Validation of vocational qualifications

Final Report

Alpha*Plus* Consultancy Ltd

February 2014 **Ofqual/14/5373**

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

Executive summary

This is the final deliverable from the Validation of Vocational Qualifications project, carried out as part of Ofqual's Validity Programme. The project was carried out by Alpha*Plus* Consultancy Ltd in collaboration with City & Guilds and NCFE – two of the leading UK vocational awarding organisations with large portfolios of qualifications under regulation by Ofqual – and explores issues relating to the validity of vocational qualifications. The research questions for this project were:

1. How do awarding organisations ***currently*** justify the validity of their vocational qualification(s) (in terms of being fit for their particular purposes)?
2. What evidence to justify the validity of their vocational qualification(s) could awarding organisations ***reasonably be expected*** to produce ***in the future*** in producing a comprehensive and robust validity argument?

Four case studies have been carried out, with each case study involving validity work around a single vocational qualification. The investigative work concentrated on specific qualifications, but the focus of the research is not on the validity of the qualifications *per se*; rather, it is to investigate validity *processes*, and to show what a validity report might look like for a vocational qualification.

A key frame of reference for the project has been Ofqual's argument-based framework for validating assessments. In essence, the research has involved applying the Ofqual validation framework to investigate the evidence available to support a validity argument for the four case studies (to answer research question 1), and subsequently considering how vocational awarding organisations might reasonably be expected in future to provide evidence against the Ofqual framework (to answer research question 2).

The research found that, although there was often strong evidence for the validity of the purpose of the qualification, there was less readily available evidence around the validity of the assessments themselves. In part this is because the data which might be available (e.g. to consider reliability) for general qualifications often does not exist, or is not collected, for vocational qualifications. Despite this, each of the four case study qualifications (and their assessments) were perceived by stakeholders to have high validity. This itself may be considered a form of evidence: there is a genuine market in place for vocational

qualifications which there is not in general qualifications, and market forces could be said to play a role in ensuring some aspects of validity.

In attempting to answer the second research question, we are recommending that additional evidence which vocational awarding organisations could reasonably be expected to produce in future should have the following features:

- It should include a robust justification for the method of assessment and documentary evidence of stakeholder support for the method of assessment.
- If a qualification is based on national standards, the evidence provided by the awarding organisation should include the relevant national standards and a mapping from the national standards to the qualification at the unit level.
- For externally set and marked tests, awarding organisations should:
    - collect marking data at the item level
    - calculate estimates of reliability using suitable indices
    - where pre-determined cut scores are used, provide detailed documentation of the process followed and the rationale for the choice of cut-score(s).
- For internally marked tests and assignments etc., awarding organisations should be required to provide:
    - sample assessments and mark schemes created by centres
    - evidence documenting how grade boundaries are moderated, which should include copies of visit/moderation/verification reports where available (i.e. where at least one visit has occurred in the lifetime of the qualification)
- Awarding organisations should be encouraged to strengthen their systems over time to collect data about their candidates in order to support analyses of the performance of sub-groups of candidates. The practicalities of collecting such data need to be properly investigated, however – it may be that a sector-wide initiative is required to agree a minimum dataset to be provided by centres during candidate registration. The practicalities of centres providing the required data, however, also need to be considered. Where there is sufficient data to support an analysis of construct irrelevance, vocational awarding organisations should be expected to carry out such analyses**. Ofqual need to understand and recognise the limitations of this analysis for many vocational qualifications.**
- Standardising how and when data gets recorded in centres is a necessary prerequisite to collecting robust data to support calculation of formal estimates of

reliability for internal assessments. Further research is required to explore how awarding organisations could strengthen their systems in this area.

We believe there is an argument for two levels of validity audit: a **standard validity audit** and a **high-impact qualification validity audit**. A standard validity audit would draw upon the evidence of validity that Ofqual might reasonably expect vocational awarding organisations to generate in the normal course of their operations (including those listed above). A high-impact qualification validity audit would be carried out less frequently than a standard audit and would seek to generate and assess one or more form of additional validity evidence which would not normally be available to vocational awarding organisations in the normal course of their operations without the addition of significant costs.

One further reason to consider the implementation of validity audits in the way described relates to the levels of expertise required to fully implement and assess the Ofqual validity framework. Ofqual has produced a sophisticated and complex validation framework which will need some expertise to put into effect. It is unlikely that the vocational sector as a whole has sufficient expertise to effectively implement all aspects of the framework meaningfully. Equally, it is unclear whether Ofqual has the resources required to assess the evidence produced. Having two levels of audit will reduce the overhead required on both sides for ensuring the validity of vocational qualifications.

There are a number of areas where further research may be warranted to test how practical it would be to implement the recommendations made in this report. In each case, a small pilot involving awarding organisations and centres would help to clarify the required processes, demonstrate how these might be implemented, and identify the implications – and the barriers, too – for awarding organisations and centres.

Table of contents

1 Introduction

This is the final deliverable from the Validation of Regulated Assessments project (Ofqual contract reference OF156). The project was carried out by Alpha*Plus* Consultancy Ltd, in collaboration with City & Guilds and NCFE, two of the leading UK vocational awarding organisations with large portfolios of qualifications under regulation by Ofqual. The project explores issues relating to the validity of vocational qualifications.

1.1 Validity and validation of assessments

Assessment validity requires a qualitative judgement of the degree to which inferences based on assessment results are meaningful, useful and appropriate. Researchers have identified many different aspects of validity, including:

- **Construct validity –** the extent to which an assessment measures what it is intended to measure
- **Content validity –** the extent to which assessment content represents the required skills in the specified subject area
- **Predictive validity –** the extent to which the results of an assessment can be used to predict future behaviour or achievement
- **Face validity –** the perception that the assessment is measuring what it should be measuring (i.e. that a mathematics test looks like a mathematics test)

This list is not comprehensive, and other aspects of validity (e.g. convergent and discriminant validity, concurrent validity, consequential validity, curricular validity, systemic validity, etc. (Frederiksen and Collins, 1989; Kane 2001; Johnson, 2007)) have been proposed. These categories are a convenient way to organise and discuss validity evidence, but they are not obviously distinct concepts: evidence normally identified with construct validity, for example, may also be relevant as evidence of content validity.

Recent treatment of validity in the literature (following Messick, 1989) has tended to recognise validity as a unified concept which encompasses different aspects, such as construct validity, content validity, face validity, predictive validity, etc. (Kane, 2001; Kane, 2006; Shaw & Weir, 2007; Opposs & He, 2011). The unified view holds that individual categories of validity are in fact different fundamental aspects of a single overarching concept of validity.

There is relatively little in the literature related to the validity of vocational qualifications. Miller and Linn (2000) outline six aspects of construct validation to guide the validation of performance-based assessments. Each aspect is discussed and studies that could be conducted within the context of a large-scale educational assessment are presented, but no practical implementation is reported. Stasz (2011) focuses on the purposes of vocational qualifications (what they are for, what purposes and functions they are used for) and asks whether vocational qualifications are fit for those purposes. The paper examines conceptions of validity and their implications for the interpretation of assessments. Stasz notes that judging the validity of vocational qualifications based on performance assessments is 'complicated', and ends with an unanswered question: 'To what extent do vocational qualifications support valid inferences in relation to their purposes?'

Validating an assessment refers to the process of accumulating empirical data and logical arguments to show that inferences based on assessment results are appropriate. A variety of frameworks have been proposed for validating assessments for their intended uses.

- Messick (1989) described six aspects of construct validity that should be addressed in any validation exercise. Messick's framework has been criticised (Brennan, 1998; Kane, 2006) in terms of practicality.
- Frederiksen and Collins (1989) presented a framework which focused on the concept of 'systemically valid tests as ones that induce curricular and instructional changes in education systems … that foster the development of the cognitive traits that the tests are designed to measure'. The measures they proposed included 'directness of measurement' (related to authenticity), scope (related to content validity), reliability and 'transparency' (the degree to which the criteria of the assessment were clear to candidates).
- Linn, Baker and Dunbar (1991) proposed a set of criteria to use to evaluate the validity of new assessments, focusing on two major categories: properties of the assessment itself (internal validity criteria, including factors such as cognitive complexity and meaningfulness) and factors external to the assessment (external validity criteria, including factors such as fairness and consequential validity).
- Crooks, Kane and Cohen (1996) represent validity as a chain of linked stages where a single weak link ('threat to validity') weakens the whole chain. The stages they identify are administration, scoring, aggregation, generalisation, extrapolation,

evaluation of performance, decisions made on the basis of judgements, and the impact of assessment processes, interpretations and decisions.

- Weir (2005) and Shaw and Weir (2007) proposed an assessment validation framework based around a unified model of validity which incorporates discrete elements defining the various types of validity evidence which can be collected at each stage in the test development process.
- Kane (2006) proposed an argument-based approach to validity evaluation which requires researchers to explicitly state the proposed interpretations of the assessment. According to Kane, an *interpretive argument* has a set of inferences and assumptions that are used to support the proposed interpretation of test results for the intended use, while the *validity argument* provides an evaluation of the interpretive argument. In simple terms, Kane's approach to evaluation is to search for and evaluate all the threats to the validity of the assessment inferences.

As part of the Validity Programme, Ofqual has adopted an argument-based approach to validity (Opposs and He, 2011). The framework document states:

Given the purpose (intended use) of the assessment and the proposed interpretation of the results, an argument-based approach to validation generally involves:

- *the development of a set of clear and coherent propositions that support the proposed interpretation and intended use of the results (interpretive argument)*
- *the evaluation of the plausibility of the propositions using appropriate data collected from various stages of the assessment process and logic (i.e. to develop a validity argument based on evidence to support the proposed interpretation and intended use of the results)*

Ofqual is supporting research that validates assessments regulated by Ofqual; this project is part of that research. A key frame of reference for the project has been Ofqual's argument-based framework for validating assessments.

1.2 Aims of the research

Ofqual is conducting a research programme, the Validity Programme, to study the validity of regulated assessments in England. The primary aims of the Validity Programme are to:

- gain a better understanding of the major issues associated with the validity of regulated assessments in England
- understand the extent of validity for a selection of regulated assessments
- develop effective validity auditing procedures for Ofqual-regulated assessments

This project, part of the Validity Programme, was carried out by Alpha*Plus* Consultancy Ltd, in collaboration with City & Guilds and NCFE. The project explores issues relating to the validity of vocational qualifications. To this end, four case studies have been carried out, with each case study involving validity work around a single vocational qualification. The investigative work concentrated on specific qualifications, but the focus of the research is not on the validity of the qualifications *per se*; rather, it is to investigate validity *processes*, and to show what a validity report might look like for a vocational qualification.

The research questions for this project were:

1. How do awarding organisations ***currently*** justify the validity of their vocational qualification(s) (in terms of being fit for their particular purposes)?
2. What evidence to justify the validity of their vocational qualification(s) could awarding organisations ***reasonably be expected*** to produce ***in the future*** in producing a comprehensive and robust validity argument?

A key frame of reference for the project has been Ofqual's argument-based framework for validating assessments (Opposs and He, 2011). In essence, the research has involved applying the Ofqual validation framework to investigate the evidence available to support a validity argument for the four case studies (to answer research question 1), and subsequently considering how vocational awarding organisations might reasonably be expected in future to provide evidence against the Ofqual framework (to answer research question 2).

1.3 Methodology

Four qualifications from the partner awarding organisations were selected as case studies in this research. The qualifications which formed the basis of the case studies were:

- City & Guilds:
  - Level 3 Diploma in Pharmaceutical Science (QAN: 500/9959/0)
  - Level 2 Diploma in Professional Cookery (QAN: 500/8909/2)
- NCFE:
  - Level 1 Functional Skills English (QAN: 501/1660/5)
  - Level 3 Certificate in Principles of Customer Service (QAN: 600/2922/5)

Each of the four case-study qualifications provided an opportunity to discover the nature and extent of evidence currently available to support the validity argument for that

qualification, and enabled reviewers to develop validity reports against Ofqual's validity framework. Four qualifications can never be fully representative of all vocational qualifications but, together, these four qualifications do provide a range of qualification and assessment types.

Once the case study qualifications had been selected, there were three distinct phases to the project:

- evidence gathering
- production of individual validity reports for the four case study qualifications
- a synthesis and analysis stage encompassing lessons learnt from the four case studies

The three phases are outlined below.

Evidence gathering

A reporting template based on Ofqual's argument-based framework for validating assessments was created in the initial stages of the project (referred to in this report as the **Ofqual validity framework template**: see Appendix 1), and the evidence gathering phase was guided by the requirements of the template – researchers sought evidence to enable them to build a validity argument for each case study qualification against the Ofqual validity framework template.

The evidence-gathering phase for each of the four case study qualifications involved:

- A review of all the documentary evidence supplied by the partner awarding organisations for the four case study qualifications. The documentation was considered in relation to the questions in the Ofqual validity framework, and consideration given to how the available documentation might provide evidence to support a validity argument for each qualification. This was a somewhat iterative process, in that, where there appeared to be gaps in the documentation or where additional evidence was thought to exist (or should exist), reviewers would ask the partner awarding organisations for more information. By way of example, Appendix 2 lists the documentary evidence reviewed for the case study qualification Level 3 Diploma in Pharmaceutical Science.

- After the review of the documentation, researchers carried out on-site interviews[1] with C&G and NCFE staff to understand the awarding organisation processes more fully and to identify where further validity evidence may or may not exist for the case study qualifications. Again, the questions in the Ofqual validity framework provided the context of the investigation.

Case studies

Once the documentary review and the interviews had been completed, reviewers working in pairs compiled and assessed the validity argument for each qualification. As noted above, the validity arguments were compiled using the Ofqual template shown in Appendix 1, so that a fully populated template was generated for each case study qualification. Summary reports for the case studies are included in Appendix 3.

Synthesis and analysis

The results of the four case studies, in terms of the available evidence to support a validity argument (and in particular any common gaps or weaknesses in the available evidence across the case studies), were considered in order to reach an aggregate view of the issues around applying the Ofqual validity framework to vocational qualifications. The main aim of this synthesis and analysis work was to consider research question 2: 'What evidence to justify the validity of their vocational qualification(s) could awarding organisations ***reasonably be expected*** to produce ***in the future*** in producing a comprehensive and robust validity argument?'

[1] In the case of the Level 2 Diploma in Professional Cookery qualification, a researcher also attended a network meeting of assessors.

2 Assessment in vocational qualifications

The collection of information and data by awarding organisations to support a validity argument for their qualifications is the main area of research for this project, with particular regard to the validity of the assessments. This section will review the assessment of vocational qualifications and how relevant aspects might affect a validity argument. For context, it is worthwhile beginning with the lifecycle of qualification development.

2.1 The qualification development lifecycle

Figure 1 presents a simplified view of the lifecycle of a vocational qualification. It is broadly similar to the product development lifecycle in many other sectors. The concept of a new qualification will usually come from a clearly identified pull from the market. This may be the result of a structural change in the market such as the introduction of new or revised occupational standards, or it may be a need identified as a result of market research or customer feedback. Once the need has been identified, a business case is initiated and developed, identifying the opportunity, the required investment, the expected return on investment, etc. If this business case is approved, then the next stage is qualification design.

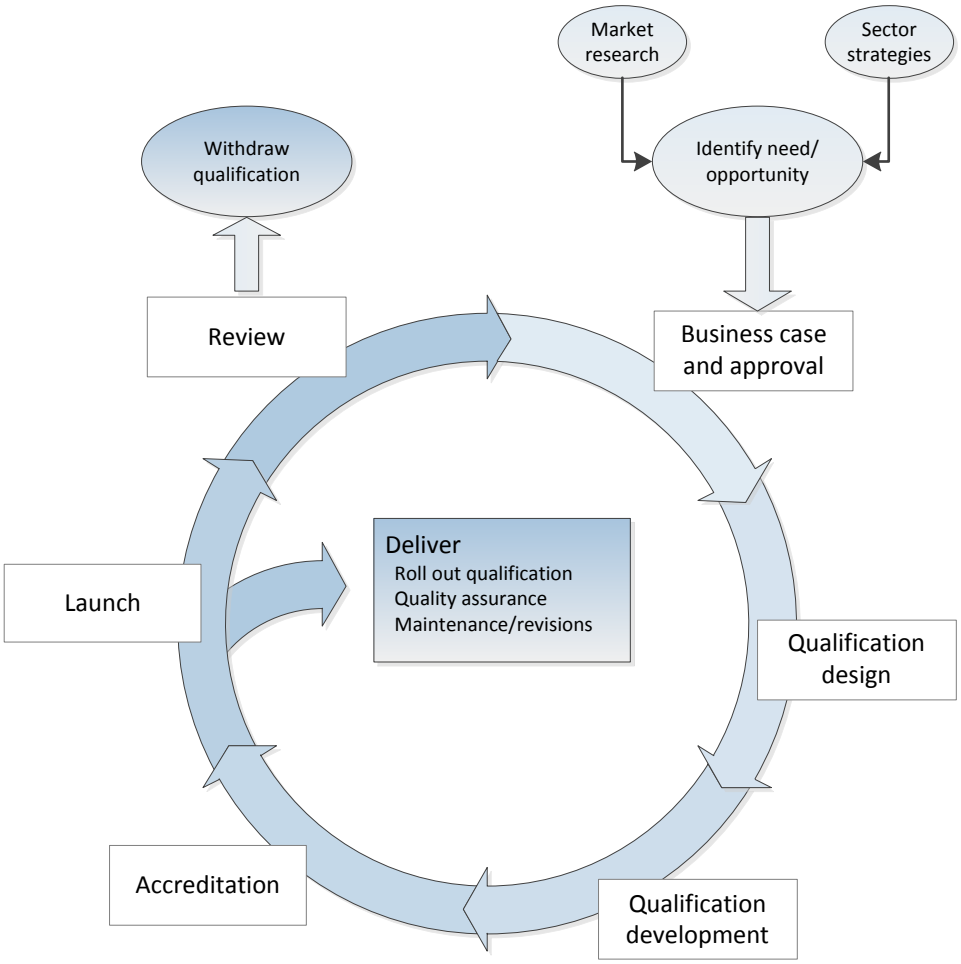

Figure 1. Simplified qualification development lifecycle

Depending on the qualification and the particular processes of the awarding organisation, qualification design, development, and accreditation may be integrated to a greater or lesser degree (this is particularly true for smaller awarding organisations). Tasks may include creating a test specification, identifying suitable units from the qualifications and credit framework (QCF), writing and submitting new QCF units, establishing the rules of combination, producing question papers, mark schemes, assignments and supporting documentation. Once developed and accredited, the new qualification will be launched and will enter the mainstream portfolio of products for the awarding organisation. Awarding organisations will continually review qualifications and portfolios of qualifications against the changing needs of the market.

This represents an extremely simplified and compressed explanation of the product development lifecycle, but it highlights an interesting issue in relation to the evidence available for establishing validity. Much of the evidence which might contribute towards a

validity argument is (or could be) assembled in the early stages of the lifecycle – identifying the need, and designing and developing the qualification.

2.2 Purpose, constructs and interpretation of vocational assessment results

The purposes of vocational qualifications are mutable and vary for different stakeholders, but the definition given in recent government legislation (referenced in Stasz, 2011) is relevant to this study:

> *The key purpose of qualifications is to show clearly and publicly the knowledge, skills and attributes that an individual has gained, especially to inform prospective employers and future providers of education and training.*

Ofqual developed and trialled its own classification of qualification purposes (Ofqual, 2009) which subsequently provided the basis of the qualification purposes on the QCF. The purposes are intended to focus on the primary outcome related to the achievement of a qualification and to span the full breadth of provision likely to be accredited. These purposes are shown in Table 1. In practice, qualifications will be designed and used for multiple reasons, not all of them always intended; an example is how ESOL qualifications are now regularly used to support a student's residency or citizenship application, rather than purely to attest to their language skills.

The purpose of a qualification is a key parameter in any consideration of validity, but, as the regulator, Ofqual's focus is firmly on evidence that the qualification *assessment* should produce a valid measure of the proficiency of an individual. Vocational qualifications, and their assessment, are often markedly different from general qualifications – and this is by design. Young (2008) in his description of change within vocational education and training defines the curriculum in terms of knowledge-based, standards-based and connective approaches. From the late 19th century to the late 1970s a knowledge-based FE curriculum was in place for the craft and technician occupations in the industrial sector, focusing on traditional academic 'subject' knowledge not applied in any workplace context, such as physics or mathematics.

Table 1. Purposes for qualifications in the QCF

Main purpose	Sub-purpose	Examples of possible qualifications
A. Recognise personal growth and engagement in learning	A1. Recognise development of Skills for Life	Functional skills qualification Skills for Life
	A2. Recognise development of knowledge and/or skills to operate independently and effectively in life, learning and work	Qualification in personal and social development Qualification in personal effectiveness
	A3. Recognise development of personal skills and/or knowledge	Qualification in music
	A4. Recognise development of employability skills and/or knowledge	Qualification in employability
B. Prepare for further learning or training and/or develop knowledge and/or skills in a subject area	B1. Prepare for further learning or training	GCSE Access to HE, or Pre-U qualification
	B2. Develop knowledge and/or skills in a subject area	Qualification in art and design Qualification in classical languages
C. Prepare for employment	C1. Prepare for employment in a broad occupational area	Qualification in economics or business Qualification in performing arts or media
	C2. Prepare for employment in a specific occupational area	Qualification for teaching assistants Qualification in hairdressing
D. Confirm occupational competence and/or 'licence to practice'	D1. Confirm competence in an occupational role to the standards required	NVQ Other competence-based qualification
	D2. Confirm the ability to meet a 'licence to practice' or other legal requirements made by the relevant sector, professional or industry body	Qualification in door supervision Qualification in accountancy
E. Updating and continuing professional development (CPD)	E1. Update knowledge and/or skills relating to legal, technical process or best practice changes/requirements	Qualification in infection control Qualification in health and safety
	E2. Develop knowledge and/or skills in order to gain recognition at a higher level or in a different role	Qualification in operational or strategic management
	E3. Develop knowledge and/or skills relevant to a particular specialisation within an occupation or set of occupations	Specialist qualification in the health sector Specialist qualification in food safety

As the number employed in science-based industries declined, there was a move in the late 1970s/early 1980s towards qualifying employees in lower-skilled roles through evidencing 'competence'. The standards-based curriculum and its assessment were clearly linked to evidencing competence of a skill in its workplace setting, with assumptions that the underpinning trade knowledge was implicitly understood, through its application. This approach was realised in the form of standards-based qualifications, standards being minimum standards required to do the job. National vocational qualifications (NVQs) were

developed as a result of this shift towards a non-academic vocational curriculum (Young, 2008). The degree of vocational relevance remains an on-going area of contention, remaining pivotal in policy development (DfES, 2006; Foster, 2005; Wold, 1998).[2]

Many vocational qualifications are assessed in the workplace and assessed by people in that context. Wolf (2011) identified three key aspects of competence-based assessment:

- The emphasis on outcomes – specifically, multiple outcomes, each distinctive and separately considered
- The belief that these outcomes can and should be specified to the point where they are clear and 'transparent' – assessors, candidates, and stakeholders should be able to understand what is being assessed and what should be achieved
- The decoupling of assessment from particular institutions or learning programmes

The first point listed above implies that candidates must demonstrate 'mastery' – that is they must satisfy *all* the assessment criteria. This is in contrast to grading and compensation arrangements of national qualifications. Mastery is often assessed through observation by assessors, and by the assessment of a candidate's portfolio, which may include evidence provided by supervisors, colleagues or managers, as well as written assignments, practical tasks, oral reports and testimony. In many cases, candidates can repeat assessment tasks until they are deemed to have demonstrated mastery. Competence-based assessments place a particular emphasis on the validity of assessment (most obviously, face and construct validity). It is not clear, however, how this validity is ensured by awarding organisations, or how it might be audited.

The second bullet point above relates to the transparency of standards, criteria and procedure – which is designed to facilitate the fairness of the assessment. The content and criteria of vocational qualifications are often derived from functional analysis of workplace tasks or occupations. Indeed, the content and criteria of vocational qualifications may be developed in partnership with employers, often as part of the development of National Occupational Standards. Thus key parameters and processes influencing the validity of

[2] The Richard Review of apprenticeships was published only shortly before the completion of this report: http://www.schoolforstartups.co.uk/richard-review/richard-review-full.pdf

vocational qualifications[3] may be determined not by awarding organisations, but by Sector Skills Councils (SSCs), Standards Setting Bodies and, ultimately, employers. The interface between awarding organisations and these other bodies is an important one when considering the validity of vocational qualifications, and of how this validity might be audited. Ofqual does not regulate the SSCs, but the awarding organisations rely heavily on the outputs of the SSCs in ensuring the validity of many qualifications. Ofqual is clear that it is the assessment provider's responsibility to demonstrate that the results from their assessments are valid for the purposes set for the assessments. The next section therefore considers assessment in the vocational context.

2.3 Key features of assessment in vocational qualifications

Both general and vocational qualifications need to be demonstrably valid, and the Ofqual validity framework outlines the various aspects of validity which regulated qualifications must demonstrate. Most of the literature on validity is, however, framed in the context of conventional tests and test results. This is the prevalent model of assessment in general qualifications, but not in vocational qualifications. In considering how a validity argument for vocational qualifications could be developed and interpreted, it is therefore important to understand some of the differences between external assessment in general qualifications and in vocational qualifications. In this section we therefore compare key features of assessments for general qualifications such as GCSEs and A levels with those for a typical vocational or professional qualification. There is inevitably an element of generalisation here but the fundamental principles apply in the vast majority of cases, and can inform any consideration of the validity of vocational qualifications.

2.3.1 Curriculum and coverage

The curriculum and coverage is the range of knowledge, skills and understanding that the candidate is expected to study and then to be assessed on.

[3] For example, with regard to the Level 3 Diploma in Pharmaceutical Science case study qualification, the General Pharmaceutical Council (GPhC) is directly responsible for the development of the learning outcomes for the qualification and for providing a technical sign-off point for the assessment content. They were not, however, involved in the development of the assessment criteria, grading criteria or mark schemes.

2.3.1.1 Curriculum and coverage in general qualifications assessment

Most general qualifications are graded, so the range of awards offered cover a wide range of abilities. Tests will contain easier or more common topics or questions that almost all candidates are expected to attempt, and then more difficult topics or questions which the examiner knows only some will attempt. Optional questions ('choose to answer one of the following six') allow the candidate to show more detailed knowledge on a topic of their choice. In many general qualifications the range of coverage is so large that it's not practical to assess it all, so each year the assessment will sample the curriculum. Over a period of years the whole assessable part of the curriculum is assessed, but not always in each year.

2.3.1.2 Curriculum and coverage in vocational qualifications assessment

The critical feature of a typical competence-based vocational qualification assessment is that it is concerned with assessing competence – is the candidate competent or not? – not in providing a range of grades or other scaled judgements. As far as possible, the curriculum covers *all* the knowledge, skills and understanding that is needed. Similarly, the assessment comprises tasks that assess *all* (or almost all – practical reasons of test duration sometimes dictate an element of the sampling seen in general qualifications) essential skills and associated understanding and does not include tasks that are optional.

Passengers will want to be confident that the assessments used to qualify a particular pilot included assessment of both taking off and landing.

2.3.2 The difficulty of the questions

2.3.2.1 The difficulty of the questions in general qualifications assessment

In order to distinguish candidates' abilities across a range of grades which cover a wide range of ability, questions (and question parts) of varying difficulty must be set. It's not acceptable in terms of ethical or statistical reliability for weaker candidates to be unable to answer most of the questions. Neither is it acceptable, however, for the assessment to include mostly relatively easy questions, and the top grades to be given to those whose work has fewest mistakes. The top grades generally reward both careful working and success in answering harder questions.

2.3.2.2 The difficulty of the assessment tasks in vocational qualifications assessment

In vocational qualifications the assessment includes tasks that are designed solely to distinguish between those who are and are not competent. There are no graded questions or tasks (e.g. hard ones to stretch the most able, easier questions/tasks to give weaker candidates a chance to show what they can do). Assessors are generally not interested in excellence – everyone has to reach the required minimum standard, and that's as far as the assessment goes. Similarly, assessors are not interested in ensuring that the assessment provide a leg-up for weaker candidates. Those vocational qualifications in use in England today (e.g. BTECs) that do include grading do so primarily to provide a degree of comparability with general qualifications rather than because it represents good practice in assessment.

2.3.3 Compensation

Compensation is inherent in all exam-based assessments where the candidate's scores from each question are totalled up and compared against a pass mark (or a range of pass marks for grades). Higher marks scored in one area compensate, in terms of the overall score achieved, for weaker marks in other areas of the test.

2.3.3.1 Compensation in general qualifications assessment

Most general qualifications allow and to an extent encourage compensation in the assessment. A candidate scoring full marks on the first half of the exam and no marks on the second half will generally achieve the same outcome as a candidate who scored half the marks available on each question. Because of the potential inequities here, general qualification assessments have complex rules for setting questions to ensure that compensatory features do not allow perverse outcomes, where candidates achieve grades while having serious deficiencies in their knowledge, skills and/or understanding.

2.3.3.2 Compensation in vocational qualifications assessment

For a vocational qualification, being particularly strong in one skill area should not compensate for weakness in another. Compensation is therefore generally not a feature of vocational qualifications, except where it is an unavoidable, implicit feature of the marking for externally assessed qualifications. In these cases, this means that different sections of

vocational qualification tests have their own pass marks (rather than one pass mark for the assessment as a whole).

Pilots who are particularly excellent at taking off, but weak at landing should not be deemed competent – their high score on the taking-off assessment elements should not compensate for the weaker score on other essential skills.

2.3.4 The cohort

2.3.4.1 The cohort in general qualifications

With most general qualifications students are required to take the assessment/examinations at particular points in the educational calendar. As a result the assessments measure what the candidate can do at that time – there is no, or only limited, opportunity for the candidate to decide when they're ready. Modular examinations and re-sitting undermine this principle to an extent but, overall, general qualification examinations tend to measure progress at a particular time, allowing comparisons to be made across the whole cohort.

Because examinations are written to provide a fair assessment for the entire cohort (based on a prediction of the range of abilities that candidates will present with, usually drawn from previous years' cohorts), a wide range of scores on the test is observed (see Figure 2 for an idealised illustration[4]). This range is regarded as positive, because it supports fairer assignment of grade boundary pass marks.

[4] This is a very simplified illustration; real data is never so well behaved. The detail of the score distribution will depend on the subject and size and nature of the cohort and may be influenced by other factors (tiering, for example).

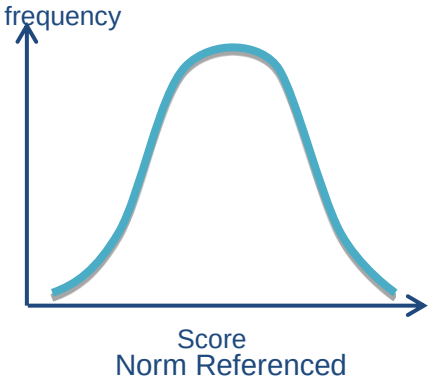

Figure 2. A wide range of scores is usually observed in general qualification examinations

2.3.4.2 The cohort in vocational qualifications

In principle, most vocational and professional assessments should be offered 'when ready', that is, when the candidate, with their tutor, decides when they have reached a level of competence to be able to undertake assessment. Some programmes that use external assessment methods do offer fixed testing opportunities (e.g. four times a year), but generally for operational, rather than educational, reasons. With candidates tending to enter only when they believe they are ready, and questions designed to measure only at or around the point of competence, the range of scores achieved tends to be skewed toward the higher scores, with relatively few candidates scoring very poorly.

There are two important implications to note:

(1) The number of candidates passing is not usually controlled – all may pass, or some, or none, depending on who presents for the assessment.

(2) With 'when ready' (or multiple instances when assessments are available) comes the requirement for multiple test papers, along with the challenge of ensuring that the pass mark each is given provides fair and repeatable judgements of competence from one instance to the next.

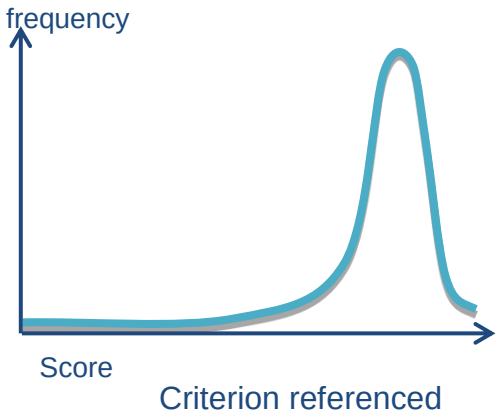

Figure 3.Idealised illustration of the range of scores in vocational qualifications, which tend to be much narrower than for general qualifications, and be skewed towards higher scores

2.3.5 Setting pass marks

Setting pass marks is more complex than may be appreciated by the general public, for two main reasons:

a) Standards in assessment (i.e. what a pass grade or an A grade means) are generally expected to have consistency from year to year for fundamental reasons of fairness and to support comparability of standards. For example, for A level candidates applying to university – some of whom will have taken a gap year, and others of whom will have come straight from Year 13 – it is important that an A grade in the two years' papers that were taken has very similar meaning in terms of describing the candidates' abilities.

b) Assessments from one year or instance to the next vary in difficulty – the same candidate would achieve different scores from one year's paper to the next. Examiners try to keep scores fairly consistent, but this is only possible within a range. So the pass marks cannot usually be set automatically – examiners generally need to look at both the questions and the candidates' performance on them in order to set the pass mark (or pass marks, in those instances where the assessment has multiple grade outcomes).

Despite (a), public pressure in examination systems also often places pressure to ensure that roughly similar numbers of candidates achieve each grade from one year to the next.

2.3.5.1 Setting pass marks for vocational qualifications

As discussed above, a knowledge-based test in a vocational qualification test aims to:

- test the entire essential curriculum

- test only at or around the point of competence (i.e. the boundary between minimally competent and only just incompetent)
- ensure that candidates can complete (almost) all essential tasks to a satisfactory standard

The pass mark for external assessments is therefore usually set quite high (typically 70% or above), reflecting the fact that candidates are expected to be able to do all the required essential skills in order to be competent (arguably, the pass mark could be set at 100%, but some allowance for human error both in the candidates' responses and in the assessment setters' tasks is usually desirable).

For longer assessments covering a wide range of skills, and with a lot of marks available, examiners sometimes feel that even with a high pass mark an unacceptable level of compensation is possible. In these circumstances, test writers may choose to break the test down into sections (usually around a particular topic) and require a pass on each section. There are other approaches to dealing with the impact of undesirable compensation: for example, 'killer questions' – which, no matter what score is achieved, the candidate fails if they answer them incorrectly.

To look at how pass marks are set in practice for vocational qualification assessments, we will consider a common scenario (many other scenarios exist):

- The assessment is an examination which produces two outcomes only: pass or fail
- The assessment is offered on demand to candidates throughout the year. As a result, the awarding organisation maintains a bank of 10 tests (these are called 'parallel forms') and ensures that (a) they are kept secure and (b) no candidate is given the same test twice
- The assessments are written in advance and a defensible pass mark is set at the outset. Reviewing pass marks after the first few hundred candidates may be required, but those candidates will have been awarded a pass or a fail already so the system must be defensible from the start, even if adjustments are made
- The pass mark set needs to be comparable across the parallel forms, so that in principle candidates should pass or fail all of the possible tests consistently if they were to attempt them all

- The assessment is designed to measure the required knowledge and understanding to ensure competence, and so the pass mark will be set at or above 70%
- Any unintended element of compensation in scoring is removed if possible

Arguably, the most secure method of setting pass marks is to require a representative group of candidates to sit the assessments as a trial and then to use this trial information to set a pass mark. In this way, before items and tests are used for live assessments (i.e. where certification is provided), good information about their performance is obtained, which allows pass marks to be set with confidence.

This approach is often not possible, however, for one or more of three main constraints:

- It is expensive to organise effective trials – recruiting candidates who are at or around the level of competence and who are not likely at some point to want to sit the test for real is difficult
- It may be difficult to maintain the security of the items
- There may not be time to undertake trialling before the assessments are needed for live use

Fortunately, alternative methods exist which do not require pre-trialling. In all these cases, however, monitoring of candidate outcomes and the performance of the assessments is much more important than if trialling had occurred. It should be noted that the various methods are not mutually exclusive; where one method is selected, other methods may be used to provide crosschecking.

Method 1: Modified Angoff

This method provides a mechanism for experts (often the question writers coupled with external experts such as teachers) to set a pass mark for an assessment without pre-trialling. It is essentially a 'wisdom of crowds' approach. The approach is as follows:

- The group of (typically, between three and five) experts independently rate each question in each of the tests for which pass marks are to be set. The rating they provide is the probability, or likelihood, that an acceptably (minimally) competent person will answer the question correctly. A minimally competent person is someone who just about adequately performs all job functions safely and requires no further training to do so.

- The experts then review each question together. A consensus is reached for the rating of each test question, commonly by taking an average of the ratings provided by each reviewer but discussing the ratings where there is a large variance or unexpected outliers. (If test information from triallists is available, this can be considered here, too, in terms of both the items' performance – usually facility and discrimination – and also comments from test takers).
- The mean, or average, of all the test question ratings (i.e. the average score that a minimally competent person would achieve) is set as the mark.

Angoff is pretty defensible as an *a priori* standard-setting approach, partly because it is widely used in North America, Australia, and in the UK (Curcin *et al*, 2009) but also because it essentially supports competence judgement processes, which are often the best available anyway. Competency judgement is rarely an exact science and usually involves an expert or group of experts deciding. The strengths and weaknesses of Angoff and modified Angoff methods are well reviewed in Ricker (2006).

Method 2: bookmark system

The bookmark approach relies on having scores from students (either obtained from a pilot or from live assessments). In many cases, a pass/fail judgement can be withheld for the first few students taking an assessment, so that standard-setting activities can be undertaken. The bookmark approach copes well with tests that include constructed response multi-mark questions (alongside multiple-choice questions). The following bookmark approach requires at least 200 student responses to the test (or all the items if a banked and randomised approach is being used to test construction):

(1) Using either Item Response Theory or Classical Test Theory,[5] the item difficulty is assessed and a 'book of items' is created with all the items in it ranging from the easiest to the hardest. Constructed response and multi-mark items appear multiple times in the book for each of the marks available – an item with a maximum score of 3 may be the 12th easiest (i.e. on the 12th page for a score of 1) but 23rd easiest (on page 23 for a score of 2) and 60th (for a score of 3). The book also contains statistical performance data.

[5] The discussion of the merits of these approaches is beyond the scope of this report.

(2) Experts work through the book and decide individually (by putting in a bookmark) the point at which the test passes beyond the level expected of a minimally competent person.

(3) The experts then discuss their findings by comparing bookmarks, and decide the cut-score accordingly.

By presenting items in rank order of difficulty, this method facilitates discussion about what items measure and what makes items more or less difficult, which is helpful for step 3.With data from candidates, this method is robust. A good analysis of its strengths and weaknesses (as well as some details of suitable implementation) is presented in Lin (2006).

Method 3: contrasting groups

The contrasting groups method is an examinee-based approach to standard setting (Cizek & Bunch, 2007). This method in particular requires that the panel of judges be highly familiar with the target test population (and tolerate a consensus or less clear view of 'minimum competence'). A group of expert judges identifies a set of candidates who are clearly not competent (e.g. at the start of their course) and another set of examinees who are clearly competent (recent graduates, practitioners, etc.). The experts try to exclude any borderline candidates. It is important that the non-competent group is carefully selected – they should be members of the target test population, not just people 'in off the street'. Both groups take the test(s), and the resulting test score distributions are plotted on the same graph. The pass mark is set at the intersection of the two distributions (see Figure 4).

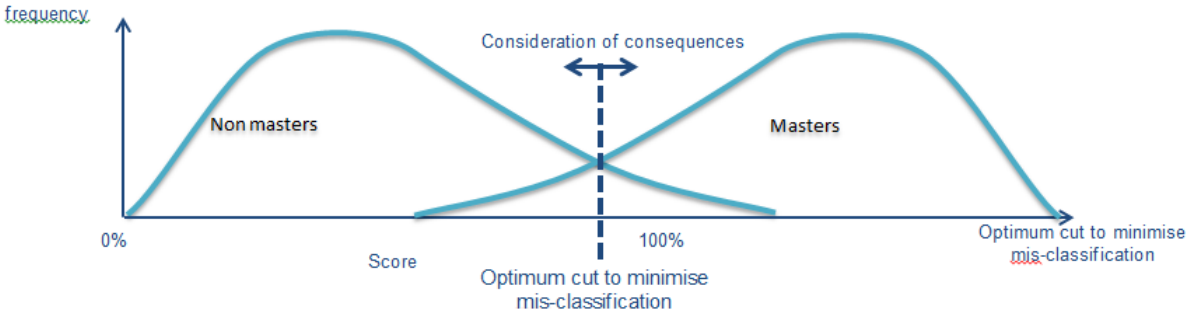

Figure 4. Setting the pass mark using contracting groups

2.3.5.2 Discussion and issues

The Angoff method is popular because it requires no student data, but it suffers in terms of reliability. The other methods are suitable if test data can be made available at the start of the programme.

A major concern for many competency assessment programmes is the risk of accrediting a candidate who is not competent, the type represented by the red shaded area in Table 2.

Table 2. The risk of accrediting a candidate who is not competent

	Fail	**Pass**
Competent	Error of rejection	Correct decision
Not competent	Correct decision	Error of acceptance

Where the risk of such misclassification is high (in terms of either the likelihood of its happening or its impact[6]), pass marks are often adjusted upward by a few marks from the point suggested by the standard-setting procedure. If such adjustments are to be made and candidate test data is available, then statistical methods exist (albeit complicated ones) to ensure that the uplift for each parallel form is equivalent.

The other main issue that standards setters need to consider when setting pass marks for a programme is how often the settings will be reviewed, to take account of changes in expectations of standards and test drift.

2.3.6 Summary

This section has outlined some of the differences between external assessment in general qualifications and in vocational qualifications, including differences in curriculum and coverage, the difficulty of the questions, compensation, the cohort, and how pass marks are set. Understanding these differences is essential to understanding the kinds of evidence which vocational awarding organisations can be expected to provide in support of a validity argument for their qualifications. This is the topic of the next section.

[6] Returning to the analogy of the pilot, the potential impact of misclassifying an inept pilot as competent is considerable, so it may be considered reasonable that the test for an airline pilot has a very high pass mark to ensure that all accredited pilots are demonstrably competent.

2.4 Assessment methods for vocational qualifications

Understanding how the wide variety of assessment methods and data available in vocational qualifications might be used to provide evidence against the various areas of the Ofqual validity framework template presents somewhat of a complex and multi-dimensional puzzle. One of the main defining characteristics of any validity argument will be the method(s) of assessment employed in the qualification. Typically, there is far less emphasis on externally set and marked tests than there is in general qualifications, so the evidence for vocational qualifications may look quite different from the evidence which is typically available for general qualifications. In this section we therefore consider the main methods of assessment which are prevalent in vocational qualifications, and draw upon the case study work to determine the kinds of evidence that vocational awarding organisations might reasonably be expected to provide for each particular method of assessment, and how this could map against the Ofqual validity framework template. Having considered the evidence relating to each form of assessment, in Section 3 we go on to consider what kinds of evidence are actually required by each area of the Ofqual validity framework template. Taken together, this provides the background to consider, in Section 4, how a validity audit might look and the implications for vocational awarding organisations.

Vocational qualifications may use (and combine) a number of different assessment methods to demonstrate sufficient evidence that the learner has met the requirements of the assessment criteria. For the purposes of considering validity evidence, these may be broadly categorised as:

> **Externally set and marked tests,** set and marked by the awarding organisation, delivered either on paper or on computer, a 'traditional' test employing one or more question types, including essay questions, structured/short-answer questions, multiple-choice and other objective questions.
>
> **Internally assessed, externally verified assessments,** set either by the awarding organisation or by the centre, internally assessed/marked by the centre, and externally verified by the awarding organisation – this form of assessment may take a variety of forms, including:
>
> - tests of knowledge

- practical assessments
- oral assessments
- assignments, projects and case studies

In addition to formal assessments, however, vocational qualifications may also accept **other forms of assessment evidence** which demonstrate candidate achievement, including:

- records of assessor observation
- witness testimony (relating to evidence against performance standards)
- records of professional discussions
- candidate and peer reports and other forms of candidate work
- evidence of prior learning

The particular assessment model for a given vocational qualification may draw upon one or more of the assessment methods listed above. In the remainder of this section we consider the evidence which Ofqual might reasonably require for each method of assessment and which areas of the Ofqual validity framework template the evidence could contribute to, and we highlight any specific issues or limitations.

2.4.1 Externally set and marked tests

2.4.1.1 Required evidence

We recommend that the required validity evidence for externally set and marked tests should include:

- assessment specifications
- test papers and mark schemes
- evidence of the procedures used for creating items and constructing tests
- evidence of the procedures followed when administering the assessment, and of how they are monitored
- evidence of the procedures followed by the standard-setting process – **where pre-determined cut scores are used, detailed documentation of the process followed and the rationale for the choice of cut-score(s) should be provided**
- **estimates of reliability of candidate results**

On the basis of the four case studies, there are no strong arguments against awarding organisations providing those items listed above in normal type as opposed to those in **bold type**. Vocational awarding organisations do not currently routinely collect item-level marking data (required to calculate estimates of reliability), but we do not believe there is a defensible argument against this practice for externally set and marked tests. We therefore recommend that awarding organisations should collect marking data at the item level to support calculation of reliability indices.

2.4.1.2 Specific issues

Analyses of construct irrelevance variance

Externally set and marked tests provide some of the data required to perform an analysis of construct irrelevance variance. There are particular issues, however, that make any such analysis problematic for vocational qualifications:

- Typically, the awarding organisations will just not have the required data to carry out any such analysis. For example, they will have little or no access to prior attainment data to enable them to look at how assessments performed for different subgroups of candidates. More fundamentally, there is an argument that many vocational candidates are 'having a second chance' and prior attainment shouldn't therefore be used as a baseline for understanding achievement in their context. Candidates who may have performed poorly at school can find FE more appropriate for their learning style and be successful. This undermines the concept of examining outcomes based on prior attainment.
- The sample of candidates is very often skewed:
    - by gender – overwhelmingly female learners register for hairdressing qualifications, overwhelmingly male for construction qualifications, etc.
    - by age – the age range of candidates is often far more diverse than for general qualifications
    - by ability – the range of abilities may be far greater[7] than for general qualifications

[7] A concrete example of this is an NCFE member of staff who is qualified to MBA level and is currently taking a Level 2 qualification in Health and Fitness.

- Even the concept of a cohort in vocational qualifications is not straightforward. As noted above, many occupational roles are imbalanced. If very few males do hairdressing, then how do we judge the attainment of males in hairdressing? Should they attain at the same level as females, or should we accept/expect a lower level of performance from male hairdressers on grounds of positive discrimination? Or should we believe that males in hairdressing are probably exceptional/highly motivated, and therefore expect their performance to be likely to be higher? What about people who register, but don't complete? They are not explicitly failures, but they are not achieving. It is difficult to be clear about exactly how to define the null hypothesis for a bias study in a gender-imbalanced occupation. Similar problems arise with other demographic data – the combination of biased samples and of lower populations of candidates than in general qualifications means that statistical data needs to be treated with caution.
- In some rare cases some candidates may indeed be unavoidably disadvantaged. For example, some Level 2 qualifications require the learners to refer to industry standard regulations which are written at a Level 3 level of language. Candidates with language skills below Level 3 will be disadvantaged, but there is no reasonable alternative – they need to be able to refer to and understand the regulations.

While recognising the constraints outlined above, however, it is not appropriate to conclude that vocational awarding organisations should ignore this aspect of the validity for all their assessments. We therefore recommend that:

- **Awarding organisations should be encouraged to strengthen their systems over time to collect data about their candidates in order to support analyses of the performance of sub-groups of candidates. The practicalities of collecting such data need, however, to be properly investigated – it may be that a sector-wide initiative is required to agree a minimum dataset to be provided by centres during candidate registration. The practicalities of centres providing the required data, however, also need to be considered.**
- **Where there is sufficient data to support an analysis of construct irrelevance, vocational awarding organisations should be expected to carry out such analyses.**
- **Ofqual need to understand and recognise the limitations of this analysis for many vocational qualifications (in terms of the potential for sample bias and limitations of sample size outlined above).**

Evidence of concurrent validity

As with construct irrelevance variance, the main issue here is that, typically, vocational awarding organisations will not currently have the required data to carry out any such analysis. Furthermore, the nature of vocational qualifications and the diversity of the candidature make it difficult to see how they could obtain relevant data in the future.

2.4.2 Externally set, internally marked tests

In principle, vocational awarding organisations might be expected to provide very similar evidence for externally set/internally marked tests as they will for externally set/externally marked tests. However, there are practical issues which affect the evidence they might be expected to provide in relation to:

- evidence of processes for checking the marking
- evidence of the procedures followed by the standard-setting process
- estimates of reliability of candidate results

We consider the specific issue of each in turn below.

2.4.2.1 Specific issues

Evidence of the procedures followed by the standard-setting process

For internally marked tests there is currently no awarding organisation standard-setting process as it would be understood from the general qualification perspective. If the awarding organisation, however, provides tests with pre-determined grade boundaries (see Section 2.3.5.1), then in effect standard setting is effected at the test development stage, and there should be a requirement to supply evidence of the efficacy of this approach. For internally marked tests where the awarding organisation provides tests with pre-determined grade boundaries, therefore, we recommend the following:

- **where pre-determined cut scores are used, awarding organisations should be required to provide detailed documentation of the process followed and the rationale for the choice of cut-score(s)**

For internally marked tests where the awarding organisation provides tests with only notional grade boundaries or grade descriptors, then we recommend that awarding organisations should be required to provide:

- evidence of the assessment guidelines accompanying the test which are provided to centres

- **evidence of how the grade boundaries are moderated by the awarding organisation**

The evidence listed in bold above was not available to the case studies, but we believe it is reasonable that awarding organisations should provide it as part of the validity argument for this form of assessment.

Estimates of reliability of candidate results

For internally marked tests, awarding organisations could provide estimates of the reliability of results if they were able to collect item-level data from centres; in principle, this is not an unreasonable requirement to put on awarding organisations or on centres. One of the issues raised in the case study work, however, has been about the way that centres record assessment results, and the implications this has on estimates of reliability.

The problem is that different centres tend to record data differently, and even to record different data. For example, while some centres may record only the final results of a candidate against each assessment criterion, other centres will record the outcomes of each 'assessment' (formative or otherwise) against each assessment criterion. Such centres may therefore have multiple different entries against some/all assessment criteria. In terms of reliability, these data may look wildly inconsistent and contradictory: e.g. a candidate has passed assessment criterion 1 and assessment criterion 2 for assessment 1 but no others; the same candidate has **not** passed assessment criterion 1 or assessment criterion 2 for a second assessment but has passed assessment criterion 3. The reality in the centre might, however, be that there was no opportunity to provide evidence against assessment criterion 3 in the first formative task, but there was in the second.

The main issue is that, if the way that data gets recorded in centres is not standardised (and it currently is not), then the assumptions which underpin the calculation of reliability indices based on that data are not valid. The situation is therefore more complex than simply requiring vocational awarding organisations to collect item-level data from centres – standardising how and when data gets recorded in centres is a necessary prerequisite if standard estimates of reliability are to be used. An alternative would be to calculate simpler statistics such as means and variances of item scores where data recording is not standardised.

Standardising how and when data gets recorded in centres needs to be approached with care, however: it is possible that enforcing too rigid a framework could result in losing some of the formative benefits of allowing centres to work in the different ways they do. As an aside, the use of e-portfolios offers one way to standardise without centres necessarily even realising that it is happening. However, this is not an approach which could be imposed on awarding organisations or centres. Further research is required to explore how awarding organisations could strengthen their systems in this area.

2.4.3 Internally set and marked tests

In the case of internally set and marked tests, the focus of the validity evidence will be on the nature of the awarding organisations processes, including:

- evidence that the awarding organisation provides adequate assessment guidelines to centres
- evidence documenting how the results of the assessments are moderated. **This should include copies of visit/moderation/verification reports (the terminology differs), if available**

We think this could be augmented by a requirement to provide:

- **sample assessments and mark schemes created by centres**

The evidence listed in bold above was not available to the case studies, but we believe it is reasonable that awarding organisations should provide it as part of the validity argument for this form of assessment.

In addition, there is a strong argument that awarding organisations should begin to investigate how they can collect more formal data to evidence marker reliability for internally marked assessments. As with internally marked tests, however, there is the issue of standardising the way centres record achievement.

2.4.4 Practical assessments, assignments, projects and case studies

Where practical assessments have been externally set, they will be locally marked and externally verified. While recognising that it is an important aspect of this form of assessment to allow for a wide range of evidence to be accumulated by the candidate, and also that variability may be introduced by differences in the facilities available at centres,

assessments of this nature still require clear and well-structured marking guidelines. Where they are externally set, the evidence produced by the awarding organisation should include:

- assessment specifications
- assessments and marking guidelines
- evidence of the procedures used for creating assessments

Whether externally or internally set, validity evidence must also focus on the awarding organisations' processes for ensuring reliable and valid assessment, including:

- evidence that the awarding organisation provides adequate assessment guidelines to centres
- evidence documenting how the results of the assessments are moderated. This should include copies of visit/moderation/verification reports (the terminology differs), if available
- where assessments are internally set, sample assessments and marking guidelines created by centres

In addition, there is a strong argument that awarding organisations should begin to investigate how they can collect more formal data to evidence inter-marker reliability. The work of Harth and Hemker (2011) is relevant here, and Ofqual has commissioned further work in this area. The issue of standardising the way centres record achievement is, however, also relevant here: see Section 2.4.2.1.

2.4.5 Other forms of assessment evidence

As indicated in Section 2.4, other forms of assessment evidence which may be submitted to demonstrate candidate achievement include:

- **Records of assessor observation** can be required to ensure that candidate performance, either over time or on specific occasions, meets the required standards under realistic conditions (e.g. in the workplace). Assessment usually involves a checklist of performance criteria, with evidence recorded for all criteria, either for individual candidates or a group of candidates. Assessment by assessor observation can be the only effective way of assessing some skills (especially interpersonal skills, or where there are health and safety procedures to be observed).

- **Witness testimony** may be used to provide supporting evidence towards a competence-based qualification where it is not possible for an assessor to be present.
- **Records of professional discussions** are typically where the assessor talks to a candidate about how they carry out an activity, and identifies any learning outcomes and assessment criteria which are covered by the actions of the candidate.
- **Candidate and peer reports and other forms of candidate work** could be a formal report or an informal diary or log, or some other example of candidate work produced to give evidence of competence.
- **Evidence of prior learning** requires the centre's internal verifiers to judge whether the available evidence of prior achievement meets the full requirements of the relevant standards and related assessment guidance.

It is difficult to generalise about the validity evidence which should be provided for these forms of assessment – the context and detail of the assessment will vary greatly by qualification. The awarding organisation should, however, be required to produce sufficient evidence to demonstrate that the assessment method is appropriate, that the administration of the assessment is appropriate, and that the monitoring and standardisation of assessment is adequate for an Ofqual reviewer to be able to judge the validity and likely reliability of the approach. Examples of appropriate evidence might include:

- justification for the method of assessment (including which learning outcomes and assessment criteria this type of assessment is used for, and why) **and documentary evidence of stakeholder support for the method of assessment**
- the qualifications/experience of any assessors/witnesses
- how results are moderated by the awarding organisation

2.4.6 Summary

In this section we have considered the evidence which Ofqual might reasonably require for different methods of assessment, and which areas of the Ofqual validity framework template the evidence could contribute to. Figure 5 summarises how the evidence from different forms of assessment might map against the headings in the Ofqual validity framework template.

In the next section, we will draw upon this when we consider each section of the Ofqual validity framework template in turn and consider the evidence which awarding organisations may be able to provide.

Validity framework area	External assessment	Internally assessed, externally verified assessment							Additional forms of evidence			
	Externally set and marked tests	*Externally set, internally marked tests*	*Internally set and marked tests*	*Practical assessments*	*Oral assessments*	*Assignments*	*Projects*	*Case studies*	*Assessor observation*	*Witness testimony*	*Professional discussions*	*Candidate and peer reports*
Alignment between assessment and the curriculum/syllabus												
Test specification and assessment methods are appropriate	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Test content, mark scheme, and component weightings are appropriate	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Accuracy and reliability of scores												
Assessment development and administration procedures are appropriate	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Construct-related evidence is adequate	✓											
Score reliability/ generalisability is adequate	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Alignment of boundary scores with performance standards												
The standard setting procedures are appropriate	✓	✓										
The boundary scores are set appropriately and there is good correspondence between performance on the assessment and the defined performance standard	✓	✓										
The outcomes are accurate	✓	✓										
The operational standards are perceived to be appropriate by the main stakeholders												
Indicator of future performance												
The performance on the assessment is a good predictor for future performance												
The performance on the assessment is perceived by employers and university admission officers to be useful in predicting future performance												
The course is perceived by employees and/or students to be helpful for their future work or study												

Figure 5. Showing how the evidence from different forms of assessment might map against the headings in the Ofqual validity framework template. Black ticks indicate strong evidence, weaker ticks indicate some evidence

3 Application of the Ofqual validity framework to vocational qualifications

3.1 Application of the Ofqual framework to the case study qualifications

To recap, the qualifications which formed the basis of the case studies were:

- City & Guilds
  - Level 3 Diploma in Pharmaceutical Science (QAN: 500/9959/0)
  - Level 2 Diploma in Professional Cookery (QAN: 500/8909/2)
- NCFE
  - Level 1 Functional Skills English (QAN: 501/1660/5)
  - Level 3 Certificate in Principles of Customer Service (QAN: 600/2922/5)

As indicated in the methodology section, evidence was gathered (from documents and interviews) for each of the four case study qualifications. Once the evidence gathering was complete, reviewers working in pairs compiled and assessed the validity argument for each qualification, and produced a validity report for each qualification. The validity reports were compiled using the Ofqual template shown in Appendix 1, so that a fully populated template was generated for each case study qualification. Summary reports for the case studies are included in Appendix 3.

3.2 Applying the Ofqual validation framework

In this section, we draw upon the mapping of assessment methods presented in Section 2.4 and the lessons learnt in the four case studies in order to consider how vocational awarding organisations might provide evidence against each section of the Ofqual validity framework template. This, therefore, is a consideration of research question 2 against the Ofqual framework.

3.2.1 Alignment between assessment and the curriculum/syllabus

3.2.1.1 Test specification and assessment methods are appropriate

Questions within this section of the Ofqual validity framework template are:

- How does the assessment specification reflect the learning outcomes?
- How do the assessment criteria reflect the required standards?
- Are the methods of assessment appropriate for the construct to be assessed?

Awarding organisations should be able to provide sufficient evidence to enable Ofqual to come to an informed judgement in this area. Depending on the method of assessment, evidence to answer these questions will draw from:

- assessment specifications
- unit and qualification specifications
- test papers and mark schemes
- assessment guidelines provided by the awarding organisation to centres

In considering whether the methods of assessment are appropriate for the construct to be assessed, Ofqual will often come across the issue of the awarding organisation's choice of internal as opposed to external assessment. Vocational qualifications make extensive use of internal assessment with external verification of a portfolio of evidence. Where this form of assessment is used in the assessment of practical skills, Ofqual should find little problem in coming to a view on the appropriateness of the method of assessment. Portfolio-based assessment is also, however, used on occasion to provide evidence of the supporting knowledge required to apply skills. There are a number of issues which Ofqual reviewers will need to be aware of:

- Where supporting knowledge and understanding is assessed as part of a competence-based qualification, a portfolio is often used to ensure a holistic approach and to limit the burden of assessment.
- Where assessment of knowledge and understanding to support the development of skills is developed as part of a knowledge-based qualification, externally set tests may be an option. The wording of the learning outcomes (often the result of a shared development involving SSCs) may not always, however, support external

testing or may limit the opportunity to test by external assessment methods. Much, therefore, depends on how the learning outcomes have been worded.

One other form of evidence we considered here is documented support from stakeholders that the method of assessment is appropriate. Ofqual is unlikely to wish to rely on such evidence (i.e. letters of support) fully, but it may be particularly useful in accredited qualifications where Ofqual reviewers are unlikely to have specific knowledge (which, given the great diversity of vocational qualifications, is far more likely than for general qualifications). We would suggest that the qualification development stage would represent a good opportunity for awarding organisations to garner such evidence in support of their decision on assessment methods.

3.2.1.2 Test content, mark scheme, and component weightings are appropriate

Questions within this section of the Ofqual validity framework template are:

- How are the assessment and the results perceived to be accurate and appropriate by the main stakeholders?
- How appropriate is the mark scheme in terms of the weightings assigned to individual tasks in a component and to individual components in relation to their relative importance within the assessed domain?
- How are the assessment tasks judged to reflect the assessed domain of contents by assessment and content experts as specified by the test specification?
- How are the assessment tasks judged to reflect the main objectives of the curriculum by curriculum developers and content experts? Are they assessing the appropriate knowledge and skills required by the learning outcomes?
- How representative is the assessed domain of content of the target domain of content defined by the entire curriculum/syllabus?

Evidence to answer all but the first of these questions will draw from:

- assessment specifications
- unit and qualification specifications
- test papers and mark schemes
- assessment guidelines provided by the awarding organisation to centres

Based on the evidence of the case study work, however, awarding organisations are unlikely to have carried out any evaluative work post-delivery to determine how content experts

view the assessment tasks or how the assessment and the results are perceived by the main stakeholders (although they would argue that their centre visits provide an informal route by which such information is garnered). Any such work is generally outside current practice for vocational awarding organisations, and would add costs which would ultimately be borne by centres and learners. More fundamentally, however, the awarding organisations' argument about the content experts' opinion of the assessment tasks is likely to be that subject experts are already used to develop, deliver and moderate the assessment tasks. Based on the case study work, we believe that for many qualifications the argument that content experts are deeply involved in defining and implementing the assessment process may be sufficient, providing that the awarding organisation can provide evidence of the credentials and involvement of subject experts and of their monitoring procedures. For high-impact qualifications, however, evaluative work post-delivery may be deemed appropriate (see Section 4). A similar argument may be advanced as to how the assessment and the results are perceived to be accurate and appropriate by the main stakeholders. For many qualifications there is stakeholder involvement in the delivery and assessment processes, and the fact that there is a genuine market in place for vocational qualifications goes some way to providing evidence of stakeholder support (funding effects notwithstanding). Again, for high-impact qualifications, however, there is an argument in favour of formal evaluation. The options for how evaluative work such as this might be implemented are considered in Section 4.

3.2.2 Accuracy and reliability of scores

3.2.2.1 Assessment development and administration procedures are appropriate

Questions within this section of the Ofqual validity framework template are:

- What are the procedures used for creating items and constructing tests?
- Are standardised procedures followed when administering the assessment, and how are they monitored?

Awarding organisations should be able to provide sufficient evidence to enable Ofqual to come to an informed judgement in this area. However, as noted in Section 2.4.2.1, standardising how and when data gets recorded in centres is an area which requires consideration.

3.2.2.2 Construct-related evidence is adequate

Questions within this section of the Ofqual validity framework template are:

- Internal structure of the assessment – How did the items/tasks perform within individual components, and how are components related? If statistical models were used, was their use appropriate (how well does the test data meet the model assumptions and fit the model)?
- Analyses of construct irrelevance variance – How did the assessment tasks and individual components perform for different subgroups of candidates, taking into account their differences in ability?
- Evidence of convergent and discriminant validity – Evidence of convergent and discriminant aspects of validity may also be collected for analysis if relevant data is available.
- Evidence of concurrent validity – How well are results from the assessment correlated to outcomes from other assessments that measure similar constructs at the same time or in close proximity (if data is available)?

In relation to the first question (internal structure of the assessment), where the assessment method involves external tests, as noted in Section 2.4.1.1, awarding organisations should be required to collect and analyse item level data. For the reasons noted in Section 2.4.2.1, this is more problematic where tests are internally marked – standardisation of how and when data is recorded in centres is required to support collection and analysis of internally marked item-level data. For other forms of internally assessed and externally moderated assessment (portfolio, etc.), the concept of item-level data does not exist, and the issue of how achievement is recorded is perhaps exacerbated. In addition, the assessment will generally be one of competence against the assessment criteria, where (on the QCF) all assessment criteria must be satisfied to achieve each individual unit.

In relation to analyses of construct irrelevance variance, as noted in Section 2.4.1.2, for externally marked tests we are recommending that awarding organisations are encouraged to strengthen their systems over time to collect more data about their candidates in order to support analyses of the performance of sub-groups of candidates; we are also recommending, however, that Ofqual need to understand and recognise the limitations of this analysis for many vocational qualifications. For internally assessed and externally moderated assessment (portfolio, etc.), the argument advanced by awarding organisations

is likely to be based on their procedures for moderation, and that the non-prescriptive nature of portfolio-based assessment ensures that different subgroups of learners are able to provide a wide variety of relevant evidence against the assessment criteria, which mitigates against bias. Here, too, there is a strong case for awarding organisations to collect data about their candidates in order to support analyses of the performance of sub-groups of candidates. As noted in Section 2.4.1.2, however, further consideration would need to be given to the practicalities of how this could be achieved.

In relation to the other questions in this section of the Ofqual validity framework template (evidence of convergent and discriminant validity and concurrent validity), as noted in Section 2.4.1.2, these are potentially difficult areas for awarding organisations to address routinely, as is discussed further in Section 4.

3.2.2.3 Score reliability/generalisability is adequate

Questions within this section of the Ofqual validity framework template are:

- Mark scheme – Is the mark scheme appropriate for consistent interpretation to maximise consistency in marking between markers?
- Marker training, marking standardisation, and marking quality monitoring – What procedures are followed to ensure marker reliability?
- Component reliability – Estimates of marker-related, test-related and overall component level reliability.
- Composite reliability – How are components weighted when aggregating component scores? How are components correlated? How reliable are the aggregated scores or composite scores?

Where the assessment method involves externally set and marked tests, as noted previously, awarding organisations should be required to collect and analyse item-level data in order to provide evidence against these questions (although, for QCF qualifications where learners must pass each unit to achieve the qualification, investigations of composite reliability are meaningless). For externally set and internally marked tests, again the issue of standardisation of the recording of results will need to be addressed before data on internal marking can be properly collected and analysed.

Workplace-based assessments, where the assessment is often carried out by people within that context, present particular problems for studies of reliability (and hence validity). The

assessor has to judge whether a candidate has produced sufficient evidence to demonstrate mastery (with reference to the assessment criteria). Reliability in this context therefore relates to the consistency of classification decisions (i.e. mastery demonstrated or not) rather than measurement error in relation to scores. Having a second assessor check the evidence seen by the first is often not possible. The evidence itself may be transient, or may need to be observed over a protracted period. Furthermore, it would seldom be practically possible to have more than one assessor. Internal moderation is therefore the process used to achieve notional 'second' assessment, as it is a process which checks the assessor's decision. Best practice also means that there are standardisation meetings so that assessors come together to check their assessment decisions across a range of portfolios, situations and learners.

Currently, therefore, awarding organisations' arguments in this area are likely to focus on process, and to be:

- that assessment guidelines provided to centres are adequate and clear
- that any sample mark schemes/marking guidelines provided to centres are appropriate for consistent interpretation
- that moderators are properly qualified and trained to ensure marking standards are appropriate and maintained
- that moderation processes are adequate and properly documented

Recent research (Harth and Hemker, 2011) has, however, investigated the reliability of workplace-based assessment, and Ofqual has commissioned further work to examine this area. In the study by Harth and Hemker, the researchers developed a methodology for data collection which involved the use of centre-devised assessment records from candidate portfolios and of internal verifier (IV) reports. The records accessed included:

- assessment observation records and feedback sheets
- achievement sheets linking evidence to criteria
- student self-reflective accounts signed off by tutors/witnesses
- planning, feedback and judgement records
- internal verifier reports

Inter-rater agreement reliability and inter-'item' reliability estimates were devised for the three qualifications being studied. Inter-rater reliability in this study relates to the

judgements made by the assessor and the internal verifier. Researchers found that inter-rater (assessor/internal verifier) agreement was high, as was inter-'item' reliabilty (although this could only be estimated for one of the qualifications under study).

There is therefore scope for awarding organisations to begin to collect data to estimate the reliability of workplace-based assessments. Given that Ofqual has commissioned other work in this area, we will not consider these matters any further here, other than to note that the need for centres to standardise how and when data gets recorded is also a necessary prerequisite for this kind of data collection and analyses.

Reliability is therefore an area where vocational awarding organisations can do more, and Ofqual has commissioned other work in this area. For many qualifications, however, the value this will bring will be limited: cohorts may be too small to provide statistical significance. Conversely, the fact that most assessments will only have a single cut-score can be expected to contribute to reliable assessment.

3.2.3 Alignment of boundary scores with performance standards

3.2.3.1 The standard-setting procedures are appropriate

Questions within this section of the Ofqual validity framework template are:

- What procedures are followed by the standard-setting process?
- How are statistical information and professional judgement used to ensure that the standards are set appropriately? The standard-setting process should pay attention to the defined performance standards.

There are a number of ways awarding organisations may choose to provide evidence against the questions in the section of the Ofqual validity framework template.

For externally set and marked tests, awarding organisations should be able to provide full details of standard-setting procedures, including any statistical analysis techniques employed. Based on the case study work, it is likely that few vocational awarding organisations actually use statistical techniques. In part, this seems to be an issue of culture: relatively few vocational qualifications involved graded external assessments, and so the statistical standard-setting tools which are routinely employed for general qualifications are not part of most vocational awarding organisations' armoury. There are some additional

reasons, however, why statistical standard setting is not deployed even for externally set and marked tests:

- Cohorts may be too small to support meaningful statistical decisions
- The sample of candidates is very often severely biased (e.g. by gender, age, or other factors)
- Awarding organisations will have little or no access to prior attainment data

Nevertheless, where tests are externally set and marked, it seems reasonable that awarding organisations should be required to provide some kind of statistical background when establishing cut-scores, even if small cohort sizes mean that the only real value is when aggregating the outcomes over time in order to provide a sanity check and a longitudinal perspective.

As noted in Section 2.3.5.1, however, one key difference between general and vocational qualifications is the widespread use of pre-determined cut-scores for (usually objectively scored) tests in vocational qualifications. Typically cut-scores (usually, but not always, a single pass mark) are determined in advance for each test, using some form of Angoff method (discussed in more detail in Section Setting pass marks for vocational qualifications). In this situation, the evidence which the awarding organisation will provide should include details of:

- **the process followed and the qualifications of the people making the judgements**
- **the assessment instrument itself**[8]
- **how the standard of the items are monitored**

It is recognised that this approach to standard setting is significantly different from that adopted in general qualifications. The approach is widespread in vocational qualifications, however, and has some justification in the literature (Curcin *et al*, 2009; Idle, 2008; Ricker, 2006). We have outlined above the evidence we think should be provided but, if this is an area of concern for Ofqual, it may warrant more research to consider specifically the nature of the evidence required to demonstrate validity.

[8] The Angoff method is difficult to apply in practice for constructed response multi-mark questions (Idle, 2008)

For other forms of internally assessed and externally moderated assessment (portfolio, etc.), the awarding organisation's argument is likely to be based on the assessment of competence against the standards. For QCF qualifications, for example, awarding organisations should be able to demonstrate (from the units themselves) that there are clear and specific assessment criteria associated with each learning outcome, that the qualification specification includes assessment guidance which indicates the range of acceptable evidence mapped against the assessment criteria, and that their processes for moderation encompass any issues related to specific assessment criteria or learning outcomes for each unit of the qualification.

3.2.3.2 The boundary scores are set appropriately and there is good correspondence between performance on the assessment and the defined performance standard

The only question within this section of the Ofqual validity framework template is:

- How do the boundary scores (which define the minimum scores on the assessment that are required for the candidates to be classified as meeting the relevant performance standards – the operationalisation of the performance standards) correspond to the pre-defined performance standards in terms of the breadth and depth of curriculum coverage (involving scripts inspection)?

For QCF qualifications, there must be evidence of competence for each assessment criterion for each learning outcome of each unit which makes up the qualification; the argument is therefore that boundary scores are correlated with the appropriate performance standards (which may in turn be linked to national standards). If a qualification is supposedly based on national standards, we recommend that the evidence provided by the awarding organisation should include:

- **the relevant national standards**
- **a mapping from the national standards to the qualification at the unit level**

3.2.3.3 The outcomes are accurate

Questions within this section of the Ofqual validity framework template are:

- What is the likely error in boundary scores resulting from potential inconsistency in standard setting?
- What is the level of classification accuracy at the overall qualification level?

For externally set and marked tests, awarding organisations should be able to provide full details of standard-setting procedures, including any statistical analysis techniques employed. As noted in section 3.2.3.1 however, it is likely that few vocational awarding organisations actually use statistical techniques. If the awarding organisation provides tests with pre-determined grade boundaries (see Section 2.3.5.1), then in effect standard setting is effected at the test development stage, and there should be a requirement to supply evidence of the efficacy of this approach.

3.2.3.4 The operational standards are perceived to be appropriate by the main stakeholders

The only question within this section of the Ofqual validity framework template is:

- How are the operational standards defined by boundary marks perceived to reflect the performance standards defined for the curriculum by the main stakeholders (inspection of actual candidates' work)?

For qualifications based on national standards, the argument of awarding organisations is likely to be that the operational standards are based directly on national performance standards which have been judged by subject experts to define the required competence levels. For other vocational qualifications, the argument may be that the development of the qualification and assessment arrangements was carried out with the involvement of stakeholders. Based on the case study work, however, there is likely to be little or no post-delivery evaluation evidence available. How such work might be carried out is discussed further in Section 4.

3.2.4 Indicator of future performance

3.2.4.1 The performance on the assessment is a good predictor for future performance

The only question within this section of the Ofqual validity framework template is:

- How well can the results from the assessment predict outcomes from assessments that measure similar constructs in the future (if data is available)?

Based on the case study work carried out, vocational awarding organisations are unlikely to have any data available to provide evidence against this question. Prior attainment data and destination data are generally not available. In some circumstances at most, awarding

organisations may be able to provide some evidence of progression to a similar qualification at a higher level (as long as both qualifications were carried out with the same awarding organisation).

This is a question which would appear much easier to answer for general qualifications, where the cohorts are much more homogenous than is the case with vocational qualifications and both prior attainment and destination data tend to be available.

3.2.4.2 The performance on the assessment is perceived by employers and university admission officers to be useful in predicting future performance

The only question within this section of the Ofqual validity framework template is:

- How useful are the results from the assessment in predicting future performance at university or in workplace?

As with the previous question, destination data is generally not available to vocational awarding organisations. For qualifications based on national standards, awarding organisations may make the argument that the qualification is based on the relevant national standards, which have been judged by subject experts to define the required skills to ensure progression in the workplace.

Based on the case study work, there is likely to be little or no post-delivery evaluation evidence currently available. Section 4 considers how such evidence might be obtained.

3.2.4.3 The course is perceived by employees and/or students to be helpful for their future work or study

Questions within this section of the Ofqual validity framework template are:

- How has what they have learnt helped them with their work?
- How has what they learnt helped them with their study?

Vocational awarding organisations do not currently carry out any post-delivery evaluation which would enable them to answer these questions. This might be an area where awarding organisations could modify their systems to obtain such evidence at relatively little cost. Online surveys are inexpensive to deliver, and it is not too much of a stretch to see how learners could be surveyed post-delivery with an online survey generating purely quantitative data which would support straightforward (and automated) analysis. The pre-requisite for this, however, would be that awarding organisations held the email addresses

of their learners, which is not generally believed to be the case at present. Centres will not supply personal learner information to awarding organisations, citing data protection issues. Based on previous attempts, relying on learners to opt to supply feedback is likely to result in response rates which are low or non-existent. This is an area where awarding organisations would like to engage more with learners; however, more research is required in order to understand what can reasonably be expected and how it can be achieved.

3.2.5 Conclusions

Understanding how the wide variety of assessment methods and data available in vocational qualifications might be used to provide evidence against the various sections of the Ofqual validity framework template is complex. In this section, we have considered the issues from the perspective of the kinds of evidence vocational awarding organisations might reasonably be expected to provide for each particular method of assessment, and also asked what kinds of evidence are actually required by each area of the Ofqual validity framework template and how difficult this might be for vocational awarding organisations to provide. In the next section, we attempt to summarise the requirements which might be put on vocational awarding organisations and consider how Ofqual might implement audits of the validity of vocational qualifications.

4 Approaches to auditing the validity of vocational qualifications

To recap, the research questions for this project are as follows:

- How do awarding organisations ***currently*** justify the validity of their vocational qualification(s) (in terms of being fit for their particular purposes)?
- What evidence to justify the validity of their vocational qualification(s) could awarding organisations ***reasonably be expected*** to produce ***in the future*** in producing a comprehensive and robust validity argument?

In the case study work carried out, both partner awarding organisations provided unfettered access to the available documentation for the qualifications under study. Each of the four case-study qualifications therefore provided an opportunity to discover the nature and extent of evidence currently available to support the validity argument for that qualification. It has become clear in the course of the research that the forms of evidence required can be divided into three categories:

- **Evidence which vocational awarding organisations already have available –** examples of such evidence include units, qualification handbooks, assessments and assessment guidelines, documentation of quality assurance processes, etc.
- **Additional evidence which vocational awarding organisations could collect relatively easily –** examples of this evidence include documentary evidence of stakeholder support for the method of assessment, item-level data for externally set and marked tests, sample assessments and mark schemes created by centres, copies of visit/moderation/verification reports, etc.
- **Additional evidence which vocational awarding organisations would find more difficult or expensive to collect routinely –** such examples include formal data to evidence marker reliability for internally marked assessments, quantitative and qualitative evidence of how assessments and results are perceived to be accurate and appropriate by stakeholders, evidence of how useful results from assessments are in predicting future performance in the workplace, etc.

There is, therefore, a judgement to be made on the third category of evidence: to what extent should vocational awarding organisations be expected to modify their current processes in order to routinely gather additional validity evidence which would add (perhaps significantly in some cases) to the costs of qualifications?

Salient points for consideration include:

- It is difficult to quantify what the additional costs would be for vocational awarding organisations to collect the third category of evidence identified above. Entry numbers for many vocational qualifications are far lower than for general qualifications, however, and so the overhead of collecting these forms of validity evidence will, for most qualifications, be proportionally higher. This is particularly true for the 'long tail' (i.e. the large numbers of qualifications with relatively low entries) of vocational qualifications, which constitute a significant part of many vocational awarding organisations' businesses. There may also be significant structural costs required, e.g. changes or additions to vocational awarding organisations' core IT systems. Investment required in structural changes of this nature will detract from investment in other areas of development.
- Vocational awarding organisations would also argue that there is a genuine market operating for vocational qualifications which there is not in general qualifications, and that market forces (notwithstanding the impact of funding) play a key role in ensuring certain aspects of validity; the involvement of stakeholders in qualification development and delivery that we have seen in the case study work provides some support for this view.
- Ofqual already recognises certain qualifications as being high impact. All regulated qualifications need to be valid, but there is arguably a stronger case for strengthening the collection of validity evidence for some qualifications than others.
- The studies required to produce some forms of validity evidence (e.g. evidence of convergent and discriminant validity, reliability studies of observation-based assessment, evidence of concurrent validity, etc.) require specialist expertise to design and conduct, and are like to prove inherently difficult for many vocational awarding organisations to implement currently. In addition, centres may well need to be incentivised to cooperate with these forms of research, and it is not clear how this could be done.

Based on these points, and on the case study work carried out, we believe there is an argument for two levels of validity audit: a **standard validity audit** and a **high-impact qualification validity audit**.

A **standard validity audit** would draw upon the evidence of validity that Ofqual might reasonably expect vocational awarding organisations to generate in the normal course of their operations. This would include the following (items listed **in bold** are those which were

not available to the case studies but which we are recommending vocational awarding organisations could reasonably be expected to produce):

- Justification for the method of assessment (including which learning outcomes and assessment criteria this type of assessment is used for and why) **and documentary evidence of stakeholder support for the method of assessment**
- For external assessment:
    - assessment specifications
    - test papers and mark schemes
    - assessment guidelines
    - evidence of the procedures used for creating items and constructing tests
    - evidence of the procedures followed when administering the assessment, and how they are monitored
    - **estimates of reliability of candidate results**
- For internal assessment:
    - **sample assessments and mark schemes created by centres**
    - evidence documenting how the results of internal assessments are moderated, **to include copies of visit/moderation/verification reports, if available**
    - **evidence of how the grade boundaries are moderated by the awarding organisation**
    - the qualifications/experience of any assessors/witnesses
- Evidence of the procedures followed by the standard-setting process – **where pre-determined cut scores are used, detailed documentation of the process followed and the rationale for the choice of cut-score(s) should be provided.**

A **high-impact qualification validity audit** would be carried out less frequently than a standard audit. For specific qualifications, a high-risk audit would seek to generate and assess one or more forms of additional validity evidence, which would not be available to vocational awarding organisations in the normal course of their operations. A high-impact audit may involve one or more of the following:

- A review of assessment tasks by independent content and assessment experts to judge how they reflect the assessed domain and how they reflect the main objectives of the curriculum

- Reliability studies:
    - collection of data to estimate marker reliability for internally marked assessments
    - collection of data to estimate the reliability of observation-based assessments in the workplace
    - re-test studies
- Stakeholder consultation exercises:
    - investigation of how operational standards are perceived to reflect the performance standards by stakeholders
- Future performance investigations:
    - how well results from assessments predict outcomes from assessments that measure similar constructs, and how useful results are in predicting future performance in the workplace
- Learner studies:
    - how what learners have learnt has helped them with their work and/or study
- Post-delivery evaluation of the efficacy of the awarding organisation's procedures for setting pre-determined cut-scores
- Convergent, discriminant and concurrent validity studies

The concept of a high-impact qualification validity audit may also relate well to the concepts of risk-based regulation and high-impact qualifications. Validity audits could be implemented as follows:

- In the first instance, a standard audit could be carried out across a selection of an awarding organisation's qualifications, with one or more high-impact qualification validity audits carried out on selected high-impact qualifications
- A single validity report for the awarding organisation would be compiled containing:
    - individual validity reports for each qualification investigated
    - an aggregate view of the key findings and any requirements for improvements to awarding organisation processes

Improvements in awarding organisation processes as a result of the findings of one validity audit would provide improvements across the range of qualifications.

One further reason to consider the implementation of validity audits in the way described relates to the levels of expertise required to fully implement and assess the Ofqual validity

framework. The Ofqual validity framework is a sophisticated and complex validation framework, which will need some expertise to put into effect. It is unlikely that the vocational sector as a whole has sufficient expertise to effectively implement all aspects of the framework meaningfully. Equally, it is unclear whether Ofqual has the resources required to assess the evidence produced. Having two levels of audit will reduce the amount it will cost both sides to ensure the validity of vocational qualifications.

5 Further work

There are a number of areas where further research may be warranted to test the practicality of implementation of the recommendations made in this report:

- Standardising how and when data gets recorded in centres would help provide reliability evidence. Standardising how and when data gets recorded in centres needs, however, to be approached with care: it is possible that enforcing too rigid a framework could result in losing some of the formative benefits of allowing centres to work in the different ways they do.
- Some forms of validity evidence would need to involve centres and/or stakeholders, or would require awarding organisations to collect data about learners. In each case, there are practical issues to be addressed (for example, would centres need to be incentivised to cooperate with this type of research and, if so, how?). Furthermore, gathering certain kinds of validity evidence (for example, evidence of convergent and discriminant validity) is likely to be difficult for vocational awarding organisations, either for reasons of cost or because the studies would be more technically challenging and the awarding organisation may not have the expertise to devise and implement a suitable study.

In each case, a small pilot involving awarding organisations and centres would help to clarify the required processes, how these might be implemented, and what the implications and the barriers would be for awarding organisations and centres. Ultimately, it may prove to be the case that, for the most technically challenging questions in the Ofqual validity framework, questions may need to be carefully tailored to the specific sector and might need a more qualitative response around key stakeholder perceptions/experiences. Ofqual's framework may need to be broadened slightly to allow for this.

Similarly, although this project has investigated the validity of vocational qualifications using Ofqual's validity framework template as it is currently worded, in practice vocational awarding organisations may find the process more accessible if a version of the template tailored to the vocational context was produced. For example, the wording of the Ofqual validity framework template often seems to make the implicit assumption that the method of assessment is externally marked tests (i.e. the unqualified references to items and mark schemes, marking standardisation, etc.). This is unlikely to be helpful to vocational awarding organisations trying to understand the evidence they need to provide to demonstrate

validity. It may help vocational awarding organisations if the wording of the Ofqual validity framework template reflected the different forms of evidence which could be accepted based on the method of assessment.

6 References

AQA (2012), Uniform marks in A-level and GCSE exams and points in the Diploma, http://store.aqa.org.uk/over/stat_pdf/UNIFORMMARKS-LEAFLET.PDF

Berk, R.A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. Review of Educational Research, 56, 137-172.

Cambridge Assessment (2011), Setting grade boundaries in the UK and maintaining standards across exam boards, http://www.cambridgeassessment.org.uk/ca/digitalAssets/200020_Factsheet_2_-_Setting_Grade_Boundaries.pdf

Cedefop (European Centre for the Development of Vocational Training) (2010). 'Changing Qualifications: A Review of Qualifications Policies and Practices'. Luxembourg: Publications Office of the European Union.

Cizek, G. J., & Bunch, M. B. (2007). Standard Setting: A guide to establishing and evaluating performance standards on tests. Thousand Oaks, CA: Sage

Cizek, Gregory J. (Ed) (2011), Setting Performance Standards. Foundations, Methods, and Innovations, 2nd Edition. Routledge, 2011

Crooks, T.J., Kane, M.T. & Cohen, A.S. (1996). Threats to the valid use of assessments. Assessment in Education: Principles, Policy and Practice, 3, 3, 265–285

Curcin, Milja, Black, Beth, Bramley, Tom (2009). Standard maintaining by expert judgment on multiple-choice tests: a new use for the rank-ordering method. Paper presented at the British Educational Research Association annual conference, University of Manchester, September 2009.

DfE (2012), The evaluation of the impact of changes to A levels and GCSEs - Final Report, https://www.education.gov.uk/publications/standard/publicationDetail/Page1/DFE-RR203

DfES (2006). Further Education: Raising Skills, Improving Life Chances, London: The Stationery Office, Cm 6768

Foster, A. (2005). 'Realising the Potential - A review of the future role of further education colleges'. DfES and LSC.

Frederiksen, J.R. & Collins, A. (1989). A systems approach to educational testing. Educational Researcher, 18, 9, 27–32.

Harth, Helen and Hemker, Bas T. (2011) 'Reliability of Vocational Qualifications'. http://www2.ofqual.gov.uk/files/reliability/11-03-16-On-the-reliability-of-results-in-vocational-assessment.pdf

Hubley, A. M., & Zumbo, B. D. (1996). A dialectic on validity: Where we have been and where we are going. Journal of General Psychology, 123, 207-215.

Hubley, Anita M. and Zumbo , Bruno D. (2011). Validity and the Consequences of Test Interpretation and Use. Social Indicators Research, v103 n2 p219-230 Sep 2011

Idle, Sally (2008). An investigation of the use of the Angoff procedure for boundary setting in multiple choice tests in vocational qualifications.http://www.iaea2008.cambridgeassessment.org.uk/ca/digitalAssets/180443_Idle.pdf

Johnson, M. (2008) 'Assessing at the borderline: Judging a vocationally related portfolio holistically'. Issues in Educational Research, 18 (1), 26-43.

Johnson, N. (2007). A consideration of assessment validity in relation to classroom practice. Cambridge Assessment. International Association for Educational Assessment Annual

Conference, Baku, Azerbaijan, September 2007. http://www.iaea.info/documents/paper_1162b25504.pdf

Kane, M. T. (2001). Current concerns in validity theory. Journal of Educational Measurement, 38, 319-342.

Kane, M.T. (2006).Validation. In: R.L. Brennan (Ed.), Education Measurement (4th ed.). Westport: American Council on Education/Praeger.

Lin,Jie (2006). The Bookmark Standard Setting Procedure: Strengths and Weaknesses, http://www2.education.ualberta.ca/educ/psych/crame/files/standard_setting.pdf

Linn, R.L., Baker, E.L. & Dunbar, S.B. (1991). Complex, performance-based assessment: expectations and validation criteria. Educational Researcher, 20, 8, 15–21

MacCann, R. G. and Stanley, G (2006). The Use of Rasch Modeling To Improve Standard Setting, Practical Research and Evaluation, Volume 11 Number 2, January 2006. http://pareonline.net/pdf/v11n2.pdf

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed.). New York: American Council on Education.

Newton, Paul E. & Shaw, Stuart D. (2012). The meaning of validity: consensus, what consensus?. http://www.cambridgeassessment.org.uk/ca/digitalAssets/201813_Validity_Seminar_1_Final.pdf Accessed 25-01-2013

Ofqual (2009). Identifying purposes for qualifications in the Qualifications and Credit Framework. http://www2.ofqual.gov.uk/downloads/category/112-qualifications-and-credit-framework-qcf?download=355%3Aidentifying-purposes-for-qualifications-in-the-qualifications-and-credit-framework-february-2009

Ofqual (2008), 'Regulatory arrangements for the Qualifications and Credit Framework'. http://www.ofqual.gov.uk/files/Regulatory_arrangements_QCF_August08.pdf

Ofqual, How are A Level Grades Awarded,http://www2.ofqual.gov.uk/downloads/category/143-presentations-briefings?download=796 %3Ahow-are-a-level-grades-awarded

Ofqual (2001). GCSE, GCE, Principal Learning and Project Code of Practice, http://www.ofqual.gov.uk/for-awarding-organisations/96-articles/247-code-of-practice-2011

Opposs, D. and He, Q. (2011) 'The Ofqual Assessment Validity Programme'. Ofqual.

Ricker, Kathryn L. (2006) Setting Cut-Scores: A Critical Review of the Angoff and Modified Angoff Methods, The Alberta Journal of Educational Research Vol. 52, No. 1, Spring 2006, 53-64

Shaw, S.D. & Weir, C.J. (2007). Examining Writing: Research and Practice in assessing second language writing, Studies in Language Testing,Volume 26, Cambridge: University of Cambridge Local Examination Syndicate/Cambridge University Press.

Stasz, Cathy (2011). The Purposes and Validity of Vocational Qualifications. SKOPE Research Paper No. 105 November 2011

Torrance, H., Colley, H., Garratt, D., Jarvis, J., Piper, H., Ecclestone, K and James, D. (2005). 'The Impact of Different Modes of Assessment on Achievement and Progress in the Learning and Skills Sector', London: Learning and Skills Research Centre.

Weir, C.J. (2005). Language testing and validation: An evidence-based approach. Basingstoke: Palgrave Macmillan

Wolf, A. (1998). 'Portfolio assessment as national policy: the National Council for Vocational Qualifications and its quest for a pedagogical revolution', Assessment in Education, 5 (3), 413-445.

Wolf, A. (2011). 'Review of Vocational Education – The Wolf Report'. DFE-00031-2011

Young, M.F.D. (2008). Bringing Knowledge Back In: From social constructivism to social realism in the sociology of education. Oxon: Routledge.

7 Appendix 1 Ofqual's validity framework template

Case study:

Awarding organisation:

Overview
Strengths
Weaknesses
Other points of interest
List of evidence reviewed

1. Alignment between assessment and the curriculum/syllabus

Test specification and assessment methods are appropriate			
	Question	**Validity argument**	**View / judgement**
1a	How does the assessment specification reflect the learning outcomes?		
1b	How do the assessment criteria reflect the required standards?		

1c	Are the methods of assessment appropriate for the construct to be assessed?		

Test content, mark scheme, and component weightings are appropriate			
	Question	**Validity argument**	**View / judgement**
1d	How are the assessment and the results perceived to be accurate and appropriate by the main stakeholders?		
1e	How appropriate is the mark scheme in terms of the weightings assigned to individual tasks in a component and to individual components in relation to their relative importance within the assessed domain?		
1f	How are the assessment tasks judged to reflect the assessed domain of contents by assessment and content experts as specified by the test specification?		
1h	How are the assessment tasks judged to reflect the main objectives of the curriculum by curriculum developers		

	and content experts? Are they assessing the appropriate knowledge and skills required by the learning outcomes?		
1i	How representative is the assessed domain of content of the target domain of content defined by the entire curriculum/syllabus?		

Accuracy and reliability of scores

Assessment development and administration procedures are appropriate			
	Question	**Validity argument**	**View / judgement**
2a	What are the procedures used for creating items and constructing tests?		

2b	Are standardised procedures followed when administering the assessment, and how are they monitored?		

Construct-related evidence is adequate			
	Question	**Validity argument**	**View / judgement**
2c	Internal structure of the assessment: How did the items/tasks perform within individual components and how are components related? If statistical models were used, was their use appropriate (how well do the test data meet the model assumptions and fit the model)?		
2d	Analyses of construct irrelevance variance: How the assessment tasks and individual components performed for different subgroups of candidates, taking into account their differences in ability?		

2e	Evidence of convergent and discriminant validity: Evidence of convergent and discriminant aspects of validity may also be collected for analysis if relevant data is available.		
2f	Evidence of concurrent validity: How well are results from the assessment correlated to outcomes from other assessments that measure similar constructs at the same time or in close proximity (if data is available)?		
Score reliability/ generalisability is adequate			
	Question	**Validity argument**	**View / judgement**
2g	Mark scheme: Is the mark scheme appropriate for consistent interpretation to maximise consistency in marking between markers?		

2h	Marker training, marking standardisation, and marking quality monitoring: What procedures are followed to ensure marker reliability?		
2i	Component reliability: Estimates of marker-related, test-related and overall component level reliability.		
2j	Composite reliability: How are component weighted when aggregating component scores? How are components correlated? How reliable are the aggregated scores or composite scores?		

Alignment of boundary scores with performance standards

The standard setting procedures are appropriate			
	Question	**Validity argument**	**View / judgement**
3a	What procedures are followed by the standard setting process?		

3b	How are statistical information and professional judgement used to ensure that the standards are set appropriately? The standard setting process should pay attention to the defined performance standards.		

The boundary scores are set appropriately and there is good correspondence between performance on the assessment and the defined performance standard			
	Question	**Validity argument**	**View / judgement**
3c	How do the boundary scores (which define the minimum scores on the assessment that are required for the candidates to be classified as meeting the relevant performance standards - the operationalization of the performance standards) correspond to the pre-defined performance standards in terms of the breadth and depth of curriculum coverage (involving scripts inspection)?		

The outcomes are accurate			
	Question	**Validity argument**	**View / judgement**
3d	What is the likely error in boundary scores resulting from potential inconsistency in standard setting?		

3e	What is the level of classification accuracy at the overall qualification level?		

The operational standards are perceived to be appropriate by the main stakeholders			
	Question	**Validity argument**	**View / judgement**
3f	How are the operational standards defined by boundary marks perceived to reflect the performance standards defined for the curriculum by the main stakeholders (inspection of actual candidates' work)?		

Indicator of future performance

The performance on the assessment is a good predictor for future performance			
	Question	**Validity argument**	**View / judgement**
4a	How well can the results from the assessment predict outcomes from assessments that measure similar constructs in the future (if data is available)?		

The performance on the assessment is perceived by employers and university admission officers to be useful in predicting future performance			
	Question	**Validity argument**	**View / judgement**

4b	How useful are the results from the assessment in predicting future performance at university or in workplace?		

The course is perceived by employees and/or students to be helpful for their future work or study			
	Question	**Validity argument**	**View / judgement**
4c	How has what they have learnt helped them with their work?		

4d	How has what they learnt helped them with their study?		

8 Appendix 2 – Evidence reviewed for Level 3 Diploma in Pharmaceutical Science

Level 3 Diploma in Pharmaceutical Science

Awarding organisation: City and Guilds

City & Guilds Generic Documentation

- Checklist for evaluating QCF units v0 5 Feb 11
- Developing Assignments (staff and consultants' manual)
- Developing qualifications- learning outcomes and assessment criteria
- DEVELOPMENT-CONSULTANT-PROFILE-V01
- Item Writing Principles
- Making assessment choices
- Multiple Choice item bank reviews handbook
- Principles and policies of assessment
- Setting and editing Short and Structured Response Question Papers

Level 3 Diploma in Pharmaceutical Science Documentation

- Level 3 Diploma in Pharmaceutical Science v10
- Pharmacy_Assessor_guidance_-_January_2012
- Meeting notes with AO staff and stakeholders
- Information provided by AO staff via telephone and email
- Assessments and mark schemes for the following units:
    - 5356-002_Biological_Principles_for_Pharmacy
    - 5356-003_Microbiology_for_Pharmacy
    - 5356-004_Human_Physiology_for_Pharmacy
    - 5356-005_Action_and_Uses_of_Medicines
    - 5356-006_Gastrointestinal_and_Nutritional_Medicines
    - 5356-007_Cardio-Respiratory_Medicines
    - 5356-008_Central_Nervous_System_Medicines_and_Anaesthesia
    - 5356-009_Infections_Immunological_Products_and_Vaccines
    - 5356-010_Endocrine_and_Genito-Urinary_Medicines(1)

- 5356-011_Malignant_Disease_Immunosuppressive_and_Musculoskeletal_Medicines
- 5356-012_Eye_Ear_Nose_and_Dermatological_Medicines
- 5356-013_Community_Pharmacy_Practice
- 5356-014_Professional_Development_in_Pharmacy
- 5356-015_Communicating_in_Pharmacy
- 5356-016_Dispensing_and_Supply_of_Medicines
- 5356-017_Pharmaceutics
- 5356-018_Pharmacy_Law_Ethics_and_Practice
- 5356-019_Making_Medicines_for_Pharmacy

9 Appendix 3 - Summary reports for each case study qualification

Although the investigative work concentrated on specific qualifications, the focus of the research is not on the validity of the qualifications *per se*; rather, it is to investigate validity *processes*, and to show what a validity report might look like for a vocational qualification. However, it is perhaps useful to include here the summary sections of the validity reports for each case study qualification.

9.1 Level 2 Diploma in Professional Cookery

Awarding organisation: City & Guilds

Summary

'The stated purpose of this qualification is to allow candidates to learn, develop and practise the skills required for employment and/or career progression in the catering and hospitality sector. It contributes to the knowledge and understanding of the related Level 1 NVQ Diploma in Food Preparation and Cooking (7131) and Level 2 NVQ in Professional Cookery (7132), whilst containing additional skills and knowledge which go beyond the scope of the National Occupational Standards (NOS). It is the view of the professional chefs who deliver the Diploma that it is much more relevant to the demands of a commercial kitchen than the NVQ.

'The underpinning assessment arrangements seem to be fit for purpose. There is strong evidence of stakeholder involvement in its early development and this continues to be the case through organisations such as the Professional Association of Catering Educators (PACE), which acts as a conduit of advice and guidance to City & Guilds (C&G) from the professionals in the field. It is clear too from evidence from the professional chefs, that the involvement of employers and the general public (through their access to the colleges' training restaurants) is a key factor in quality assuring the Diploma. There is no evidence to suggest that the standards are not appropriate to the requirements of the industry nor that the assessment package is failing to address the criteria. The fact too that the majority of students following the Level 2 Diploma progress to Level 3 after which they find employment often within the local employer network is further evidence.

'From interviews with the awarding organisation it is clear that the qualification is supported by SSC and PACE, that both were involved in the development and review, with support provided for the accreditation of the qualification. However there is no documentary evidence of this in the provided documentation, and so for the future purposes of a validity audit, the awarding organisation should obtain written verification from the SSC, the chefs and from PACE to confirm both their initial and continuing involvement in this qualification, with particular regard to the role they have in the assessment development.

Strengths

- *'The Diploma has been developed with the full involvement of professionals from the sector, and has their continued involvement to ensure that the required standards are maintained. Professional chefs are responsible for the assessment of the students and the fact that the qualification is so strongly supported by employers in the sector supports valid assessment.*

Weaknesses

- *'Although the documentation is comprehensive in most respects, it would be helpful if the mapping references to the NOS that are alluded to could actually be included.'*

9.2 Level 3 Diploma in Pharmaceutical Science

Awarding organisation: City and Guilds

Summary

'The qualification is required to support the mandatory recognition of fitness to practise as a pharmacy technician in conjunction with the related NVQ. The qualification consists of 19 mandatory units. These units are each assessed through short answer and context specific, task-based questions suited to the knowledge-based nature of the qualification and the requirement for the application of skills in a specific context. In adhering to the General Pharmaceutical Council (GPhC) standards the content of the qualification includes significantly greater detail within the knowledge requirements than is included within the related national occupational standards (NOS). This is in keeping with the technical nature of the job role and the purpose of the qualification.

'The changes to the assessment approach used for this current version of the qualification have introduced a greater aspect of controlled assessment than was previously available. The awarding organisation is intending to investigate how further changes to the assessment approach may be incorporated in future versions to enhance this further and to provide opportunities for more detailed statistical monitoring and moderation of the assessment itself. Conversations with the awarding organisation indicate that this change to the assessment approach was implemented in order to better reflect the assessment of knowledge and understanding outlined in the learning outcomes and in keeping with current awarding organisation practice for many of its other vocational qualifications. This is a transitional approach with a view to considering a move to online assessment sometime in the future. (The awarding organisation did not consider an assignment based approach appropriate for online assessment.)

'Discussions with the GPhC confirmed that they were directly responsible for the development of the learning outcomes for the qualification and for providing a technical sign-off point for the assessment content. The GPhC were not involved in the development of the assessment criteria, grading criteria or mark schemes.

'The grading scheme is based on a percentage pass mark of 60% per unit with no definition of essential successful outcomes within each. Discussions with GPhC and the awarding organisation highlighted that this is an area they are going to discuss in more detail moving forward and address in any subsequent assessment development, possibly to address through the inclusion of the definition of essential versus desirable outcomes. It is noted that with reference to the purpose for this qualification, any concern raised about what might be considered a relatively low pass mark could possibly be offset by the requirement for the combination of successful achievement and certification of this qualification in conjunction with the related NVQ, to warrant candidates being registered as fit to practise.

'With regards to stakeholder involvement and the technical expertise involved and monitored with the development process, the awarding organisation has clearly defined policy and protocols in place to manage and document this process. However, due to the succession of personnel it was not possible to provide evidence of this from the previous development cycle. Discussion with the GPhC confirmed that they were intrinsically involved in the detailed development of the learning outcomes and provided technical sector expertise to support

this. The learning outcomes and assessment criteria were developed jointly by the awarding organisations (City & Guilds and Edexcel), using sector and assessment experts (e.g. external verifiers for this sector). Each awarding organisation developed their own assessment method and instrument.

Strengths

- *'The design of the qualification and the evidence provided clearly supports the description of this qualification as an integral part of the licence to practise requirements for pharmacy technicians. This is further supported by the essential nature of the stakeholder endorsement by the regulatory body, the GPhC and the adherence to the content requirements predetermined and confirmed by the GPhC. The submission for GPhC was detailed and comprehensive and the assessments were scrutinised by them in the build-up to the launch of assessments.*
- *'In adhering to the GPhC standards the content of the qualification includes significantly greater detail within the knowledge requirements than is included within the related NOS.*
- *'Units are each assessed through short answer and context specific, task based questions suited to the knowledge-based nature of the qualification and the requirement for the application of skills in a specific context.*

Weaknesses

- *'There appears to be no process in place to support statistical/in-depth monitoring of performance of the assessment tools. The awarding organisation's intended review may help address this point.*
- *'At present feedback on the assessment and the content of the qualification is reliant on the centre and external verifier system which may present a weakness in terms of for example, item monitoring and standardisation over time etc. Again, the awarding organisation's intended review may help address this point.*
- *'The choice of a single predetermined pass mark for multiple assessments would appear to be a concern. This has been acknowledged by the awarding organisation, and will be reviewed.'*

9.3 Level 1 Functional skills: English

Awarding organisation: NCFE

Overview

"The functional skills qualifications in English are designed to assess the practical skills that allow people to use English in real life contexts. The qualification has a set of qualification criteria and subject criteria which define in some detail many aspects of the required assessments. These include the skills to be assessed and within them the skill standards (essentially the learning outcomes) coverage and range, whether the assessment should be internally or externally set, assessed and/or marked, and where internal, the nature of the controls. In addition, the weighting of some of the skills standards and the proportion of open to closed-response questions are set out. Each awarding organisation (AO) wishing to offer any functional skills qualification had to submit their proposed qualification together with a detailed set of sample assessment materials and guidance documents, covering both the externally assessed elements and those using controlled assessment, to the regulator. These are then subject to close scrutiny by subject experts appointed by the regulator prior to accreditation.

This has important implications for some aspects of the validity argument, which will be explored further in the report below. For example, a good deal of the validity argument must arise from the regulatory documents and the accompanying accreditation process. Clearly, an AO with a qualification developed under this regime has a responsibility to ensure that their assessments stay in line with those accredited (i.e. they are parallel assessments) and to provide evidence about how the qualification is functioning. But, in essence, issues associated with the case for the qualification, its development and general structure, the content standards, the assessment methods, and even the nature of the assessment tools used, all lie with the regulator, or at least the regulatory process. For example, assessment is compensatory within a skill (i.e. attainment in reading depends on overall performance in the reading test) but there is no compensation across skills (i.e. candidates have to pass all three skills to gain the overall award).

Strengths

- *The regulatory documents and accreditation process impart some important strengths to the qualification. In particular, not only has the process supplied a validation test of the sample assessment materials, but the need to be very clear as to how they comply with the regulations means that they can be easily mapped to the*

various criteria requirements, meaning that in turn subsequent live tests can also be so mapped. In the case of this qualification, each assessment is checked against an assessment grid, which is entirely derived from the criteria requirements and covers the various skill standards, the requirements for coverage and range, the relative weightings defined for the individual assessments, and the balance of closed to open response questions.

- *The processes by which the tests and controlled assessments are set is explicit and documented, making it much easier to evaluate in terms of its support of valid and reliable assessments.*

Weaknesses

- *The AO has considerable experience in running vocational qualifications but it is relatively new to managing assessment processes such as those required for functional skills. It has therefore been a steep learning curve for them in terms of the sort of procedures and approaches that best support effective assessment. This is particularly visible in the evaluation of the assessments. There appears to be no process in place to support statistical/in depth monitoring of performance of the assessment tools."*

9.4 Level 3 Certificate in Principles of Customer Service

Awarding organisation: NCFE

Summary

"This qualification is primarily designed for learners working in a customer service related role who are looking to further develop their knowledge of customer service. There is a clear and justifiable rationale set out for the qualification. The original development of the qualification was undertaken working with the Institute of Customer Service (ICS), and the final qualification specification was supported by Skills CFA (July 2011).

Assessment is by an internally assessed and externally moderated portfolio. The units which make up the qualification are based on the relevant NOS. Each unit specifies the required learning outcomes and the associated assessment criteria. Units are pass/fail. The

Qualification Specification details acceptable forms of evidence mapped against the assessment criteria. Assessors and internal moderators confirm that the evidence provided by learners is appropriate to satisfy each of the assessment criteria and therefore demonstrates mastery of the associated learning outcomes. Learners must achieve all three mandatory units to achieve the qualification. Centres are required to carry out internal moderation and standardisation, and external (by the awarding organisation) standardisation of assessment is achieved by (a) moderation visits by experience and well qualified moderators, (b) running standardisation events for moderators, and (c) the provision of comprehensive guidance and support to centres on issues of assessment and moderation.

The qualification and related assessment methodology would appear to be in line with accepted best practice for portfolio based assessment within vocational qualifications. The links to the national occupational standards (NOS) and the involvement of suitably qualified personnel in the assessment and moderation processes supports the view that the assessments are generally valid. However there are gaps in the evidence which is available to support a comprehensive validity argument, in particular with regard to reliability.

Strengths

- *Links with the NOS support the argument for the validity of the qualification.*
- *The assessment criteria purposefully reflect the required standards and provide a clear link between standards and assessment.*
- *The qualification specification lists the kinds of evidence which are likely to be acceptable, and maps these against the relevant assessment criteria.*
- *There are clear requirements on the qualifications and experience required of the delivery, assessment and moderation personnel.*
- *The method of assessment reflects the nature of the teaching and learning process for this work-related qualification.*
- *There are formal and documented procedures in place to monitor implementation of the required procedures for assessment and standardisation.*
- *The fact that there is only a single cut-score for each unit, and the fact that learners must provide a range of evidence to support the assessment of each learning outcome, can be expected to contribute to reliable assessment.*

Weaknesses

- *The mapping to the NOS is not obvious from the supplied documentation.*
- *In line with the other case study qualifications, there is a lack of documentary evidence to support a validity argument against particular aspects of the Ofqual framework. As with the other case study qualifications, this is particularly true of reliability, where the AOs approach to ensuring reliability (which is in line with industry best practice) does not currently provide the kind of data required to calculate reliability estimates.*
- *NCFE follow the same best practice as all other AO's with regards to portfolio based assessment. Ultimately, based on the nationally accepted approach to assessment of accredited qualification which operates on the basis that personnel must have the nationally accredited Assessor qualifications required to assess regulated qualifications, there is a strong reliance on the subject expertise and professional judgement of the assessor (moderated by the awarding organisation). This is not necessarily a weakness (arguably for many aspects of validity it is a strength) but it does raise questions about reliability."*

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 2014

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