. • Can be calculated from total scores alone. • Will work in the case of candidates receiving different items. • Could be inserted into Livingston's index for tests where different candidates receive different items.	• Somewhat archaic – rarely used in modern research. • Does not model inter-item unreliability in the case where different candidates receive different items.
Cronbach's alpha	Internal consistency	• Simple to compute using off-the shelf statistical packages. • Widely understood. • Models inter-item unreliability.	• May understate reliability when test score distribution is 'narrow' or 'bunched'. • Makes no statement about misclassification. • Criticised by authoritative articles – especially that it may underestimate reliability. • Needs all candidates to answer same questions to function.
Guttman's lambda-2	Internal consistency	• Simple to compute using off-the shelf statistical packages. • Moderately well understood. • Models inter-item unreliability. • Suggested to be closer to 'true reliability' than alpha.	• May understate reliability when score distribution 'narrow' or 'bunched'. • Makes no statement about misclassification. • Needs all candidates to answer same questions to function.
Rasch-based person separation reliability index (REL)	Pseudo-internal consistency	• Simple to compute (using specialist software). • Models inter-item unreliability. • Based on trait measures, which amount to 'adjusted scores' in the case where different candidates receive different items. • Could be inserted into Livingston's index for tests where different candidates receive different items (and such a use would model inter-item unreliability in the case where different candidates received different items).	• Needs specialist software to compute. • Rasch models tend to be contentious among psychometricians. • VQ tests are designed to assess all LCs and ACs – not to develop a latent trait scale. Therefore, not clear as a matter of principle that these tests *should* fit the Rasch model. • Person separation reliability index not the main way that IRT conceives of reliability.
Relative G coefficient	Generalizability	• Simple to compute (using specialist software). • Models inter-item unreliability.	• By definition equivalent to alpha and returns equivalent results to alpha in two-factor situations, such as candidates taking a fixed form test. This may make alpha more attractive for researchers who do not have specialist software.

Index	Type of index	Strengths	Weaknesses/limitations
Phi	Domain dependability (absolute measurement)	• Simple to compute (using specialist software). • Models inter-item unreliability. • 'Nested' design can be used to overcome limitation of crossed design requiring all candidates to answer all items. • Use of absolute rather than relative error variance makes it more appropriate for CRT applications.	• Needs specialist software to compute. • When nested designs are used, candidates who did not answer all questions have to be discarded. This could be a disadvantage if this involves too many candidates.
Livingston's index	Squared-error loss agreement	• Squared-error loss agreement indices are – as a matter of principle – the most appropriate indices for 'internal reliability' analyses of CRTs with cut scores. • Index is written sufficiently generally to permit the 'plugging in' of various internal consistency indices. • Index can return values in the case where different candidates receive different items. • Could be valuable for some researchers – eg those without specialist g-theory software.	• Index has been criticised as making too many CTT/NRT assumptions for a CRT application (superseded by phi(lambda)). • 'Plugging in' KR-21 and REL is an 'idiosyncratic 'approach'; close analysis of internal algebra may suggest unresolved inconsistencies.
Brennan & Kane's phi (lambda)	Squared-error loss agreement	• Squared-error loss agreement indices are – as a matter of principle – the most appropriate indices for 'internal reliability' analyses of CRTs with cut scores. • Researchers suggest that phi (lambda)'s non-reliance on CTT/NRT assumptions gives it the edge over Livingston's index. • 'Nested' design can be used to overcome limitation of crossed design requiring all candidates to answer all items.	• Needs specialist software to compute. • When nested designs are used, candidates who did not answer all questions have to be discarded. This could be a disadvantage if this involves too many candidates.
SEM	Standard Error of Measurement	• Can be calculated straightforwardly, given the existence of an estimate of measurement error variance. • Is often a default output of analysis programs. • Recommended as preferable to reliability coefficients by Cronbach, when he reflected on his life's work. • Can be used to provide confidence intervals around test scores, including cut scores.	• Does not amount to 'an internal reliability estimate' and therefore may be out of scope for this project. • Needs careful interpretation; scores within CIs are not necessarily misclassifications.

Table 2: Reliability indices used in the project, with strengths and weaknesses

Interpretation of reliability indices

Reliability analysis is a deeply quantitative enterprise. However, quantification alone is insufficient to constitute effective reliability research. Nicols & Smith (1998, p. 34) argue for:

> a reconceptualization of reliability that reflects the importance of the theoretical expectations of the test specialist and the learning and problem solving of the test takers.

Parkes (2007, p. 6) suggests that reliability research should adopt recent innovations from validity research and attempt to construct reliability arguments. Such arguments would have the following six stages:

1. A determination of the social and scientific values of dependability, consistency, accuracy, etc. most relevant to the scenario at hand.
2. Clear statements of the purpose and the context of the assessment.
3. The definition of a replication in the particular assessment.
4. A determination of the tolerance or level of reliability needed.
5. The evidence.
6. The Judgment: Pulling it all together

Baird et al (2011, p. 22) suggest that the following information should be presented in reliability research reports:

Facets – which replications have been included in the study (eg raters, items, occasions)	
Conceptual	• **Comparison** – are observed scores being compared with observed, or is there a claim about comparison with true scores? • **Generalisation** – what is claimed or known about the generalisation of the reliability estimate?
Design	• **Level** – has study been conducted at a component or qualification level? • **Administration** – procedures for the assessment and study and claim about ecological validity of those (eg blind or non-blind presentation to raters) • **Representativeness** – of facet objects (eg raters), stimuli and data • **Method** – analysis (eg G-theory, IRT) and reporting (eg percentage consistently classified) • **Scale** – number of points and grades

Table 3: Checklist for reporting reliability claims

It is suggested that the extensive review section above, and the detailed description of procedures in the following Method section, addresses the concerns implied by Baird et al in their table above. When findings are presented below, they will also be contextualised with a clear argument as to their meaning.

Taxonomies

The organised description of (groups of) phenomena by means of taxonomies is one of the fundamental building blocks in the scientific method. A taxonomy will be beneficial for this reliability project because it will help describe tests that meet criteria set out by Ofqual, and enable a range of tests to be selected for reliability analysis. The information required to describe tests and form the taxonomy could be gathered from the following sources:

- the Register of Regulated Qualifications (RRQ) – database of active qualifications accredited by Ofqual (see: http://register.ofqual.gov.uk/)
- Regulatory IT System (RITS) – database of active and inactive qualifications accredited by Ofqual (see: http://www.ofqual.gov.uk/rits)
- SAP – City & Guilds' database on active and inactive accredited and non-accredited qualifications
- Business Warehouse (BW) – City & Guilds' database on income, registrations and completions (relevant information is fed into SAP)
- Qualification Management Database (QMD) – City & Guilds' database relating to the qualification development process (relevant information is fed into SAP)
- Other supplementary sources such as City & Guilds' website and internal product catalogue

However, some of these are not directly compatible with each other either in terms of format or in terms of the fields they hold. The decisions to select particular sources and the issues relating to these will be discussed in the next section for each indicator that forms part of the taxonomy.

Method

This section starts with a detailed discussion of the taxonomy development. This taxonomy forms the basis for the test selection approach which is explained immediately afterwards. The procedures used to collect, prepare and quality assure data for selected tests are then described. This section ends with an outline of the reliability indices and software packages used in this study.

Taxonomy development

Assessment type

A variety of assessment methods are used in VQs. Whilst many units are assessed using assignments, portfolios and practical assessments, which are based on judgements, it is something of a myth that there are no tests (cf. Baird, et al, 2011, p. 57). Assessments such as multiple choice (MC) and short answer question tests are also used in VQs. As the reliability project focuses on MC and short answer tests, where test-related unreliability is expected to represent the major source of measurement error, identifying such tests amongst the array of assessments used in VQs is a crucial part of the taxonomy.

In the QMD and SAP databases, similar categories are used to classify different types of assessments[17] with only one type assigned to each unit at any one time. On the other hand the Ofqual database RITS uses fewer and slightly different categories[18] but allows multiple assessment types to be attached to each unit; making it difficult to identify eligible units for the purpose of this study. Hence SAP is a better option for identifying assessment type.

There are two main types of MC tests used within City & Guilds (see Figure 1, Column C3 – below, at p. 36):

[17] Assignment, Case Study, Centre Devised, Individual Work, Multiple Choice, Oral/Spoken, Portfolio, Practical, Practical/Oral, Profile, Project, Proof Reading Test, Proxy, Short Answer, UPK Test and Written.

[18] Aural Examination, Coursework, e-assessment, Multiple Choice Examination, Oral Examination, Portfolio of Evidence, Practical Demonstration / Assignment, Practical Examination, Task based Controlled Assignment and Written Examination.

i. **End-of-unit MC tests** assess candidates' knowledge and understanding, do not require assessor judgement and the performance on the test determines the overall result for the unit being assessed. These are classified as 'Multiple Choice' in the SAP database. Item data are available for the majority of these tests (exceptions noted subsequently).

ii. **Underpinning Knowledge (UPK) tests** assess all knowledge areas within a unit. The generated score report forms part of the portfolio of evidence required to achieve the unit. Knowledge elements not demonstrated through the MC test can then be assessed using other appropriate methods eg oral questioning or short answer questions. Due to their nature, UPK tests are classified as an assessor-marked assessment eg 'portfolio' in RITS but as 'Underpinning Knowledge Tests' in SAP.

Similarly, short answer questions are also used to assess knowledge and understanding. These tests are categorised as written exams, assignments or another assessment type in the databases mentioned above, because no such category exists in RITS and only one short answer question test is found in SAP. This makes it difficult to distinguish short answer tests from those based on more detailed essay type questions that require assessor judgement. Furthermore, there is no candidate item response data available for short answer question tests due to the way these are administered. For these reasons, short answer question tests will not be assessed as part of this study.

Test mode

Tests in VQs may be delivered through different media – for example, on screen or on paper (e-assessment is much more prevalent in this sector than in academic qualifications). This is an important part of the taxonomy because the flow and availability of item data tends to vary between online and paper MC tests (see Figure 1 C2).

- Paper MC tests are mainly end-of-unit exams – item data are not available for centre-assessed paper MC tests (more details to follow)

- Online MC tests can either be end-of-unit or UPK tests – item data are available for all end-of-unit online MC tests but not for UPK tests

It might be assumed that units classified as both ‘e-assessment’ and ‘Multiple Choice Examination’ in RITS are online MC tests but in fact these can also include UPK tests. Furthermore, there is uncertainty relating to whether or not all those units solely classified in RITS as ‘e-assessment’ are MC tests and all those solely classified as ‘Multiple Choice Examination’ are in fact paper based. In comparison, there is an indicator in SAP that distinguishes online from paper-based units of assessment and so this is the preferred indicator.

Fixed-form or randomly-generated tests

Fixed-form tests present the same questions to all candidates whereas randomly-generated tests are based on a random selection of items from an item bank. City & Guilds’ randomly-generated tests are generally based on a tight specification with a predefined number of items delivered to candidates from within item pools relating to the relevant ACs. The distinction between a fixed-form and a randomly-generated test is a significant one because this in part determines which reliability indices are appropriate, and can be produced, for which tests.

There are no clear indicators for fixed form and randomly-generated tests in any of the databases explored. However, discussions with colleagues revealed that:

- end-of-unit paper MC tests are almost always fixed, although different versions of the test may be available at any one time
- end-of-unit online MC and UPK tests are predominantly randomly-generated tests with a few exceptions (mainly Key Skills tests) – (see Figure 1 C4)

On-demand or dated tests

Paper-based, fixed-form MC tests can either be held on set dates (dated entry) or made available on-demand (see Figure 1 C5). According to the SAP database, many dated and on-demand paper MC tests are scanned and machine marked but some are also marked internally by Assessment Centres accredited by City & Guilds. Although data for the machine-marked paper MC tests can be obtained, data for centre-marked tests are not readily available.

In comparison, online end-of-unit MC and UPK tests are primarily available on-demand with a few exceptions (mainly Key Skills tests). Item-level data are available for all end-of-unit online on-demand MC tests but there are no such data for UPK tests.

The figure below illustrates the relationship between the indicators discussed so far.

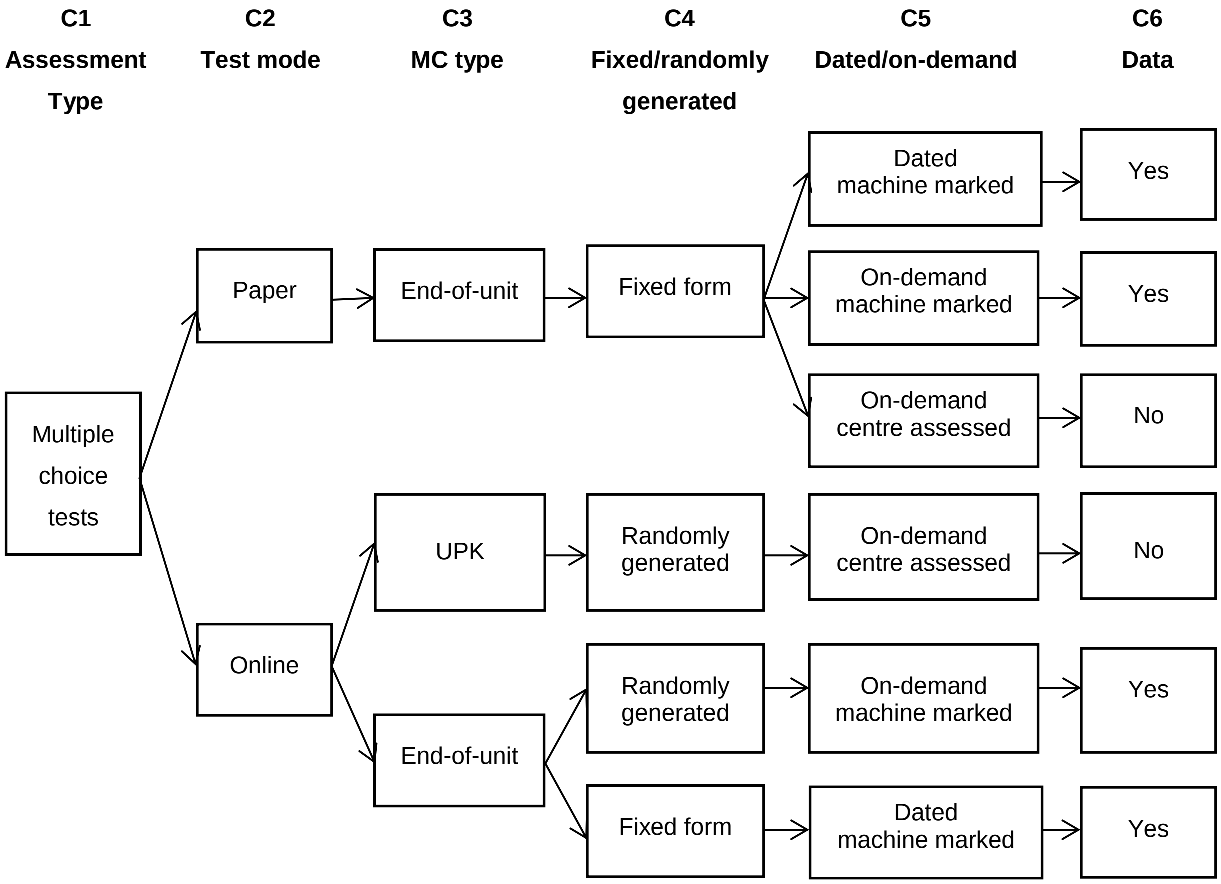


Figure 1: Relationship between selected variables in the taxonomy

In addition to the above set of indicators, there are additional factors that need to be considered as part of the taxonomy, such as industry area, level, uptake, risk etc.

Industry area

Several sources show the main industry or sector area for each unit of assessment but there is variation between these. For example:

- a list of 15 broad and 48 detailed sectors form the Sector Subject Area Classification System (SSAC)[19] used in RITS and RRQ
- over 30 sectors and 100 sub-sectors are available within QMD to indicate main industry area relating to each unit
- around 30 disciplines and over 100 sub-disciplines are used in SAP, which are similar (not identical) to those in QMD, but are available at the qualification rather than unit level

Although the number of categories used in RITS and RRQ appear to be more manageable than the other sources, they cannot be used. This is because units of assessment were identified and selected for this project primarily based on information sourced from SAP. A further practical consideration is the fact that linking information between City & Guilds internal systems and external databases is challenging due to the way that units are set up. For instance, RITS and RRQ allow multiple assessment methods to be assigned to one unit but internal City & Guilds systems require each type of assessment to be set up as a separate unit. Therefore one unit in RITS could relate to multiple units in City & Guilds' systems. For these reasons, it is recommended that internal sources should be used to obtain industry information and that QMD was preferable because it provides this information at the unit rather than the qualification level.

Level

For similar reasons to those mentioned above, information about unit (QCF) level was obtained from the internal City & Guilds source such as SAP.

[19] Available at: http://www.ofqual.gov.uk/standards/142-statistics-articles/429-sector-subject-area-classification-system-ssac

Uptake – completions and bookings

Between October 2010 and September 2011, the number of candidates sitting City & Guilds MC tests ranged from zero to a little over 40,000 per test. City & Guilds SAP database draws on test completion data for each unit of assessment from BW so could be used to identify units with low, medium and high take-up. Completions are not recorded for UPK tests as they are centre-assessed and form part of another main assessment method within a unit. However, bookings (ie number of candidates booked to take the test) are recorded for these tests, which could give an indication of uptake if required.

Licence to practise

Licence to Practise (LtP) is any requirement, including standards, voluntary or statutory, to which employers and employees in a sector must adhere. The LtP provides opportunities for good companies to demonstrate quality and provides reassurance for consumers. This term is used to describe a range of controls, which can include completion of a specific qualification, passing independent industry assessments, passing medical tests, undergoing criminal background checks and engaging in a specified minimum amount of continuing professional development activities. Such controls are imposed because of the threats potentially arising from poor performance in certain industries and where, consequently, mismanagement could have dire consequences (City & Guilds, 2011b). Hence, it is proposed that units relating to City & Guilds qualifications that lead to LtP status were included in this study. An indicator for LtP is available within RRQ at the qualification level ie ‘purpose’, but as this field does not contain any information for many qualifications so was not seen as a suitable source. Therefore information relating to LtP was obtained from the City & Guilds LtP team and the SAP database.

Apprenticeships

There has been growing attention and emphasis placed on apprenticeships over recent years within City & Guilds, the Government and beyond. The rise in apprenticeships is evident with some ‘442,700 apprenticeship starts in the 2010-2011 academic year compared with 279,000 the previous year’ (BBC, 2011). Given recent activity and interest relating to apprenticeships, the inclusion of such a field in the taxonomy would also be worthwhile. However, finding a reasonable indicator in the databases is by no means an easy task because of the way apprenticeships are set up (they typically comprise of an NVQ, key skill units and a technical certificate or equivalent). Therefore a variety of sources (eg SAP, City & Guilds web site and internal product catalogue) had to be used to find apprenticeship schemes in the first instance. This was followed by identification of the qualifications that form part of each scheme and the units that relate to these qualifications.

Risk

Although it would have been interesting to compare reliability indices of tests based on the level of risk associated with the unit, a robust risk indicator was not available in any database. The risk indicator in QMD is available for overall suites (groups) of qualifications rather than for individual units and the decisions around classification as high risk or otherwise are not well documented. A new risk indicator was recently added to QMD to identify risky qualifications as well as the reasons for the risk but information relating to this field is not yet populated and so could not be used for this project.

Performance information

Test performance information such as mean score, score range and variance was also considered to be an appropriate inclusion to the taxonomy. However, this information was not readily available and would have been time consuming and cumbersome to collate for all MC tests. Such statistics were therefore derived from actual data for tests selected for the reliability analysis.

In summary, data to populate the indicators for the taxonomy are available but are located across a number of databases and there is a lack of consistency between them. The table below shows the availability of information across different sources and the variations that exist.

Indicators	RRQ†	RITS†	SAP	QMD	BW
Assessment type*		✓	✓	✓	
Test mode*		✓	✓	✓	
Fixed form/randomly generated					
On-demand/dated			✓		
Industry area*	✓	✓	✓	✓	✓
Level*	✓	✓	✓	✓	✓
Completions/bookings			✓		✓
Licence to practise*	✓	✓	✓	✓	
Apprenticeship*		✓	✓	✓	
Risk				✓	

*These fields vary slightly between the different databases.
†The fields in RRQ come from RITS, but certain fields are not available in the publicly available extract pertaining to all units.

Table 4: Availability of indicators across databases

Test selection for reliability analysis

City & Guilds had over 2,000 end-of-unit MC tests listed in internal databases in November 2011 (SAP and BW). However, many of these were centre assessed, inactive (i.e. no longer used) or active but with little or no recent take-up. Therefore these did not meet the data and analysis requirements of this project. Once these were removed from the list this left 493 tests that had issued results for at least 100 candidates in any one year over the past three years.

However, there was a further requirement, that the tests studied should be focussed on levels 1, 2 and 3. Tests at these levels constitute the majority of MC tests in City & Guilds; some 432 of the 493 tests were at these levels. Therefore the final selection of tests for analysis was made from this long list of 432 tests.

In line with the main taxonomy indicators, these tests were split into four main groups:

1. Paper-based, fixed form dated tests
2. Paper-based, fixed form on-demand tests
3. Online, fixed dated tests

4. Online, randomly generated on-demand tests

The table below shows the number of tests available for analysis within each group with a breakdown by level, risk, licence to practise and apprenticeship scheme.

Test mode	Fixed/ randomly generated	Dated/ on-demand	Total available	Level 1	Level 2	Level 3	High risk	Medium risk	Low risk	Licence to Practise	Apprenticeship
End-of-unit Paper	Fixed	Dated	49	19	27	3	0	0	34	0	12
		On-demand	44	13	29	2	5	7	31	5	18
	Total paper		***93***	**32**	**56**	**5**	**5**	**7**	**65**	**5**	**30**
End-of-unit Online	Fixed	Dated	14	7	7	0	0	0	14	0	14
	Random	On-demand	325	46	184	95	23	29	252	18	121
	Total online		***339***	**53**	**191**	**95**	**23**	**29**	**266**	**18**	**135**
GRAND TOTAL			***432***	***85***	***247***	***100***	***28***	***36***	***331***	***23***	***165***

Notes: risk could not be determined for some units based on available information.

Table 5: Tests available for analysis by taxonomy indicators

In total, 68 tests were selected and analysed. The approach to the test selection considered the characteristics of the individual tests within each group in order to ensure that a range of tests were chosen that would represent the range of mode and test types, difficulty and risk levels, and tests related to licence to practise and apprenticeship schemes. The resulting sample of tests drawn using these indicators is shown in the table below; a complete list of tests is shown in Appendix 2 (below, at p. 82).

Test mode	Fixed/ Randomly generated	Dated/ on-demand	Total analysed	Level 1	Level 2	Level 3	High risk	Medium risk	Low risk	Licence to Practise	Apprenticeship
End-of-unit Paper	Fixed	Dated	27	16	10	1	0	0	20	0	9
		On-demand	15	4	10	1	4	1	10	4	8
	Total paper		***42***	***20***	***20***	***2***	***4***	***1***	***30***	***4***	***17***
End-of-unit Online	Fixed	Dated	5	2	3	0	0	0	5	0	5
	Random	On-demand	21	7	6	8	2	1	16	2	9
	Total online		***26***	***9***	***9***	***8***	***2***	***1***	***21***	***2***	***14***
GRAND TOTAL			***68***	***29***	***29***	***10***	***6***	***2***	***51***	***6***	***31***

Notes: risk could not be determined for some units based on available information.

Table 6: Tests analysed by taxonomy indicators

Furthermore, the chosen tests covered a variety of industry areas, including: beauty and complementary therapies; building services; business skills; construction and energy; engineering; hairdressing; hospitality and catering; IT; land based services; learning; manufacturing; retail and warehousing; security; skills for work and life; sport and recreation; transport maintenance; and travel, tourism and aviation. Only five industry areas were not covered, and this was because they had relatively low numbers of tests available for analysis.

The industry levels of the tests are shown in Table 7, below.

Industry	***Level***			***Grand Total***
	1	***2***	***3***	
Beauty Therapy	1		1	2
Building Services		4	1	5
Business Skills	3	3		6
Construction and Energy	6	2	1	9
Engineering	2			2
Hairdressing	2			2
Hospitality and Catering	2	3		5
IT	1	1	1	3
Land Based Services	1	4		5
Learning		1		1
Manufacturing Industry			1	1
Retail and Warehousing	1			1
Security		4	2	6
Skills for Work and Life	6	4		10
Sport and Recreation		1		1
Transport Maintenance	1		2	3
Travel, Tourism, Aviation	3	2	1	6
Grand Total	***29***	***29***	***10***	***68***

Table 7: Industry areas of analysed tests, pivoted against level

This table illustrates one limitation to the types of reliability analysis that could be conducted based on the set of selected tests. It might be useful to compare the reliability of different industry sectors' tests, for example. However, Table 7 suggests that this would not be appropriate. This is for the following reasons:

- There are only a few tests from each industry sector, and thus any reliability analyses would only be based on a small sample of tests, making robust comparisons difficult to sustain.
- The features of tests are confounded, so that, for instance, if we found that industry area A appeared more reliable than industry area B, we would not know whether that apparent finding was actually an artefact of some other feature – such as: level, mode of testing (randomly-generated or fixed test), test length, and so on.

The take-up of tests was broadly taken into account when selecting tests. Based on results issued over the past three years, each of the 432 MC tests was classified into one of three groups:

- ***small*** (below 50th percentile) – 216 tests (from which 18 tests were drawn for analysis)
- ***medium*** (between 50th and 75th percentile) – 108 tests (from which 26 tests were drawn for analysis)
- ***large*** (above 75th percentile) – 108 tests (from which 24 tests were drawn for analysis)

There were relatively fewer of the smaller tests with sufficient numbers of candidates who had taken a particular version of the test, and this accounts for the smaller number of these tests in the selection taken through to analysis[20].

Data collection

Item level data for **paper MC tests** were obtained online from the City & Guilds SAP database for paper-based machine-marked MC tests. The item data report included: unique unit number; unique student number; test version (for on-demand tests); scheduled test date (for dated tests); actual test date; question number; candidate response; answer key; right/wrong indicator (yes/no); total questions; grade; marks (total raw score); candidate attendance status (present, absent etc.); unique centre (customer) number; and centre name. The data were exported into Excel in readiness for analysis. Cut scores for tests were also acquired through the SAP database.

For **online MC tests**, item data were obtained through the existing Pearson VUE CATGlobal.com online portal. This portal allows building and publishing of tests, content management and delivery, and reporting of results and other relevant information. The item data report includes a basic search tool, which provides access to item level data. The following information was obtained for each test: test name; form number; item number; answer key; candidate response; candidate number; and overall raw score. The data was extracted into the default text file format. Cut scores for online tests were found through another part of the CATGlobal.com portal.

Data preparation, issues and quality assurance

A number of steps were taken to prepare the data for analysis but these varied slightly by type of test because the raw data format, method of analysis and analytical software

[20] This is unsurprising, given the cut-off of at least 100 candidates in any one year. Also, larger data files are generally preferable for reliability analysis.

varied by type of test. Details of the steps taken, along with quality assurance procedures and any issues faced are reported for the different types of tests in Appendix 1 (below at pp. 80ff).

Calculation of reliability indices

Of the several reliability indices discussed in the review section, Cronbach's alpha, Guttman's lambda-2, coeff_G, phi and phi (lambda) were considered to be the most appropriate reliability indices for fixed-form tests and so were applied to all 47 such tests. In comparison, phi and phi (lambda) were the most relevant measures for randomly-generated tests and so were used to analyse all 21 randomly-generated tests in the sample. In addition, the Rasch-based person separation index, KR-21, Livingston's index with Rasch and Livingston's index with KR-21 were tested on three randomly-generated tests.

The table below gives an overview of the number and types of tests to which each reliability index was applied and the main software package used to carry out the required analysis.

Reliability index	Tests analysed by type		Main software package used for analysis			
	Fixed	Randomly generated	EduG[1]	SPSS[2]	Winsteps[3]	Excel[4]
Cronbach's alpha	47 tests	-		✓		
Guttman's lambda-2	47 tests	-		✓		
Coeff_G[21]	47 tests	-	✓			
Phi*	47 tests	21 tests	✓			
Phi (lambda)*	47 tests	21 tests	✓			
Rasch-based person separation	-	3 tests			✓	
Kuder-Richardson 21 (KR-21)	-	3 tests				✓
Livingston's index with Rasch	-	3 tests				✓
Livingston's index with KR-21	-	3 tests				✓
Standard Error of Measurement†	3 tests	-				✓

* Different estimation designs used for fixed form and randomly-generated tests.

† SEM estimates were output from EduG and confidence intervals were calculated using Excel.

1. **EduG** is a specialist statistical software application used to perform generalizability analysis based on the Analysis of Variance (ANOVA) and Generalizability Theory (G-Theory) (SSREWG, 2010).
2. **IBM SPSS** is a standard off-the-shelf statistical package that provides numerous analysis possibilities, although generalizability coefficient calculation is not a standard menu-based option.
3. **Winsteps** is a specialist program primarily used to perform analysis based on the Rasch model and item response theory (IRT) (Linacre, 2009).
4. **Microsoft Excel** is a spreadsheet based computer program, which allows data manipulation, calculations, graph creation and complex formulas to be embedded.

Table 8: Number of tests by reliability indices and software packages used

[21] Coeff_G is definitionally and empirically identical to alpha in the case where all candidates answer the same test items (fixed-form tests).

Due to the nature of fixed-form and randomly-generated tests, different estimation designs were adopted when running phi analysis for these tests. Where all candidates were presented with the same items (fixed-form tests), the crossed design was used for analysis. On the other hand, a nested design was used for analysis of randomly-generated tests i.e. an 'items nested within candidates (i : c)' design, which acknowledged that different candidates were asked to answer different sets of questions. However, this design requires candidates to answer the same number of questions, which is why candidates who had taken a randomly-generated test but did not have answers recorded for the full required number of questions were excluded from the analysis. EduG permits both crossed and nested designs to be applied (the latter if all candidates answer the same number of questions).

For quality assurance purposes, some of the reliability indices were applied in more than one software package. For instance:

- Cronbach's alpha was run for 26 of 47 fixed form tests in both SPSS and Excel
- phi and phi (lambda) were run for 30 of 68 tests in both EduG and Excel

In general, we prefer to use specialist analysis software to produce indices; this being a more effective and generally safer way of proceeding. However, in this project, Excel proved a useful way to understand the 'under-the-bonnet' workings of indices, and in addition provided the opportunity for a further quality check (allowing us to check that the same answer was obtained using the two different types of software).

In addition to reliability indices, basic test statistics were derived for all tests – including score mean, variance, range, kurtosis etc. The distribution of scores for VQ MC tests was of particular interest and was explored using kurtosis (Pearson, 1905) as part of this study.

We took the following definition of kurtosis from the SPSS statistical software:

> A measure of the extent to which observations cluster around a central point. *For a normal distribution, the value of the kurtosis statistic is zero*. Positive kurtosis indicates that, relative to a normal distribution, the observations are more clustered about the centre of the distribution and have thinner tails until the extreme values of the distribution, at which point the tails of the leptokurtic distribution are thicker relative to a normal distribution. Negative kurtosis indicates that, relative to a normal distribution, the observations cluster less and have thicker tails until the extreme values of the distribution, at which point the tails of the platykurtic distribution are thinner relative to a normal distribution.

We used that program to estimate the kurtosis values of distributions, and the Standard Error of Kurtosis (SEK) estimate. We categorised distributions as mesokurtic if we had 95 per cent confidence that the kurtosis value was such that it was likely to be from the normal distribution. Where the kurtosis value was below that range (given its value and SE) we categorised the distribution as platykurtic (flatly peaked) and where it was higher we categorised it as leptokurtic (slender, or steeply-peaked).

Furthermore, the relationship between test score distribution and reliability was investigated through comparison of reliability indices for tests with different levels of kurtosis. The results of these analyses are presented in the next section.

The values of phi and phi (lambda) on a small selection of tests were also compared. This was done by showing differing values of those indices for a small number of tests, and by discussing how the indices differed – as a function of the value of phi and phi (lambda), and the distance between the mean score and cut score. This contrast was used to make conclusions about the usefulness of phi and phi (lambda).

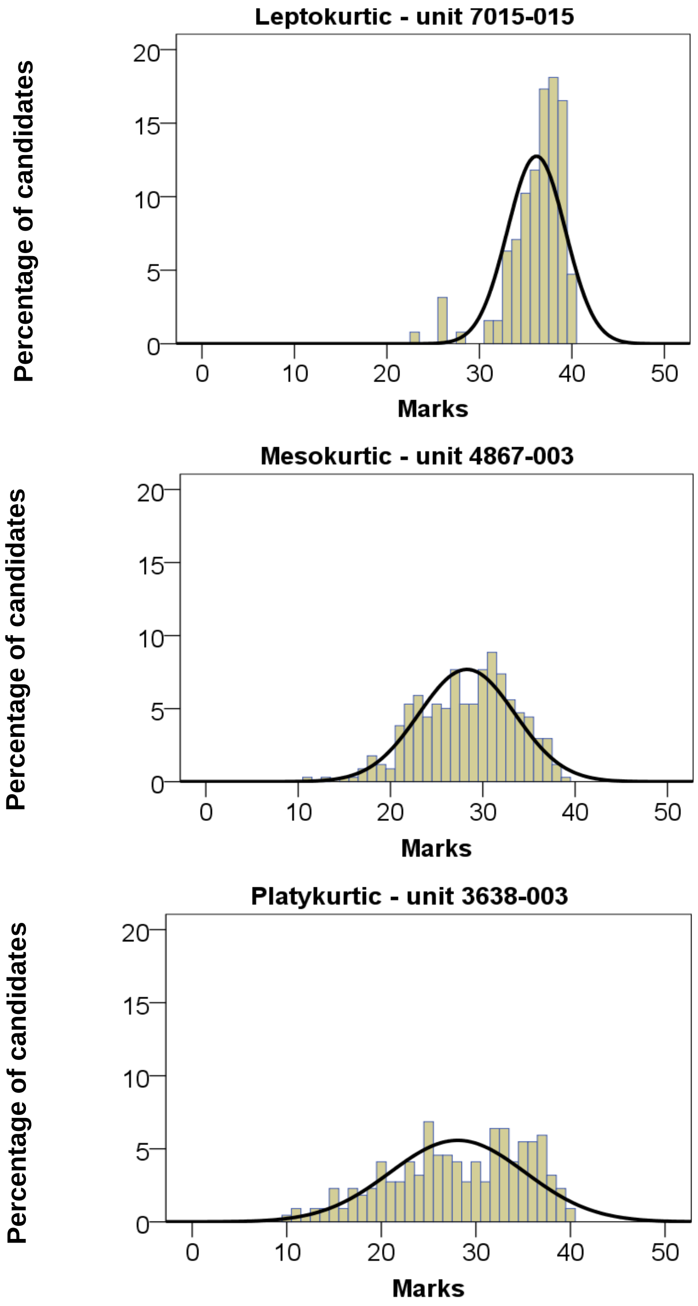
Findings

Score distribution

Kurtosis was used to explore the distribution of the MC tests analysed.

- Leptokurtic is a score distribution that is highly peaked around the mean
- Mesokurtic is a moderately peaked score distribution (typical of the normal distribution)
- Platykurtic is a flatter score distribution

Here is an example of each from the tests analysed:


Leptokurtic - unit 7015-015

Mesokurtic - unit 4867-003

Platykurtic - unit 3638-003

Figure 2: Histograms of a leptokurtic, mesokurtic and platykurtic score distribution

As shown in the table below, slightly more of the paper-based, fixed dated tests in the sample had a mesokurtic score distribution than a leptokurtic distribution whereas all the other test types tended to have a leptokurtic distribution. Few tests were classified as ‘platykurtic’. The kurtosis value, SEK, and the type is given for each test in Appendix 3 (below, at p. 85).

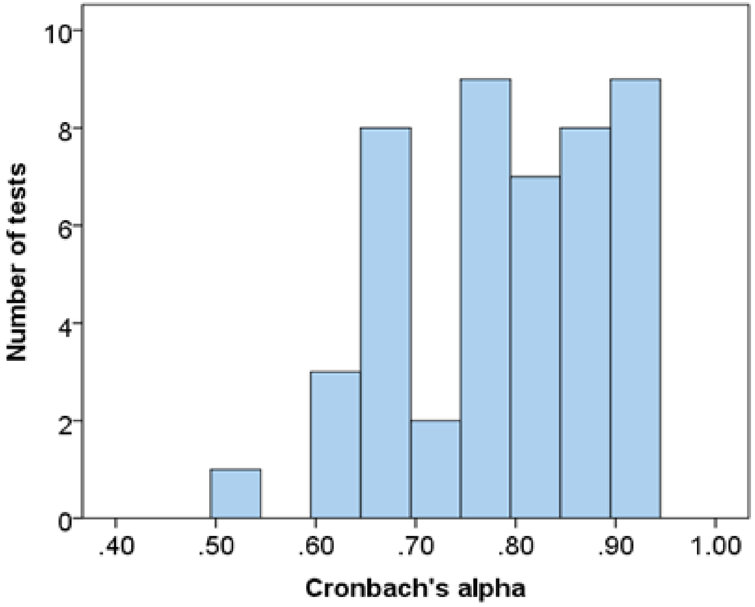
Test type	Leptokurtic	Mesokurtic	Platykurtic	Total
Paper fixed dated	10	14	3	27
Paper fixed on-demand	12	3	0	15
Online fixed dated	4	1	0	5
Online randomly-generated on-demand	11	7	3	21
Total	37	25	6	68

Table 9: Number of tests analysed by type of distribution

Reliability

Cronbach’s alpha

Cronbach’s alpha was applied to all 47 fixed-form tests, and produced a mean value of 0.79. This is similar to Bramley & Dhawan’s (2010) estimates of the reliability of GCSE and A-level qualifications, where the mean Cronbach’s alpha was 0.82 for AS units and 0.81 for GCSE units. Of the 47 tests, 23 had an alpha value greater than 0.8. The minimum and maximum values were 0.52 and 0.94 respectively.



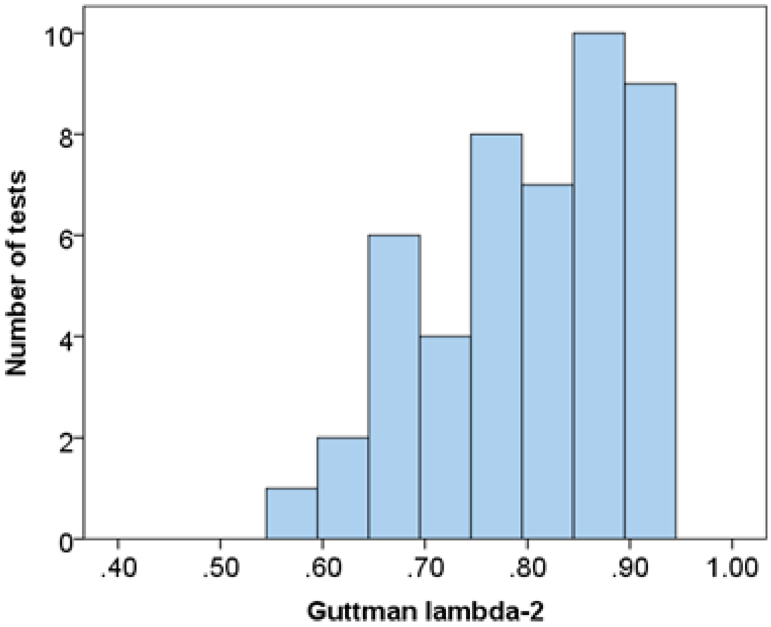
Number of tests	47
Mean	0.79
Median	0.80
Std. Deviation	0.11
Minimum	0.52
Maximum	0.94
Lower quartile	0.69
Upper quartile	0.87

Figure 3: Distribution of Cronbach’s alpha for fixed form tests

The relative generalizability coefficient (coeff_G) is identical in definition and effect to Cronbach’s alpha in the case of fixed-form tests, and therefore only alpha findings are reported.

Guttman’s lambda-2

The distribution and summary statistics for Guttman’s lambda-2 are very similar to those found for Cronbach’s alpha, with a mean value of 0.80.

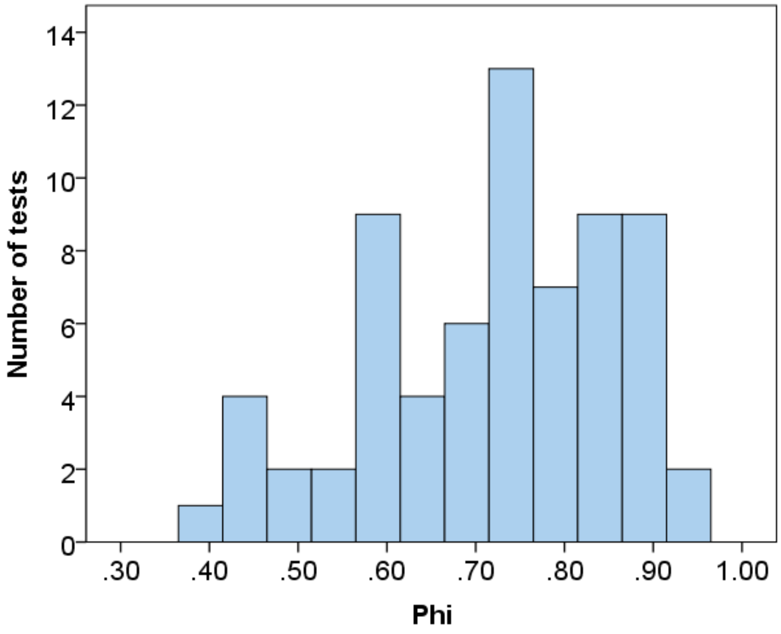


Number of tests	47
Mean	0.80
Median	0.81
Std. Deviation	0.10
Minimum	0.57
Maximum	0.94
Lower quartile	0.71
Upper quartile	0.88

Figure 4: Distribution of Guttman’s lambda-2 for fixed form tests

Phi

This is the reliability index which, as would be predicted from its meaning[22] shows the lowest value, with a mean of 0.72. Although the maximum value of 0.94 mirrors Cronbach's alpha and Guttman's lambda-2, the lowest phi value is much lower at 0.39.



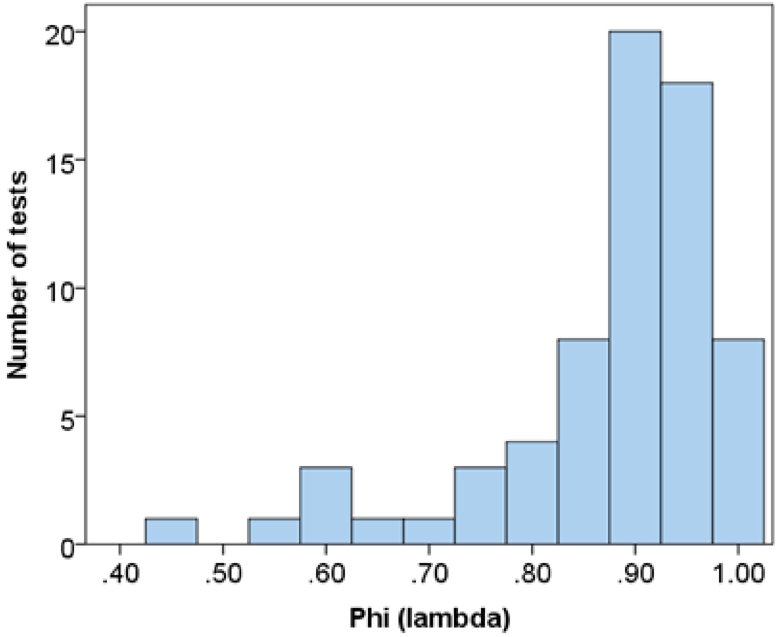
Number of tests	68
Mean	0.72
Median	0.73
Std. Deviation	0.14
Minimum	0.39
Maximum	0.94
Lower quartile	0.60
Upper quartile	0.83

Figure 5: Distribution of phi for all tests

[22] Given that it has an extra term in its denominator, ie the between-item variance.

Phi (lambda)

In comparison to previous indices, phi (lambda) produced the highest mean, at 0.88, and the distribution below shows relatively more tests with a value over 0.8. It should be remembered that the value of phi (lambda) will change if the cut score, lambda, is changed.



Number of tests	68
Mean	0.88
Median	0.90
Std. Deviation	0.12
Minimum	0.45
Maximum	0.98
Lower quartile	0.85
Upper quartile	0.96

Figure 6: Distribution of phi (lambda) for all tests

Comparison of reliability indices

The figure below compares mean reliability indices for each type of test.

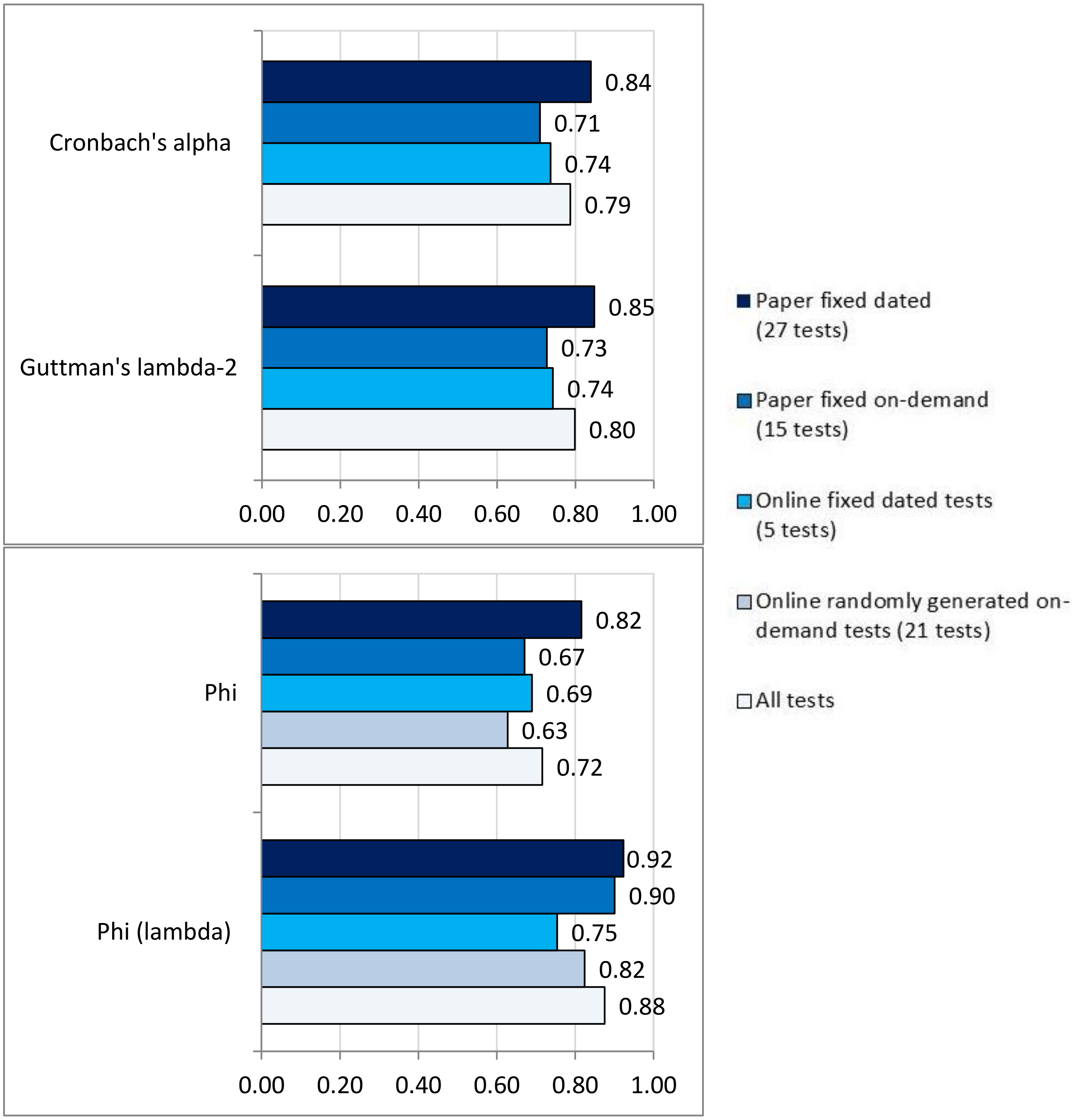


Figure 7: Mean reliability indices by test type

Figure 7 illustrates that paper-based, fixed-form tests have higher mean reliability than other types of test across all indices – ranging from 0.82 to 0.92. In comparison, the mean reliability of paper-based, fixed-form on-demand tests is around 0.70 across all indices with the exception of phi (lambda), which is 0.90. Online randomly-generated tests on the other hand have the lowest mean of 0.63 for phi in comparison to all indices and test type but the phi (lambda) value is relatively higher at 0.82 for these tests. The mean for online fixed form dated tests needs to be treated with caution due to small sample size, and the fact that other features may affect the reliability values for the tests, other than just test type.

The mean reliability index was derived for leptokurtic and mesokurtic tests. The table below shows that there was no major difference in means between these types of tests for any of the reliability indices.

Reliability index	Leptokurtic	Mesokurtic
Cronbach's alpha	0.78	0.79
Guttman's lambda-2	0.79	0.80
Phi	0.71	0.73
Phi (lambda)	0.90	0.86

Table 10: Mean reliability indices for tests with leptokurtic and mesokurtic distributions

Differing relationships between phi and phi (lambda)

To illustrate the differing relationships between the domain dependability index, phi, and the squared error loss index, phi (lambda)[23], the 21 randomly-generated on-demand tests (see Appendix 3, below at p. 89) were further analysed.

The phi co-efficients for these tests were given a rank order number from 1 (lowest) to 21 (highest). Also, a standardised index was derived (not reported here) to show the squared distance between the mean and the cut score, as a function of total test score. As in the case of the phi coefficients, the tests' values on this coefficient were given a rank order number.

The three tests in Table 11, below, were selected to illustrate differing relationships between phi and phi (lambda).

[23] In principle similar relationships could be demonstrated between Livingston's squared error loss index and the internal consistency index to which it was related.

Unit number	Items to answer	Mean score	Score variance	Kurtosis	Kurtosis type	Cut score	Phi	Phi (lambda)	Phi rank order	Distance between mean and cut score rank order
3667-102	25	20.7	6.3	1.1	Lepto	16	0.45	0.88	4	16
6926-022	40	32.1	21.1	1.6	Lepto	15	0.72	0.98	15	21
7564-303	80	46.3	143.3	-0.6	Meso	40	0.87	0.9	21	4

Table 11: Selected tests to show relationship between phi and phi (lambda)

The relationships between phi and phi (lambda) for the three illustrative tests can be described as follows. Test 3667-102 has the fourth lowest value for phi amongst the 21 randomly generated tests. This is probably because the score variance on the test is low. (The distribution of scores on 3667-102 is also leptokurtic.) However, the mean score of candidates on this test is quite distant from the cut score; being the sixteenth most distant amongst the 21 selected tests. Thus, the phi (lambda) value is quite high. This appears to be intuitive, since, if average scoring is distant from the cut score, misclassification is likely to be low, and thus a reliability coefficient should return a high value. So, in the case of 3667-102, phi (lambda) appears to provide useful extra information in addition to phi.

Test 6926-022 has one of the higher phi values among the randomly-generated tests (0.72, or fifteenth highest out of 21). Additionally, the standardised distance between the cut score and the mean score is the highest of all the 21 tests. The value of phi (lambda) (0.98) reflects this distance, being close to unity. As in the case of 3667-102, the squared error loss coefficient provides useful extra information to the domain dependability coefficient.

Test 7564-303 has the highest phi coefficient amongst the 21 in the sample. However, the mean score and cut score are close together (fourth lowest squared distance). In this instance, phi (lambda) does not give much additional information to phi.

KR-21, Rasch and Livingston’s index

KR-21, the Rasch person separation reliability index and Livingston’s index were calculated for three online randomly-generated tests to explore the likely value of using these as alternative approaches for such tests. The results, also with those for phi and phi (lambda), are shown in Table 12 below. Inspection of the data reveals that the values obtained for phi and for KR-21 were identical, as were those for phi (lambda) and Livingston’s index with KR-21 (cf. Haertel, 2006, p. 99).

Unit number	Phi	Phi (lambda)	KR-21	Rasch Person Rel	Infit[24]	Outfit	Livingston's index (with KR-21)	Livingston's index (with Rasch)
2079-101	0.69	0.82	0.69	0.71	0.99	0.97	0.82	0.84
4872-201	0.80	0.97	0.80	0.78	0.99	0.95	0.97	0.97
6926-022	0.72	0.98	0.72	0.71	0.99	0.90	0.98	0.98

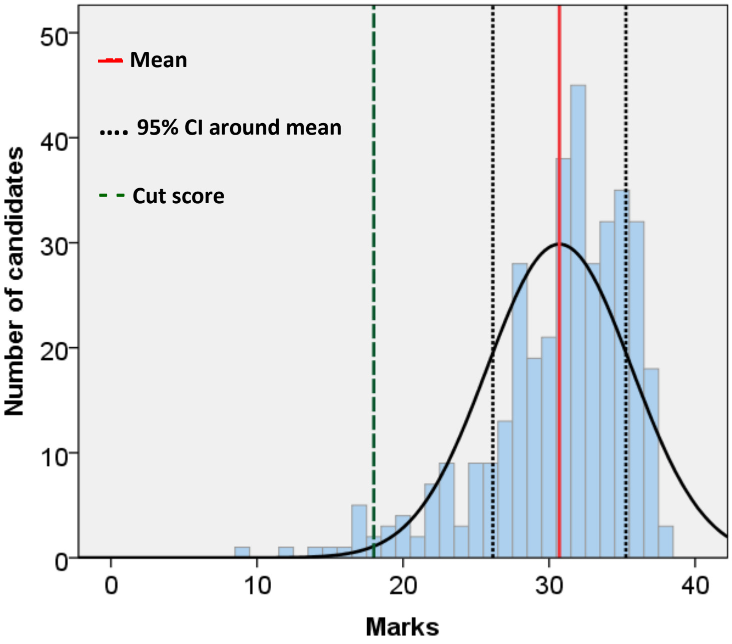
Table 12: Illustration of KR-21, Rasch and Livingston’s index

Standard Error of Measurement

The SEM around total test scores was calculated for three different tests. The score distribution, cut score, mean score and summary statistics are given below. The three examples were selected for analysis based on the following features:

- One with a prominent gap between the cut score and mean score
- One where the mean is lower than the cut score
- One where the cut score and mean are closer together

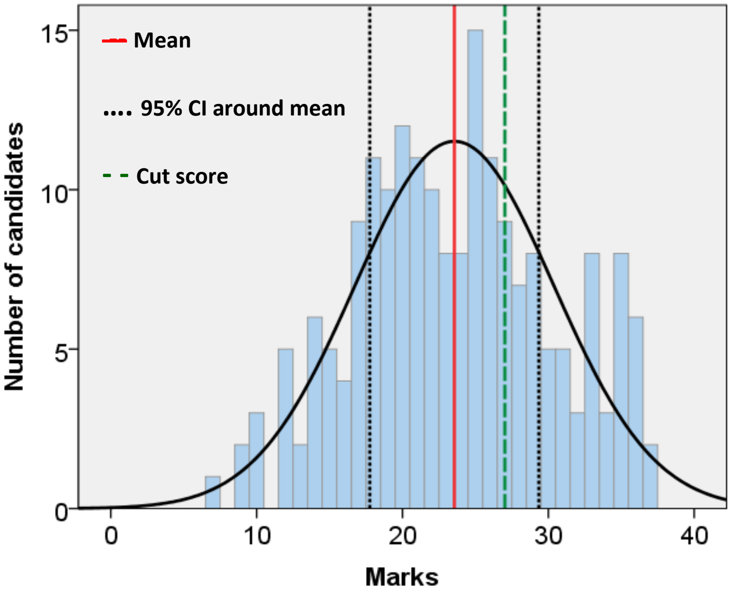
[24] All these fit statistics were of the category ‘productive for measurement’ – see Table 1, above at p. 23.



Absolute SE	0.061
Number of items	38
Mean score	30.71
Cut score	18
SEM	2.32
95% CI	± 4.55

Figure 8: Score distribution with cut score and mean score for unit 1122-001

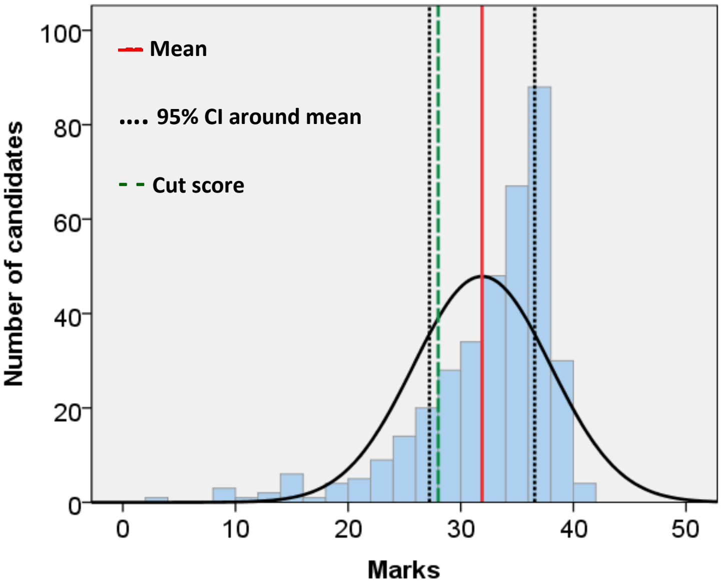
The SEM for the test in which the cut score and mean score are very separated is 2.32. So the 95% confidence interval around a total test score will be roughly ± 4.55 marks. Thus, we can state with at least 95 per cent confidence that a person who scores on the mean has a true score that is above the cut score.



Absolute SE	0.074
Number of items	40
Mean score	23.56
Cut score	27
SEM	2.96
95% CI	± 5.79

Figure 9: Score distribution with cut score and mean score for unit 3692-303

The SEM for the above test is 2.96 and the 95% confidence interval is around ± 5.79 marks. The cut score is well within the 95 per cent confidence interval around the mean. Thus, we cannot state with 95 per cent confidence that a person scoring the mean score has a true score that is below the cut score.



Absolute SE	0.060
Number of items	40
Mean score	31.89
Cut score	28
SEM	2.38
95% CI	± 4.67

Figure 10: Score distribution with cut score and mean score for unit 1892-010

The SEM for the above test is 2.38 and the 95% confidence interval is around ± 4.67 marks. The cut score is within the 95 per cent CI – but only just. Although we could be very confident that a person scoring at the mean was truly above the cut score, we could not quite state this with 95 per cent confidence.

Discussion

In this section, we return to the benefits that the project team stated would accrue when they proposed this project (see p. 4, above). We repeat them here for ease of reference:

- An opportunity to rectify the previous under-representation of VQ assessments in reliability research.
- An opportunity to sponsor research based on large sets of suitable tests.
- An opportunity to derive meaningful and coherent descriptions of tests – rectifying inconsistencies in current awarding body and regulatory databases.
- The application of a range of reliability indices to test score data, including the use of diverse indices beyond the commonly used Cronbach's alpha.
- An understanding of fundamental issues in vocational and – indeed – all forms of assessment; relationship between CRT and NRT, role of knowledge testing in VQs, empirical basis for statements about competence-based assessment, etc. (City & Guilds, 2011a, p. 11)

We now consider each of these points in turn and use them to discuss what has been learned in undertaking this work, and how the work contributes to advancing the body of knowledge relating to reliability estimation for VQs in the UK.

- **An opportunity to rectify the previous under-representation of VQ assessments in reliability research.**

Harth & Hemker (2011) researched the reliability of portfolio judgments in some VQs. The current project started from the observation that there were more tests in vocational portfolios, perhaps more than some might imagine, and set out to investigate the reliability of such tests. The current findings are therefore different to Harth & Hemker's and serve to extend our understanding of the measurement properties of VQs. We believe that this is a positive outcome of this work – both for vocational awarding organisations and for regulators.

There is still much work that remains to be done, of course. We hope that this research will be published on the web and that we can disseminate the findings more widely via conferences, journal articles and so on. In so doing, we aim to integrate our work with that of other high quality researchers working in the field of VQs, both in England (eg Johnson, 2008; Novakovic & Greatorex, 2011), and overseas (see Baird et al, 2011, pp. 53 – 59 for summary of VQ assessment research, including international work).

At present, City & Guilds researchers are also investigating other measurement properties of VQs in both internal and externally-sponsored projects. We are currently looking at areas such as: comparability (between assessment modes, over time, and of demands) and validity (creating a validation framework for observational assessment). Conducting a broad range of research is important for the vocational sector. Not to do so would be to 'concede the point' on quality. Observers (whether well-intentioned or not) would assume that an absence of research findings implied a general absence of quality in VQs. While this is not the case, in the absence of any evidence it can be difficult to argue to the contrary. The current findings (and those of Harth & Hemker, 2011) reveal that, in fact, vocational assessment outcomes are tolerably reliable at worst and highly reliable at best.

- **An opportunity to sponsor research based on large sets of suitable tests.**

The exercise has resulted in the computing and comparison of a variety of reliability coefficients for 68 different tests, based on data sets ranging in size from 82 candidate results to 7,655. The outcomes make a significant contribution to reassuring the public regarding the reliability of these tests and, therefore, the equitable treatment of those candidates who undertake these qualifications.

- **An opportunity to derive meaningful and coherent descriptions of tests – rectifying inconsistencies in current awarding body and regulatory databases.**

Information relating to qualifications, units and assessments held in internal and external databases enabled the development of taxonomies as part of this study. This has proven to be a useful exercise not only as a way of describing, classifying and selecting City & Guilds' tests as part of the research process but also because it gave the opportunity to derive and compare reliability indices for different types of tests eg paper fixed dated, paper fixed on-demand, online fixed dated and online randomly-generated on-demand.

We suggest that this taxonomy could be used in future work. Further work could include analyses to investigate the reliability of tests across other variables such as levels, industries etc. However, before this could be done more widely, some issues (mainly relating to extant databases) would need to be resolved; the main ones are summarised below:

- The way in which assessment type (multiple choice, assignment, portfolio etc.) is classified varies between databases and does not allow particular types (eg short answer question tests) to be clearly identified.

- The way in which units are set up and presented in internal and external databases make it difficult to link information between them: for example, one ‘unit’ on the Ofqual RRQ database could relate to multiple units on the City & Guilds database.
- There is inconsistency in the sector/industry classification between internal City & Guilds databases as well as between internal and external databases.
- There is no single robust source for each of the following indicators: licence to practise, risk and apprenticeship-related units.
- It is not possible to easily obtain general test performance information such as mean score, score range, variance etc. for tests from current systems.

While it is possible for City & Guilds to address some of the issues to do with its own internal systems, other bodies would need to be involved in the further development of external databases. Nonetheless the current work may usefully inform future decision-making when those databases are updated.

- **The application of a range of reliability indices to test score data, including the use of diverse indices beyond the commonly used Cronbach’s alpha.**

We reviewed the field of ‘internal reliability’ as applied to CRT, and undertook a series of analyses using the following main techniques:

- Cronbach’s alpha and Guttman’s lambda-2 to provide (relative) internal consistency coefficients
- Phi for domain dependability (‘absolute’ measurement) and phi (lambda) for cut score applications.
- Contrasting analytical designs for the case of fixed-form and randomly-generated tests.

In addition, two approaches to using Livingston’s squared error loss index and the use of the SEM were trialled on a small scale.

The exercise, and in particular the comparison across indices was very useful overall, in particular the examination of Cronbach’s alpha. We understand that it is a flawed indicator of reliability, but, as Bramley & Dhawan (2010) have noted, it is widely understood and a ‘de facto standard’. As such it is was useful to compare our results with those of previous reliability studies[25]. Further, it is cheering for City & Guilds, and for the vocational sector in general, that our findings were nearly identical to those derived from A levels and GCSEs.

In contrast, we found Guttman's lambda-2 less useful, and we would be less likely to use it operationally for practical purposes within City & Guilds. It is not so widely used as alpha, and, further did not provide substantially different results to that better-known index, and hence did not deliver on the promise envisaged by Sijtsma (2009) of getting closer to the 'true reliability'. At present, we see little reason to recommend its operational use within vocational awarding organisations.

We found the phi and phi (lambda) coefficients to be potentially suitable for modelling reliability in VQ tests. The mean value of phi across the studied tests was the lowest amongst all the coefficients calculated (0.72). This is because phi focuses on absolute measurement, i.e. on how well a test locates individuals absolutely on a test score scale, rather than on relative measurement, which is simply to do with how well a test places individuals on the score scale relative to one another (absolute location irrelevant). Since between-item variance, as well as candidate-item interaction variance, contributes to measurement error in absolute measurement, values of phi are likely to be lower than values of alpha. This is entirely to be expected, given the definition of the index.

In contrast, phi (lambda) provided the highest mean values amongst the coefficients that we calculated. Again, we should not be surprised by this; this essentially amounts to a design feature of squared error loss coefficients (cf. equation 9, above, at p. 25). However, phi (lambda) offers the potential of modelling error variance within a data set, whilst at the same time providing absolute interpretations of domain dependability, with an adjustment which takes into account the distance between the mean score and the cut score. This arguably makes it the best index to answer the question set out in the specification for this project.

We illustrated the workings of phi (lambda) (and its relationship to its logical precursor, phi) in Table 11 and surrounding discussion. From that evidence, it would appear that phi (lambda) has the positive potential to represent the low likelihood of misclassification when average scores are distant from the cut score. However, reflecting on that finding

[25] Effectively comparing awarding organisations with each other, rather than comparing candidates or tests.

(which indeed flows inexorably from the terms in the phi (lambda) equation) leads one to troubling questions about the nature of misclassification as an indicator of measurement quality. If candidate scores are distant from the cut score (passing standard), conventional loss indices (squared error or threshold loss) will be higher. However, such a test would also be poorly targeted. Writing poorly-targeted tests just to boost a loss function would be wrongheaded. Several directions are possible to investigate how to conceptualise classification consistency as something different to mean-to-cut-score distance. Such directions might include conditional SEMs (SEM being inversely related to measurement information) and 'trait-measure-based' reliability indices rather than 'raw-score-based' indices (Andrich, 1982; Linacre, 1997).

Given the unresolved issues in the foregoing discussion, it is possible that, were a vocational awarding organisation to use the phi (lambda) index as part of their operational procedures, they may find that this index's combination of three elements (absolute measurement, cut score location and mean score to cut score distance) makes it harder to understand measurement properties. For this reason, awarding organisations may also wish to look at misclassification, domain dependability and SEM/CIs separately to understand the measurement properties of their tests. It remains to be seen whether the 'everything bundled into a single index' feature of the squared error loss coefficients is in fact a strength, or a factor that might obscure meaning.

Alongside Brennan & Kane's phi (lambda) index, we calculated several examples of Livingston's index. We acknowledge that Livingston's index is less suitable to the case of CRT – on account of its underpinning assumptions about norm-referencing – and that 'plugging' a range of internal consistency coefficients into Livingston's index may lead to unresolved inconsistencies in the variance terms that make up that index. However, bearing such health warnings in mind, awarding body staff may find it useful to calculate Livingston's index for their tests; when KR-21 is used as the internal consistency coefficient it gives similar results to phi (lambda), and it can be calculated relatively easily using a spreadsheet without recourse to specialist software[26].

[26] Of course, any index can be calculated using a powerful modern spreadsheet program. However, as a general rule, it is both easier and better as a matter of principle to use specialist software. But in cases where such software is not available, it is possible to use the spreadsheet, and we found it particularly easy to do so for the Livingston's index.

We particularly emphasise – having done this work – that Livingston's index using the Rasch person separation index needs further work. In particular, users would need to understand the meaning of variance terms within the Rasch person separation index. That in itself might be no bad thing; it might make Rasch theory's treatment of reliability more comprehensive and convincing. It might lead to wider work as well; one can imagine an interrogation of Rasch model assumptions from a G-theory or experimental design perspective that might lead to new insight around the Rasch model's claims to sample-independent measurement.

Finally, we calculated the SEM for three data sets and calculated 95% CIs around mean scores. Our colleagues writing in the sister publication to the current one (Johnson et al, 2012) assert the centrality of SEM in reporting the score precision of CRTs. We agree that SEM is an important indicator of score precision that awarding organisations should use. However, the specification of the current project meant that we only exemplified SEM with a small number of tests. Even these few examples provoked interest, however.

The amount of score precision that is appropriate for test scores in VQs is a matter that depends upon context. For example, we know that – under QCF – free-standing units can be combined with other units to form awards, certificates and diplomas. Such qualifications are of different lengths and contain differing amounts of assessment. Thus, the amount of measurement impression that one might be prepared to tolerate around a score might be lower – say – in the case of a Diploma in which it was known that there was a great deal of other assessment. This might be especially true if that other assessment was of the portfolio type which Harth & Hemker (2011) demonstrated to be highly reliable.

Implicit in our comparison of the utility of different indices is the suggestion that awarding bodies should use reliability research to enhance their operational procedures. That is City & Guilds' intention. We believe that the insights that have emerged from this research can be used to inform and support decision-making in a range of contexts such as results determination (City & Guilds, 2008) and item bank reviews. In doing this, we will seek insight from the small literature that discusses how to present statistical information derived from assessment to users in the most informative and meaningful manners (eg Zenisky, Hambleton & Sireci, 2009; Hattie, undated).

- **An understanding of fundamental issues in vocational and – indeed – all forms of assessment; relationship between CRT and NRT, role of knowledge testing in VQs, empirical basis for statements about competence-based assessment, etc.**

In this part of the discussion section, we look at some of the underlying measurement issues that flow from this research. Firstly, we look at the measurement approaches that can be used to investigate reliability, then we go on to discuss the characterisation of reliability as replication.

The indices used in this research come from three measurement traditions: Classical Test Theory (CTT), Item Response Theory (IRT) and Generalizability theory (G-theory). As we have noted, the limitations of CTT have been widely documented. And yet, we take the view at the end of this research that CTT nonetheless retains some use for awarding organisations. Cronbach's alpha is simple to calculate using standard software such as SPSS, and results can be compared between awarding bodies. It seems a starting point for reliability analysis; although hopefully not the end point for those with genuine commitment to understanding how their tests are measuring.

We also used G-theory and IRT approaches to reliability estimation. For the City & Guilds researchers on this project, this was our first experience of using G-theory[27] (although we had read about it extensively). Our reflection is that we found the claims of G-theory to absolute measurement, and especially the principled treatment of potential sample-to-population bias, to be highly congruent with CRT and competence-based measurement. In truth, the project provided a brief introduction to G-theory through the derivation of the base form of indices. It remains to be seen whether we will find an operational use for some of the more sophisticated applications, such as estimating the contribution of a wider range of error variance components (see Cardinet, Johnson & Pini, 2010, p. 14).

In contrast, we perceived IRT (especially the Rasch model that we used) to be less consonant with the core philosophy of CRT. The central aim of Rasch – to array candidate abilities along a single latent trait – seems to us to be at odds with the central aims of CRT and competence-based assessment. Those are about demonstrating competence and knowledge in broad domains; this is why G-theory is in principle

[27] Although Assessment Europe was a key member of the project team because of their expertise with G-theory.

superior – it really is ‘the domain’ that matters. Having said this, however, we also found that most of our data sets fitted the Rasch model well. So, while in principle the Rasch approach is at odds with the underpinning philosophy, in practice it appears to fit the data. It is difficult to know what to conclude from this finding. One interpretation is that CRTs and NRTs are only distinguishable at the level of score interpretations and there is no way of ‘spotting’ a CRT or NRT merely by observing the test and the data set generated from it. Or, alternatively, it may be that the Rasch model is an empirical generalisation (Bass, 1985); that is, a pattern in data that repeats itself robustly in a (wide) range of circumstances. Over time this may form the basis for further inductive theory building (Locke, 2007).

Two fundamental issues about reliability emerged from the findings of this research. They relate, firstly, to the values of reliability coefficients when data sets have relatively low variance (scores are ‘bunched’ in one part of the potential scoring scale, rather than spread all along the ostensible scale). Secondly, they relate to the values of reliability coefficients when the mean score is close to the cut score. The impact of each of these two separate issues is magnified when they co-occur.

By definition, reliability coefficients that contain a term for population (or universe) score variance will tend to be lower when that variance was lower. However, when we investigated the values of coefficients on meso- and leptokurtic data sets (see Table 10, above at p. 57) we did not find the more broadly spread distributions to be more reliable. We treat this finding with caution. It may be simply that kurtosis is not an ideal proxy for spread of scores (Wuensch, 2011). Also, a more thoroughgoing G-theory analysis would likely show that it is not just between-candidate variance that determines coefficient values but also the amount of candidate-item interaction variance, plus between-item variance in the case of absolute measurement coefficients. Such thorough treatments of test-score variance might provide a more nuanced understanding of reliability than simply worrying about narrowness of spread.

Further research might also seek to account for differences in reliability findings for tests delivered (or generated) via different modes (Figure 7, above at p. 56). Such research would need to look more deeply at differences in reliability indices between fixed-form and randomly-generated tests. In doing so, it would need to confirm that there was a substantial difference in reliability co-efficients for these two modes of test generation, and that the differences found in this research were not due to some other cause, such as the nature of the subjects being tested. Research on this issue could be conducted using a G-theory analysis looking at several facets, including inter-item variance. Alternatively, comparability studies could be conducted (for instance, using the Rasch model) side-by-side with reliability analysis.

Certainly, however, we would not wish to see CRT designers introduce more broadly-spread score distributions. We take the view that the narrowly-spread score distribution is a design feature of a CRT and should not be tampered with to inflate reliability coefficients.

Similarly, we observed situations where the cut score is close to the mean score on a test (see Figure 10, above at p. 61). On those occasions, there does – on the face of it – appear to be a major risk of misclassification. In some cases, candidates appear to have learned 'just enough to pass'. An observer might be concerned that such, 'just-passing' candidates might forget what they had learned and be likely to be classified as 'fail' if they were to re-sit the test. This is a genuine concern, but it misunderstands the nature of VET. Candidates can be assessed many times (see: Harth & Hemker, 2011 and Baird et al, 2011, p. 53). Also, and especially in the case of workplace qualifications, assessment forms part of continuing professional development, and candidates do not 'stand still'; they learn more and practise their skills in different contexts. Definitions of reliability as replication (such as the Ofqual/Newton definition cited in this paper at p. 13) imply that the fact that people change between potential (or simulated) replications is a derogation from the most principled case; that in an ideal world it would be possible to observe two situations between which the subject (the people) did not change. In fact, we say that the case of VET shows that replication in this sense is not possible, and nor should it be. We would want people to change (learn) between the two observations.

Further, the steps that one might take to increase reliability coefficients are – as in the case of constrained distributions – not always desirable. In order to 'move' the mean score away from the cut score, the test content could be manipulated (for instance to add more very easy or very difficult items). But we would not recommend such an action, because although it might increase the value of reliability coefficients and decrease the proportion of 'misclassifications', it would also make the test less well targeted and less efficient. These features are properties of measurement that high quality awarding organisations should aspire too. In short, more reliable tests are not always better tests in validity terms – though both high validity and high reliability can be simultaneously achieved.

Conclusions and recommendations

Overall, the work presented a positive assessment of the reliability of vocational tests. By building on the work by Harth & Hemker (2011) and in parallel to the City & Guilds partnership's other project in the current round of tenders, the research contributes to the emerging body of reliability research on VQs. Through dissemination of the findings through this report and conferences and articles we hope this will provide a platform for the further development of research into vocational assessments. We note that this report may not be in a format that is entirely accessible to officers who are directly involved in managing assessments on a day-to-day basis and our next actions will be to consider how best to present this information in a format that will be accessible, informative and meaningful to users.

Based on the findings of this study, we make the following additional conclusions and recommendations:

- The research demonstrated that it is possible for VQs to provide reliable outcomes. The mean values for Cronbach's alpha in this project were very similar to those returned in reliability analysis of GCSEs and A levels.

- Selecting a sample of tests for analysis was inhibited by inconsistencies between regulatory and awarding body databases. City & Guilds and Ofqual should review and update their respective databases in light of the issues raised in this report and work collaboratively to do so where this is mutually advantageous.

- The classical test theory index Cronbach's alpha is a starting point for reliability analysis. It has limitations – particularly when applied to CRTs – but its ubiquity and ease of calculation mean that all credible awarding organisations ought to be able to calculate and interpret it.

- We did not find Guttman's lambda-2 index to add much value over and above that of alpha, and would not recommend its use operationally by awarding organisations.

- The phi coefficient seemed intuitive to us, proved relatively straightforward to calculate and appeared to have a basis in principle that was suitable for CRT applications. We recommend that vocational awarding organisations investigate its use operationally.

- Phi (lambda) has similar philosophical bases to phi and thus appears to be useful. It also has the potential to ‘bundle’ absolute measurement and an adjustment to account for a cut score in a single index. This may make it a very efficient index for our purposes. However, awarding organisations may consider that this ‘bundling’ creates confusion and may prefer to keep the cut score and dependability aspects of reliability estimation distinct.

- Livingston’s index is a less defensible squared error loss index than phi (lambda), but it could be used (with suitable health warnings) by an awarding body that was not able to calculate phi (lambda).

- The derivation of SEM and associated CIs is a reputable technique to show the reliability of test outcomes in terms of score precision. The interpretation of the amount of score precision necessary depends upon context. High score precision in other assessed elements within a qualification might mitigate low precision within the measurement from a knowledge test.

- We noted two concerns for estimating the reliability of CRT outcomes which need to be thought through. When scores were ‘bunched’ in only a section of the scoring scale, reliability indices ought – in principle – to be low. Also, a mean score that was very close to a cut score ought to be a concern – suggesting high levels of misclassification. Such concerns suggest that reliability analysis is best conducted in tandem with informed validity research.

References

Andrich, D. (1982) An index of person separation in latent trait theory, the traditional KR-20 index, and the Guttman scale response pattern. *Education Research and Perspectives*, 9(1), 95-104.

Andrich, D. (2004) Controversy and the Rasch model: a characteristic of incompatible paradigms? *Medical Care*, 42(1), 1-7.

Baird, J., Black, P., Béguin, A., Pollitt, A. & Stanley, G. (2011) *The reliability programme: final report of the Technical Advisory Group*. (Coventry: Ofqual).

Barnes, J. (1992) The accreditation of prior learning, *Education Today*, 42(1), 11-17.

Bass, F.M. (1995) Empirical generalizations and marketing science: a personal view, *Marketing Science*, 14(3), part 2 of 2, G6-G19.

Bees, M. & Swords, M. (Ed.s) (1990) *National Vocational Qualifications and Further Education*. (London: Kogan Page/National Council for Vocational Qualifications).

Bramley, T. & Dhawan, V. (2010) *Estimates of the reliability of qualifications.* (Coventry: Ofqual).

Brennan, R.L. & Kane, M.T. (1977) An index of dependability for mastery tests. *Journal of Educational Measurement*, 14(3), 277-289.

Brennan, R.L. (2001). *Generalizability theory*. (New York: Springer-Verlag).

Brennan, R.L. (2003) Coefficients and indices in generalizability theory. *Center for Advanced Studies in Measurement and Assessment (CASMA) Research Report Number 1*. http://www.education.uiowa.edu/centers/Libraries/CASMA_Research/01casmareport.sflb.ashx.

British Broadcasting Corporation (BBC) (2011) *Apprenticeships rise by a half, data shows*. http://www.bbc.co.uk/news/education-15477990.

Brockmann, M., Clarke, L. & Winch, C. (eds) (2011). *Knowledge, Skills and Competence in the European Labour Market. What's in a vocational qualification?* (Abingdon and New York: Routledge).

Brown, J.D. (1990) Short-cut estimators of criterion-referenced test consistency. *Language Testing*, 7(1), 77-97.

Callender, J.C. & Osburn, H.G. (1979) An empirical comparison of coefficient alpha, Guttman's lambda - 2, and MSPLIT maximized split-half reliability estimates. *Journal of Educational Measurement*, 16(2), 89-99.

Cardinet, J., Johnson, S., & Pini, G. (2010) *Applying Generalizability Theory with EduG*. (New York: Routledge).

City & Guilds (1993) *City and Guilds of London Institute: a short history, 1878 – 1992.* (London: City and Guilds of London Institute).

City & Guilds (2008) *Result determination manual*. (Version 1.1, Dec 08) (Unpublished document).

City & Guilds (2011a) *Tender for estimation of internal reliability: OF126*. Unpublished tender submitted to Ofqual, September 2011.

City & Guilds (2011b) *Licence to practise: special report*. http://www.cityandguilds.com/21094.html.

Cronbach, L.J. & Shavelson, R.J. (2004) My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement*, 64, 391-418.

Cronbach, L.J. (1951) Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297-334.

Cronbach, L.J., Gleser, G.C., Nanda, H. & Rajaratnam, N. (1972). *The dependability of behavioral measurements: theory of generalizability for scores and profiles*. (New York: Wiley).

Embretson, S.E. & Reise, S.P. (2000) *Item response theory for psychologists.* (Mahwah, NJ: Lawrence Erlbaum Associates).

Evans, D. (2007) *The history of technical education: a short introduction* (second edition). http://www.tmag.co.uk/extras/history_of_Technical_Education_v2.pdf.

Evans, D. (2008) *History of technical and commercial examinations: a reflective commentary*. http://www.tmag.co.uk/extras/history_of_Technical_Commercial_Exams08.pdf.

Feldt, L.S. & Brennan, R.L. (1989) Reliability in Linn, R.L. (Ed.) *Educational measurement* (third edition). (Washington, DC: American Council on Education/Macmillan).

Gonzci, A. (2001) Competency-based learning: a dubious past – an assured future? In Boud, D. & Garrick, J. *Understanding learning at work*. (London: Routledge).

Gravetter, F.J. & Wallnau, L.B. (1996) *Statistics for the behavioral sciences: a first course for students of psychology and education* (fourth edition). (Minneapolis/St Paul: West Publishing Company).

Greatorex, J. (2005) Assessing the evidence: different types of NVQ evidence and their impact on reliability and fairness. *Journal of Vocational Education and Training,* 57(2), 149-164.

Guttman, L. (1945) A basis for analyzing test-retest reliability. *Psychometrika*, 10, 255-282.

Haertel, E.H. (2006) Reliability in R. L. Brennan (Ed.) *Educational measurement* (fourth edition). (Westport, CT: American Council on Education/Praeger).

Hambleton, R.K., Swaminathan, K., Algina, J. & Coulson, D. (1975) Criterion-Referenced Testing and measurement: a review of technical issues and developments. Paper presented at the *Annual Meeting of the American Educational Research Association* (Washington, D. C., March 30-April 3, 1975). ERIC ID: ED107722. http://www.eric.ed.gov/PDFS/ED107722.pdf.

Harth, H. & Hemker, B. (2011) *On the reliability of results in vocational assessment: the case of work-based certification*. (Coventry: Ofqual).

Harth, H. & van Rijn, P. (2010) *Reliability issues in competence-based assessment: concepts and estimates.* Interim report submitted to Ofqual as part of Reliability Programme, April 2010.

Harvill, L.M. (1991) Standard Error of Measurement. *Educational Measurement: Issues and Practice*, 10, 33-41.

Hattie, J. (Undated) Visibly learning from reports: the validity of score reports. *Online Educational Research Journal*. http://www.oerj.org/View?action=viewPDF&paper=6.

He, Q. (2009) *Estimating the reliability of composite scores*. (Coventry: Ofqual).

Hutchison, D. & Benton, T. (2009) *Parallel universes and parallel measures: estimating the reliability of test results*. (Coventry: Ofqual).

Jessup, G. (1991) *Outcomes: NVQs and the emerging model of education and training.* (London: The Falmer Press).

Johnson, M. (2006) A review of vocational research in the UK 2002-2006: measurement and accessibility issues. *International Journal of Training Research*, 4(2), 48-71.

Johnson, M. (2008) Assessing at the borderline: judging a vocationally-related portfolio holistically. *Issues in Educational Research*, 18(1).

Johnson, S. & Johnson, R. (2010) *Conceptualising and interpreting reliability*. (Coventry: Ofqual).

Johnson, S. & Johnson, R. (2011) *Component reliability in GCSE and GCE*. (Coventry: Ofqual).

Johnson, S., Johnson, R., Miller, L. & Boyle, A. (2012) *Reliability of vocational assessment: an evaluation of level 3 electro-technical qualifications*. (Forthcoming, Coventry: Ofqual).

Koeppen, K., Hartig, J., Klieme, E. & Leutner, D. (2008) Current issues in competence modelling and assessment. *Journal of Psychology*, 21(6), 61-73.

Kuder, G.F. & Richardson, M.W. (1937) The theory of the estimation of test reliability. *Psychometrika*, 2, 151-160.

Lang, J. (1978) *City and Guilds of London Institute: centenary 1878 – 1978.* (London: City and Guilds of London Institute).

Linacre, J.M. (1997) KR-20/Cronbach alpha or Rasch reliability: which tells the 'truth'? *Rasch Measurement Transactions*, 11(3), 580-581.

Linacre, J.M. (2009) *A User's Guide to WINSTEPS® & MINISTEP. Rasch-Model computer programs*. Program Manual 3.68.0. http://www.winsteps.com/a/winsteps3682.pdf.

Livingston, S.A. (1972) Criterion-referenced applications of classical test theory. *Journal of Educational Measurement*, 9(1), 13-26.

Locke, E A (2007) The case for inductive theory building. *Journal of Management*, 33(6), 867-890.

Meretoja, R., Isoaho, H., & Leino-Kilpi, H. (2004) Nurse Competence Scale: development and psychometric testing. *Journal of Advanced Nursing*, 47(2), 124-133.

Murphy, R., Burke, P., Content, S., Frearson, M., Gillispie, J., Hadfield, M., Rainbow, R., Wallis, J. & Wilmut. J. (1995) *The reliability of assessment of NVQs. Report presented to NCVQ*. http://www.nottingham.ac.uk/shared/shared_cdell/pdf-reports/nvqrelrep.pdf.

Newton, P.E. (2009) The reliability of results from national curriculum testing in England. *Educational Research*, 51(2), 181-212.

Nichols, P.D. & Smith, P.L. (1998) Contextualizing the interpretation of reliability data. *Educational Measurement: Issues and Practice*, 17(3), 24-36.

Norris, N. (1991) The trouble with competence. *Cambridge Journal of Education*, 21(3), 331-344.

Novakovic, N. & Greatorex, J. (2011) Comparing the demand of syllabus content in the context of vocational qualifications: literature, theory and method. *Research Matters*, 11, 25-31.

Oates, T. (2007) The constraints on delivering public goods – a response to Randy Bennett's 'What does it mean to be a non-profit educational measurement organization in the 21st Century?' A paper presented at the *International Association for Educational Assessment Annual Conference*, Baku, Azerbaijan, September 2007. http://www.iaea.info/documents/paper_1162d20f28.pdf.

Office of Qualifications and Examinations Regulation (Ofqual) (2011a) *OF126 – estimation of internal reliability specification*. (Coventry: Ofqual).

Office of Qualifications and Examinations Regulation (Ofqual) (2011b) *Annual qualifications market report: version two – August 2011*. (Coventry: Ofqual).

Office of the Qualifications and Examinations Regulator (Ofqual) (2008a) *Regulatory arrangements for the Qualifications and Credit Framework*. (Coventry: Ofqual).

Office of the Qualifications and Examinations Regulator (Ofqual) (2008b) *Operating rules for using the term 'NVQ' in a QCF qualification title*. (Coventry: Ofqual).

Osterlind, S.J. (1988) Using CRTs in program curriculum evaluation. *Educational Measurement: Issues and Practice*, 7(3), 23-30.

Panayides, P., Robinson, C. & Tymms, P. (2010) The assessment revolution that has passed England by: Rasch measurement. *British Educational Research Journal*, 36(4), 611-626.

Parkes, J. (2007) Reliability as argument. *Educational Measurement: Issues and Practice*, 26(4), 2-10.

Pearson, K. (1905) Das Fehlergesetz und seine Verallgemeinerungen dürch Fechner und Pearson. A Rejoinder. *Biometrika*, 4, 169-212.

Qualifications and Curriculum Development Agency (QCDA) (2010) *Combining knowledge and skills in units of competence-based vocational qualifications in the QCF.* (Coventry: QCDA).

Raykov, T. & Marcoulides, G.A. (2011) *Introduction to psychometric theory*. (New York: Routledge).

Ripley, M. (2004) *Speech to e-assessment conference, Tuesday, 20 April 2004, The Royal Festival Hall*.

Robinson, C. (2007) Awarding examination grades: current processes in Newton, P., Baird, J-A., Goldstein, H., Patrick, H. & Tymms, P. (Eds.) *Techniques for monitoring the comparability of examination standards.* (London: QCA).

Sands, W.A., Waters, B.K., McBride, J.R. (Eds.) (1997) *Computerized adaptive testing: from inquiry to operation*. (Washington, DC: American Psychological Association).

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, 74(1), 107-120.

Stanley, J.C. (1971) Reliability in Thorndike, R.L. (Ed.) *Educational measurement* (second edition). (Washington, DC: American Council on Education/Macmillan).

Swiss Society for Research in Education Working Group (SSREWG) (2010) *EduG user guide*. http://www.irdp.ch/edumetrie/documents/EduGUserGuide.pdf.

Tattersall, K. (2007) A brief history of policies, practices and issues relating to comparability in Newton, P., Baird, J-A., Goldstein, H., Patrick, H. & Tymms, P. (Eds.) *Techniques for monitoring the comparability of examination standards.* (London: QCA).

Thissen, D. & Wainer, H. (2001) *Test scoring.* (Mahwah, NJ: Lawrence Erlbaum Associates).

Tight, M. (2002) *Key concepts in adult education and training; second edition*. (Abingdon, Oxon: RoutledgeFalmer).

Traub, R.E. & Rowley, G.L. (1980) Reliability of test scores and decisions. *Applied Psychological Measurement*, 4(4), 517-545.

Traub, R.E. & Rowley, G.L. (1991) Understanding reliability. *Educational Measurement: Issues and Practice*, 10(1), 37-45.

Wainer, H. (Ed.) (2000) *Computerized adaptive testing: a primer* (second edition). (Mahwah, NJ: Lawrence Erlbaum Associates).

Watts, A. (2008) Independent examination boards and the start of a national system. *Research Matters*, 5, 2-6.

Wheadon, C., Whitehouse, C., Spalding, V., Tremain, K. & Charman, M. (2009) *Principles and practice of on-demand testin*g. (Coventry: Ofqual).

Wolf, A. (2011) *Review of vocational education – the Wolf report*. https://www.education.gov.uk/publications/eOrderingDownload/The%20Wolf%20Report.pdf.

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

Wuensch, K.L. (2011) *Skewness, kurtosis, and the normal curve*. http://core.ecu.edu/psyc/wuenschk/docs30/Skew-Kurt.docx.

Zenisky, A.L., Hambleton, R.K., & Sireci, S.G. (2009) Evaluating the utility of NAEP reporting practices. *Applied Measurement in Education*, 22, 359–375.

Appendices

Appendix 1: Data preparation, issues and quality assurance

Paper-based, dated, fixed form MC tests

Where a different version of these tests was run on more than one occasion during a selected period between 2009 and 2011, a single exam sitting was chosen for each test. This selection was based (in most cases) on the highest number of candidates sitting each test. Tests with only one exam date during the specified period were selected by default. Data for selected tests were then explored for errors and anomalies.

The following steps were taken or decisions were made in preparing the data:

- Indicator for right/wrong answer for each item was recoded from ‘Y’ to ‘1’ indicating correct answer and from ‘N’ to ‘0’ representing wrong answer
- Candidate response was recoded from ‘O’ to ‘0’ where no answer was given
- Items excluded from the test due to errors (coded ‘X’ in the original dataset) were excluded from analysis
- Candidates with both missing grade and zero score were excluded
- Candidates with ‘absent’ or ‘work not submitted’ status were also excluded
- Retakes were not excluded (for any test types)

The number of items and candidates excluded are shown for each test in Appendix 3. Data were saved in various formats for analysis e.g. Excel, SPSS and text.

Paper-based, on-demand, fixed-form MC tests

The same procedure was applied to paper-based, on-demand fixed-form tests. The only exception was that each test version was run over a period of time rather than on a specific date.

Online, dated, fixed-form MC tests

The raw data from SAP was imported into Excel and a version of the test was chosen based on the number of candidates. Candidates’ answers to individual items (i.e. ‘a’ ‘b’ ‘c’ ‘d’) were then converted into ones and zeros based on the answer key provided. The sum of individual items was compared to the overall score recorded to quality assure the transformation.

Data exploration revealed missing answers to individual items. These were recoded to zero; in addition any candidates for whom no answers had been recorded for any of the items in the test were excluded from the analysis (see Appendix 3, below, at p. 85). Finally, data were imported into SPSS for analysis and in addition saved as a text file to enable other relevant analyses to be performed.

Online, on-demand randomly generated MC tests

The steps taken to prepare online, on-demand, randomly-generated tests were the same as those taken for online, dated, fixed form tests, as described above. The exceptions were that:

- candidates who did not have answers for the required number of questions were excluded to enable the relevant analysis to be performed
- one item was excluded from one test due to errors and was removed from the analysis

the data file for randomly-generated tests is different to the fixed-form equivalent because it includes all items in a bank; therefore it contains missing values for items that candidates were not given as part of their test.

Appendix 2: List of tests and features

Unit number	Title	Mode	Fixed/ random	Dated/ on-demand	Level	Industry	LtP	App	Risk
2230-001	Microcomputer Technology	Paper	Fixed	Dated	1	IT			Low
6958-013	PC Technology	Paper	Fixed	Dated	2	Building Services		Yes	Low
3692-403	Reading	Paper	Fixed	Dated	2	Skills for Work and Life		Yes	Low
6958-010	Consumer and Commercial Electronics Core Studies 2	Paper	Fixed	Dated	2	Building Services		Yes	Low
6958-009	Consumer and Commercial Electronics Core Studies 1	Paper	Fixed	Dated	2	Building Services		Yes	Low
7068-001	Accommodation Operations and Services Principles	Paper	Fixed	Dated	1	Hospitality and Catering			
7067-001	Reception Operations and Services Principles	Paper	Fixed	Dated	1	Hospitality and Catering			
6165-011	Construction Technician Principles 1	Paper	Fixed	Dated	1	Construction and Energy			Low
1121-001	Retailing Principles 1	Paper	Fixed	Dated	1	Retail and Warehousing			Low
1123-001	Beauty Therapy Skills Principles 1	Paper	Fixed	Dated	1	Beauty Therapy			Low
6165-001	Core Construction Skills Principles	Paper	Fixed	Dated	1	Construction and Energy			Low
6161-037	Electrical Installation 2 Principles	Paper	Fixed	Dated	2	Construction and Energy			Low
3692-303	Reading	Paper	Fixed	Dated	1	Skills for Work and Life		Yes	Low
6165-002	Basic Construction Skills Principles	Paper	Fixed	Dated	1	Construction and Energy			Low
6161-036	Refrigeration and Air Conditioning 2 Principles	Paper	Fixed	Dated	2	Construction and Energy			Low
4867-003	Business Aspects of International Tourism 1	Paper	Fixed	Dated	1	Travel, Tourism, Aviation			
4867-002	International Tourism Principles 1	Paper	Fixed	Dated	1	Travel, Tourism, Aviation			
4867-015	International Tourism Operations 2	Paper	Fixed	Dated	2	Travel, Tourism, Aviation			
3638-011	Information and Communication Technology	Paper	Fixed	Dated	2	Skills for Work and Life		Yes	Low
1122-001	Hairdressing Skills Principles 1	Paper	Fixed	Dated	1	Hairdressing			Low
6161-001	Core Construction Skills Principles	Paper	Fixed	Dated	1	Construction and Energy			Low
3638-003	Application of Number	Paper	Fixed	Dated	1	Skills for Work and Life		Yes	Low
3638-001	Communication	Paper	Fixed	Dated	1	Skills for Work and Life		Yes	Low
3638-005	Information and Communication Technology	Paper	Fixed	Dated	1	Skills for Work and Life		Yes	Low
6165-021	Applied Scientific Techniques 2 Principles	Paper	Fixed	Dated	3	Construction and Energy			Low
7065-018	Patisserie Principles	Paper	Fixed	Dated	2	Hospitality and Catering			
7065-016	Food Preparation/Cooking (Culinary Arts) Principles 2	Paper	Fixed	Dated	2	Hospitality and Catering			
7545-201	Nutrition and Health of School Aged Children	Paper	Fixed	On-demand	2	Learning			Low

Unit number	Title	Mode	Fixed/ random	Dated/ on-demand	Level	Industry	LtP	App	Risk
0065-503	Horticulture (Sports Turf)	Paper	Fixed	On-demand	2	Land Based Services		Yes	Low
1886-202	Principles of Conflict Management	Paper	Fixed	On-demand	3	Security			Low
0067-500	Animal Care (Animal Care and Welfare)	Paper	Fixed	On-demand	2	Land Based Services		Yes	Medium
7013-013	Introductory Certificate in Conflict Handling	Paper	Fixed	On-demand	1	Business Skills			Low
0065-500	Horticulture (Seeding and Planting)	Paper	Fixed	On-demand	2	Land Based Services		Yes	Low
7015-015	Introductory Certificate in Selling	Paper	Fixed	On-demand	1	Business Skills			Low
1892-010	Working as a Security Officer	Paper	Fixed	On-demand	2	Security	Yes	Yes	High
3681-030	Safe Working Principles	Paper	Fixed	On-demand	2	Business Skills			Low
7104-603	Staff Working in Scottish Licensed Premises	Paper	Fixed	On-demand	2	Hospitality and Catering			Low
1892-012	Working as a Door Supervisor	Paper	Fixed	On-demand	2	Security	Yes	Yes	High
1892-013	Conflict Management for the Private Security Industry	Paper	Fixed	On-demand	2	Security	Yes	Yes	High
1892-009	Working in the Private Security Industry	Paper	Fixed	On-demand	2	Security	Yes	Yes	High
7014-014	Introductory Certificate in Customer Service	Paper	Fixed	On-demand	1	Business Skills			Low
3792-300	Adult Numeracy	Paper	Fixed	On-demand	1	Skills for Work and Life		Yes	Low
4475-502	Principles of providing administrative services	Online	Fixed	Dated	2	Business Skills		Yes	Low
2251-501	Working Safely in an Engineering Environment	Online	Fixed	Dated	1	Engineering		Yes	Low
3638-978	Information and Communication Technology	Online	Fixed	Dated	2	Skills for Work and Life		Yes	Low
3638-975	Communication	Online	Fixed	Dated	1	Skills for Work and Life		Yes	Low
3792-986	Adult Literacy	Online	Fixed	Dated	2	Skills for Work and Life		Yes	Low
7579-311	3D CAD	Online	Random	On-demand	3	Manufacturing Industry			Low
7564-303	Knowledge of Anatomy, Physiology and Pathologies	Online	Random	On-demand	3	Beauty Therapy			Low
1892-008	Working as a Close Protection Operative	Online	Random	On-demand	3	Security	Yes		High
3667-102	Fibre Optic Cabling in an Internal Environment	Online	Random	On-demand	2	IT			Low
4871-106	Air fares and ticketing	Online	Random	On-demand	1	Travel, Tourism, Aviation			Low
0351-200	NPTC Introduction to art and design for florists	Online	Random	On-demand	2	Land Based Services			Low
4873-399	UK Travel and Tourism Destinations	Online	Random	On-demand	3	Travel, Tourism, Aviation		Yes	Low
2800-601	Introduction to Engineering	Online	Random	On-demand	1	Engineering		Yes	Low
4101-130	Diagnose/Rectify Vehicle Transmission System Faults	Online	Random	On-demand	3	Transport Maintenance		Yes	Low
4835-201	Understanding Employment Rights and Responsibilities	Online	Random	On-demand	2	Sport and Recreation		Yes	
6217-099	Basic Construction Skills	Online	Random	On-demand	1	Construction and Energy			Low
0361-101	Safe/effective working practices in land-based industries	Online	Random	On-demand	1	Land Based Services			
4872-201	Worldwide Travel and Tourism Destinations	Online	Random	On-demand	2	Travel, Tourism, Aviation		Yes	Low

Unit number	Title	Mode	Fixed/ random	Dated/ on-demand	Level	Industry	LtP	App	Risk
4417-201	Delivery of effective customer service	Online	Random	On-demand	2	Business Skills		Yes	Low
6314-103	Construction Diploma - Plastering	Online	Random	On-demand	1	Construction and Energy			Low
7540-664	Plan for Delivery of ICT Support Services	Online	Random	On-demand	3	IT		Yes	Low
4101-114	Diagnose/Rectify Vehicle Engine System Faults	Online	Random	On-demand	3	Transport Maintenance		Yes	Low
4101-702	Carry Out Basic Routine Maintenance	Online	Random	On-demand	1	Transport Maintenance		Yes	Low
6926-022	Salon services	Online	Random	On-demand	1	Hairdressing			Low
2079-101	F Gas and ODS Regulations: Category I	Online	Random	On-demand	2	Building Services	Yes		High
2382-200	Certificate in Requirements for Electrical Installations	Online	Random	On-demand	3	Building Services			Medium

Notes:

- Some unit titles were amended slightly for presentational purposes.
- LtP = Licence to Practise and App = Apprenticeship

Table 13: List of tests and features

Appendix 3: Summary statistics for each test

Paper fixed dated tests

Unit number	Exam date/period	Items to answer	Items removed	Mean score	Score variance	Kurtosis	Kurtosis SE	Kurtosis type	Minimum score	Maximum score	Cut score	Candidates	Candidates passed	Candidates removed	Cronbach's alpha	Guttman's lambda-2	Phi	Phi (lambda)
1121-001	24/06/2010	39	1	33.5	22.5	2.6	0.4	Lepto	13	39	19	163	98.8%		0.83	0.84	0.82	0.98
1122-001	01/06/2009	38	2	30.7	24.4	1.9	0.3	Lepto	9	38	18	370	97.3%		0.81	0.82	0.79	0.97
1123-001	04/12/2009	39	1	32.8	27.4	3.2	0.4	Lepto	11	39	19	168	96.4%		0.85	0.86	0.83	0.98
2230-001	12/06/2009	75	-	49.0	125.9	0.1	0.5	Meso	21	67	30	100	92.0%		0.90	0.90	0.88	0.97
3638-001	02/03/2010	40	-	28.3	47.0	-0.1	0.3	Meso	6	40	26	242	66.9%		0.86	0.87	0.85	0.86
3638-003	02/03/2009	40	-	28.1	51.1	-0.7	0.3	Platy	10	40	23	219	77.2%		0.88	0.89	0.86	0.90
3638-005	06/05/2010	40	-	31.4	30.9	0.8	0.2	Lepto	13	40	29	409	76.8%		0.84	0.85	0.81	0.83
3638-011	07/05/2009	40	-	29.2	16.6	0.4	0.3	Meso	14	38	28	270	67.8%		0.69	0.70	0.60	0.60*
3692-303	08/06/2010	40	-	23.6	46.6	-0.6	0.3	Meso	7	37	27	197	67.5%		0.84	0.85	0.82	0.85
3692-403	02/03/2010	40	-	27.9	62.3	-1.2	0.5	Platy	11	38	25	82	61.0%		0.90	0.91	0.89	0.90
4867-002	03/12/2009	40	-	24.2	31.1	-0.9	0.3	Platy	11	37	20	315	76.8%		0.76	0.77	0.72	0.82
4867-003	03/12/2009	40	-	28.3	27.0	-0.3	0.3	Meso	11	39	20	339	95.0%	1	0.75	0.76	0.72	0.92
4867-015	01/12/2010	50	-	33.0	46.0	-0.4	0.2	Meso	13	47	25	391	88.5%		0.80	0.81	0.78	0.90
6161-001	02/06/2009	50	-	36.8	52.2	0.6	0.2	Lepto	10	48	25	387	93.0%	1	0.85	0.86	0.83	0.95
6161-036	06/12/2011	98	2	70.6	158.0	0.5	0.3	Meso	21	90	49	248	94.8%		0.90	0.91	0.89	0.97
6161-037	09/06/2009	100	-	71.5	202.4	1.5	0.3	Lepto	15	97	50	258	92.2%		0.92	0.93	0.91	0.97
6165-001	07/06/2011	50	-	38.8	28.1	0.3	0.3	Meso	20	49	25	317	98.4%	3	0.77	0.78	0.72	0.96

Unit number	Exam date/period	Items to answer	Items removed	Mean score	Score variance	Kurtosis	Kurtosis SE	Kurtosis type	Minimum score	Maximum score	Cut score	Candidates	Candidates passed	Candidates removed	Cronbach's alpha	Guttman's lambda-2	Phi	Phi (lambda)
6165-002	07/06/2011	40		29.0	28.2	-0.1	0.2	Meso	12	37	20	404	93.3%	2	0.80	0.81	0.75	0.93
6165-011	08/06/2011	99	1	72.0	204.3	1.4	0.3	Lepto	3	97	49	256	93.4%		0.93	0.93	0.91	0.98
6165-021	06/06/2011	90	-	64.4	249.6	-0.4	0.2	Meso	1	90	45	557	86.2%		0.94	0.94	0.94	0.97
6958-009	08/06/2011	70	-	50.5	136.8	0.1	0.4	Meso	13	67	35	131	88.5%		0.92	0.93	0.91	0.97
6958-010	10/06/2010	50	-	36.8	89.5	-0.6	0.4	Meso	14	50	25	124	87.1%		0.92	0.92	0.91	0.96
6958-013	16/06/2011	39	1	32.8	20.6	0.3	0.5	Meso	18	39	19	92	98.9%		0.79	0.81	0.77	0.98
7065-016	01/11/2011	96	4	70.8	136.5	0.5	0.1	Lepto	20	93	48	2,250	95.7%	7	0.89	0.89	0.87	0.97
7065-018	02/11/2011	80	-	63.3	138.5	0.5	0.2	Lepto	19	80	40	909	94.8%	1	0.92	0.93	0.92	0.98
7067-001	01/11/2011	60	-	42.3	38.4	0.8	0.3	Lepto	22	55	30	236	97.0%		0.72	0.74	0.70	0.94
7068-001	08/11/2011	60	-	40.3	33.2	0.7	0.4	Meso	22	51	30	150	94.0%		0.68	0.70	0.64	0.91

*Phi (lambda) was less than phi so phi taken as final phi (lambda) value as it is the lower bound

Table 14: Summary statistics for paper fixed dated tests

Paper fixed on-demand tests

Unit number	Exam date/period	Items to answer	Items removed	Mean score	Score variance	Kurtosis	Kurtosis SE	Kurtosis type	Minimum score	Maximum score	Cut score	Candidates	Candidates passed	Candidates removed	Cronbach's alpha	Guttman's lambda-2	Phi	Phi (lambda)
0065-500	2011	36	-	24.2	16.4	-0.1	0.2	Meso	11	32	18	452	92.9%	7	0.65	0.67	0.58	0.86
0065-503	2011	31	-	23.7	8.7	1.2	0.3	Lepto	11	29	16	270	98.5%		0.60	0.62	0.49	0.92
0067-500	2011	33	-	28.6	6.0	1.8	0.3	Lepto	19	33	16	272	100.0%	10	0.52	0.57	0.46	0.98
1886-202	2009-10	40	-	34.8	13.2	0.7	0.4	Meso	22	40	22	157	100.0%	5	0.73	0.75	0.69	0.98
1892-009	2011	25	-	21.0	10.6	3.4	0.1	Lepto	0	25	20	1,799	75.7%	7	0.75	0.76	0.72	0.74
1892-010	2010	40	-	31.9	37.0	3.4	0.3	Lepto	3	40	28	365	81.9%	18	0.87	0.88	0.85	0.89
1892-012	2011	40	-	35.9	22.4	12.9	0.2	Lepto	4	40	28	595	95.3%	26	0.86	0.87	0.86	0.96
1892-013	2011	30	-	20.9	19.4	0.6	0.2	Lepto	4	28	17	795	83.9%	2	0.79	0.80	0.72	0.82
3681-030	2011	40	-	35.2	10.3	0.9	0.3	Lepto	22	40	28	222	98.6%		0.65	0.68	0.62	0.94
3792-300	2010	40	-	31.7	35.5	1.5	0.1	Lepto	7	40	22	1,360	92.4%	30	0.86	0.86	0.84	0.95
7013-013	2009-10	40	-	36.3	12.5	2.9	0.4	Lepto	22	40	32	186	89.2%		0.76	0.77	0.75	0.90
7014-014	2011	40	-	37.3	7.2	9.3	0.2	Lepto	19	40	32	670	96.7%	8	0.67	0.68	0.66	0.93
7015-015	2010	40	-	36.2	9.8	4.2	0.4	Lepto	23	40	32	127	93.7%		0.68	0.71	0.67	0.88
7104-603	2009	28	2	22.8	8.5	0.1	0.2	Meso	11	28	19	876	91.1%		0.60	0.62	0.56	0.82
7545-201	2009-10	40	-	34.8	9.7	3.9	0.4	Lepto	21	40	26	120	98.3%		0.65	0.68	0.59	0.95

Table 15: Summary statistics for paper fixed on-demand tests

Online fixed dated tests

Unit number	Exam date/period	Items to answer	Mean score	Score variance	Kurtosis	Kurtosis SE	Kurtosis type	Minimum score	Maximum score	Cut score	Candidates	Candidates passed	Candidates removed	Cronbach's alpha	Guttman's lambda-2	Phi	Phi (lambda)
2251-501	2010	19	13.5	7.7	0.0	0.2	Meso	2	19	13	788	63.3%	2	0.64	0.65	0.57	0.59
3638-975	2011	40	33.1	25.2	5.9	0.1	Lepto	0	40	28	1,252	88.5%	2	0.83	0.84	0.80	0.90
3638-978	2010	40	28.6	27.5	2.8	0.1	Lepto	0	40	28	2,361	61.6%		0.78	0.79	0.74	0.74*
3792-986	2010	40	31.0	24.0	1.5	0.1	Lepto	0	40	27	5,966	83.2%	8	0.78	0.78	0.74	0.83
4475-502	2011	20	15.8	7.3	1.3	0.2	Lepto	2	20	14	677	82.7%	2	0.65	0.66	0.60	0.71

*Phi (lambda) was less than phi so phi taken as final phi (lambda) value as it is the lower bound

Table 16: Summary statistics for online fixed dated tests

Online randomly generated on-demand tests

Unit number	Exam date/period	Items to answer	Items removed	Items in bank	Mean score	Score variance	Kurtosis	Kurtosis SE	Kurtosis type	Minimum score	Maximum score	Cut score	Candidates	Candidates passed	Candidates removed	Cronbach's alpha	Guttman's lambda-2	Phi	Phi (lambda)
0351-200	2010	15	-	62	9.7	5.4	0.0	0.2	Meso	2	14	7	935	90.1%	6	-	-	0.39	0.73
0361-101	2010	16	-	68	13.1	5.8	2.0	0.1	Lepto	0	16	10	2,120	91.2%	31	-	-	0.63	0.86
1892-008	2010	65	-	355	52.6	24.4	4.5	0.2	Lepto	24	63	45	497	94.6%	8	-	-	0.60	0.88
2079-101	2010	40	-	184	28.4	24.9	-0.1	0.1	Meso	9	40	24	6,866	83.3%	119	-	-	0.69	0.82
2382-200	2010	30	1	204	24.3	13.5	0.6	0.1	Lepto	7	30	18	7,655	95.0%	1,431	-	-	0.68	0.92
2800-601	2009	40	-	190	27.4	25.1	0.4	0.2	Lepto	7	38	20	940	92.4%	26	-	-	0.67	0.90
3667-102	2010	25	-	148	20.7	6.3	1.1	0.3	Lepto	11	25	16	373	96.8%	5	-	-	0.45	0.88
4101-114	2009	25	-	99	16.4	9.7	-0.2	0.1	Platy	5	25	15	1,904	73.3%	27	-	-	0.44	0.53
4101-130	2009	25	-	110	17.0	11.3	-0.2	0.2	Meso	6	25	15	938	76.9%	16	-	-	0.54	0.66
4101-702	2010	15	-	151	8.7	6.2	-0.4	0.1	Platy	1	15	9	4,517	52.4%	29	-	-	0.44	0.45
4417-201	2010	30	-	182	24.4	10.4	1.8	0.1	Lepto	7	30	19	1,672	95.0%	27	-	-	0.58	0.89
4835-201	2010	25	-	100	20.2	7.3	0.9	0.1	Lepto	7	25	15	1,250	96.6%	14	-	-	0.49	0.89
4871-106	2009	40	-	148	27.2	30.6	0.1	0.3	Meso	9	39	20	290	92.1%	116	-	-	0.73	0.90
4872-201	2009	50	-	207	39.4	37.8	0.5	0.1	Lepto	14	50	25	1,868	97.8%	170	-	-	0.80	0.97
4873-399	2009	60	-	477	40.5	58.4	-0.7	0.2	Platy	19	57	33	588	84.2%	113	-	-	0.79	0.89
6217-099	2009	30	-	125	23.0	12.2	1.5	0.1	Lepto	3	30	18	1,552	92.5%	11	-	-	0.58	0.86
6314-103	2009	60	-	291	39.0	51.6	0.0	0.1	Meso	13	58	30	1,596	90.2%	62	-	-	0.75	0.90
6926-022	2010	40	-	160	32.1	21.1	1.6	0.1	Lepto	4	40	15	5,965	99.7%	161	-	-	0.72	0.98
7540-664	2010	40	-	242	25.4	23.0	0.0	0.1	Meso	8	40	26	2,504	51.2%	39	-	-	0.61	0.62

Unit number	Exam date/period	Items to answer	Items removed	Items in bank	Mean score	Score variance	Kurtosis	Kurtosis SE	Kurtosis type	Minimum score	Maximum score	Cut score	Candidates	Candidates passed	Candidates removed	Cronbach's alpha	Guttman's lambda-2	Phi	Phi (lambda)
7564-303	2010	80	-	320	46.3	143.3	-0.6	0.3	Meso	23	77	40	247	66.0%	8	-	-	0.87	0.90
7579-311	2010	30	-	100	22.7	18.8	1.0	0.3	Lepto	5	30	18	286	87.4%	8	-	-	0.73	0.87

Table 17: Summary statistics for online randomly generated on-demand tests

We wish to make our publications widely accessible. Please contact us if you have any specific accessibility requirements.

First published by the Office of Qualifications and Examinations Regulation in 2013

This publication is also available on our website at www.ofqual.gov.uk

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

2nd Floor
Glendinning House
6 Murray Street
Belfast BT1 6DN

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