A validation framework for work-based observational assessment in vocational qualifications

by

Milja Curcin, Andrew Boyle, Tom May and Zeeshan Rahman

The City and Guilds of London Institute

Ofqual/14/5374 **February 2014**

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

Acknowledgements

Many people contributed to this research and we believe it is better for it.

Firstly, we would like to thank all the people who were kind enough to share their views with us in interviews, and to let us into their salons, workshops and construction sites to observe them. Equally, we are grateful to all our questionnaire respondents, for taking the time to respond, and reply at length to open questions.

Without the help of our Portfolio colleagues, Jacky Jones, Diane Mitchell and Salim Visram, we would not have been able to organise and carry out our fieldwork in such a short space of time. Jacky and Diane – thanks also for all your insights and bearing with interminable questions. We are also very grateful to Mark Duerden and Swapnil Bhiker of Learning Assistant for providing e-portfolio data, and to Jenny Kwon from Business Intelligence for all her help with setting up the questionnaire. And thanks, Shabana Haroon, for helping out with transcribing those interviews.

Finally, we would like to thank the members of City and Guilds Assessment Research Technical Advisory Group (ARTAG) for spending the day with us giving us very useful comments on an earlier version of this report and framework. And to our colleagues, Inga Fitzgerald and Sharon Frazer, who also commented on previous versions of the report and framework, and patiently explained the various intricacies of competence-based VQs and their assessments. A big thanks is also due to Andrew Newman, Stuart Copus and Joel Bild for proofreading the final version of the report at top speed.

Any errors and misconceptions in this report remain our own.

Table of Contents

List of tables

List of figures

Executive summary

This document reports on a research project conducted by the City and Guilds of London Institute between March and December 2012 under tender number 154 let by the Office of Qualifications and Examinations Regulation (Ofqual). The project investigated the context, features, and underlying assumptions of Work-based Observational Assessment (WBOA); an assessment method that is widely used in competence-based Vocational Qualifications (VQs) in the United Kingdom, with an aim of producing a framework for its validation.

While working towards developing the validation framework, and in line with our promise to our sponsor, this research has provided the following **five key contributions**, highlighted in bold below:

1. The first review section attempts to quantify the preponderance of WBOA in accredited VQs, and describes and discusses the operational features of WBOA, and its status in the wider context of VQ assessment, performance assessment and competence-based assessment. The process of its quality assurance/verification is also briefly described and discussed. This description and discussion has enabled us to produce **a definition of WBOA**, as follows:

   WBOA can be defined as a real-life task-centred performance assessment method, where an assessor observes a candidate performing practical tasks and other activities in the workplace, and/or inspects the artefacts produced, in order to make a dichotomous judgement (achieved/not yet achieved) against assessment criteria about aspects of a candidate's competence.

2. Where appropriate, the initial review section also highlights possible threats to validity of WBOA and its results, associated with specific operational features or assumptions of the method. Together with the insights we have derived from the review of the argument-based approach to validation and several different validation frameworks, this has enabled us to produce **the interpretive argument (i.e. framework) for validation of WBOA**.
3. Empirical work to evaluate aspects of the framework (particularly the assumptions and threats to validity associated with WBOA) included stakeholder opinion collection via a questionnaire and structured interviews, and the subsequent analysis of these data. This empirical work **demonstrated the usefulness of consulting stakeholder views (and structuring the relevant instruments based on the proposed framework) as a method of gathering validity evidence.**

4. Assessor and verifier judgements from WBOA (gathered from a previous reliability study and via e-portfolios) were analysed and **a range of statistical indices were derived to show how they might feasibly provide insight into WBOA and quantitative evidence of its validity.**
5. The validation framework was amended in the light of the abovementioned empirical work, and **potential data collection methods that might be used to collect evidence of validity of WBOA were suggested in the final version of the framework.**

We believe that we have taken a useful first step towards describing and defining WBOA in terms that will be understandable to an audience wider than just VQ practitioners. We also hope that the approach we have taken of emphasising that WBOA is an assessment method that needs to comply with certain general assessment and validity principles has helped to derive a validation framework that is comprehensive and general enough, yet does justice to some of the specific properties of WBOA.

A number of issues that require further research have emerged from this study, the most important of which are outlined in our recommendations below. We hope that the current increased focus on VQs, both from the regulator and more generally, will help to drive these research strands forward for the benefit of the VET sector.

Our **recommendations** are as follows:

1. Although we believe that the proposed framework is sound and fairly comprehensive, we strongly recommend that it is fully piloted on a representative sample of competence-based VQs.
2. We also recommend that any pilots of the framework are carried out with a view to prioritising the threats which exist to validity, which will help in operational validation.
3. One possible approach to implementing some aspects of the framework into VQ awarding organisation practice might be to use it in structuring and focusing the operational verification processes (which, in any case, could be seen as continuous internal validation exercises) to reflect the framework as far as this is practically feasible.

Our other recommendations, outlined below, reflect our current view of the main validation questions (all based on the proposed framework) that future pilots should seek to answer based on validation exercises in individual qualifications. Some of our high-level recommendations (which probably go beyond WBOA-related issues) would require cross-awarding organisation cooperation and agreement from relevant Sector Skills Councils (SSCs) and funding bodies to investigate and potentially implement, e.g., combined workplace/simulated assessment in competence-based VQs. The regulator might wish to promote and facilitate debate and further research into such issues.

4. Review of constructs:
    - Design exercises to investigate whether current Learning Outcomes and Assessment Criteria assessed by WBOA appropriately reflect what practitioners actually assess and believe are important aspects of competence.
    - Investigate the possibility of expressing the construct as underlying knowledge/skills/abilities rather than lists of tasks where this is possible.
5. Review of available work-based assessment methods using insights from other fields where they are widely used (e.g., medical education):
    - Based on this review, develop criteria to judge whether the current form of WBOA used in VQs is the most appropriate for the constructs (that should be) assessed.
    - Investigate the potential of different 'versions' of WBOA (or different assessment methods altogether) to enable better targeting of assessment (e.g., those focusing on procedural skills vs. those focusing on integration of skills and less easily definable aspects of competence such as reasoning, professionalism, communication, and integration in the workplace).
    - Translate insights from these investigations into coherent database categories in order to ensure that WBOA (or its versions) is recognised and recorded as an assessment method there.
6. Review of 'orthodoxies':
    - Notwithstanding the overall support for using different forms of workplace assessment – is workplace assessment always the most appropriate? Could some carefully designed and targeted simulations not contribute useful additional evidence of competence (and mitigate negative washback)? A genuine, real-world task may not always be authentic, i.e. might sometimes not allow the assessors to see the abilities/skills that they are interested in.
    - Is observational assessment really unproblematic in competence-based VQs – or, should we be taking a hard look at levels of standardisation and agreement between assessors/IQAs/QCs and designing methods to promote and monitor this more effectively?
    - Does the final binary judgement of whether someone is competent or not yet competent really preclude the use of meaningful global scales (or comparative judgement techniques) to help standardise or help investigate the consistency of assessor judgements?
    - It is possible (though not straightforward) to produce appropriate and useful statistical indicators of WBOA quality – relevant research in this area should be continued. This might become easier if standardised e-portfolio forms were used across centres, and if the forms were designed in a way that might support relevant data collection and analysis.
    - Review outcomes-based funding arrangements insofar as these present threats to the integrity of WBOA practices and decisions (cf. recommendations by the Richard Review).

Introduction

The context for this work

This study exists in a context. Firstly, vocational qualifications (VQs) have traditionally lacked a basis in research to compare with general qualifications. For example, some leading academics in the VQ field have said:

> *… the research-based literature on VQs is thin, and … the UK lacks a substantive evidence base on VQs …* (Unwin, Fuller, Turbin, & Young, 2004, p. 3)

Further, VQs have long been the object of major government initiatives; often these are introduced at pace, judged according to political rather than technical yardsticks and can (unsurprisingly) appear to not be successful (Stone, 2012). Recent government reports have challenged the worth of many VQs (Wolf, 2011), and have proposed amendments to apprenticeships (Richard, 2012).

In this light, this research should be seen along with preceding work sponsored by Ofqual (Harth & Hemker, 2012; Boyle & Rahman, 2012; Johnson, Johnson, Miller, & Boyle, 2012) as part of a programme of research[1]. The aim of this programme is to give fair-minded observers and commentators a basis from which to develop informed opinions about the worth (or otherwise) of VQs.

This context, orientation and superordinate aim has several consequences; firstly, as all the reliability researchers cited above found, the absence of prior research requires a substantial initial step of 'scything through dense undergrowth' in order to establish first principles. This is a strength – starting from first principles is never a bad thing, but also brings challenges. It tends to make for long and convoluted research reports that seek to synthesise diverse technical and theoretical literatures and apply them to a context in which they have not been applied before. Secondly, we are dealing with prior assumptions – both those of the authors and those of the readers. Many of the concepts in assessment and psychometric literature assume external written tests, even if they purport to eschew such limitation. Further, many readers will assume that an assessment method such as work-based observational assessment (henceforth, WBOA) is 'bound' to be less rigorous than alternatives such as external assessment (Mansell, 2011).

The desire to establish a research base in a field where one has been lacking, the suspicion that some readers will have unfair preconceptions about WBOA and perhaps even the authors' status as employees of an awarding organisation lead to the moral hazard of seeking to 'go easy' on WBOA. We hope we have

[1] The other VQ validity report in the current round of tenders (AlphaPlus Consultancy Ltd, forthcoming) should also be seen as part of this programme of work.

avoided this hazard; we seek to build knowledge and be balanced. To criticise where it is merited, but in doing so to help to take understanding of vocational assessment, and here in particular of WBOA, forward and thus improve confidence in qualifications that employ it.

Purpose of the study

The purpose of this study was to propose a framework for validating one of the common assessment methods in VQs – that of WBOA. WBOA was introduced into VQs as part of the competency movement and competence-based assessment (Wolf, 1995) in the nineties. One notable characteristic of competence-based assessment is that the role of assessment methods and methodologies, and the validation of procedures for eliciting performances from candidates is neglected (McNamara, 2006, p. 38), assessment is seen as 'unproblematic' (Wolf, 1995), and there is little appreciation for the impact of assessment method on performance and outcomes although a variety of studies demonstrate that this can be enormous (McNamara, 2006, p. 38).

In this study, we aim to refocus attention on various implications that the features and procedures of WBOA as an assessment method used in VQ assessment might have for the validity of its outcomes, by developing a framework for its validation. We endeavour to situate this method in the wider context of competence-based VQ assessment as well as performance assessment. We discuss its various complexities in view of its association with these assessment contexts and traditions in order to increase understanding of its features and assumptions, as well as derive a definition of the method. Ultimately, all this will inform the proposed validation framework.

This study provides:

- a description of the preponderance and distribution of WBOA in vocational qualifications based on data available in City and Guilds databases
- a validation framework that is informed by:
    - a detailed review of key concepts and properties of WBOA based on the relevant literature and discussions with City and Guilds operational staff, leading to a definition of WBOA
    - the views of some of the key practitioners and stakeholders, collected through a questionnaire and interviews
- proposals for some key statistical indicators as quantitative evidence for validity (or lack thereof) of WBOA
- suggestions for possible evidence-collection methods to be used in validation of WBOA
- some recommendations.

We hope that the proposed framework will help to inform Ofqual's methodologies for regulating VQs in this domain. We also hope that the framework will help practitioners to engage in collecting appropriate evidence to investigate the WBOA's validity, and contribute to its improvement where required.

Validity and validation in brief

Validity is one of the most important properties of assessments and needs to be present in sufficient degree if their quality and fitness for purpose is to be guaranteed by awarding organisations. A view of validity as 'an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment' (Messick, 1989, p. 13) is now widely accepted, though not unchallenged in the educational assessment community (cf. Borsboom, Cramer, Kievit, Zand Scholten, & Franic, 2009; Lissitz & Samuelsen, 2007). This view subsumes previously proposed models of validity (criterion validity, content validity) that failed to account for the complexity and the multi-faceted nature of this concept.

However, with Pollitt (2009) (cf. Borsboom, 2005), we argue that, although, of course, the uses of assessment results need to be legitimate and appropriate, the 'result ingredients' (e.g., the appropriate construct definition, quality of test items or tasks, appropriate difficulty of test tasks, level of standardisation of assessors, etc.) all need to be of sufficient quality to ensure that the ultimate interpretation of qualification/assessment results is useful and valid. As Pollitt states:

> *…validity is a continuous quantity, a property of the assessment process, maximised at the beginning – when the test is conceived – and lost to some degree at every step along the way.*
> Pollitt (ibid., p. 4)

This view of validity can be termed 'intrinsic' (Pollitt & Ahmed, 1999, p. 1) in that 'validity consists in managing the assessment procedures in such a way that candidates' mental activities during the test will correspond as closely as possible to the mental activities of a person engaged in real life use of the knowledge or subject being assessed' (cf. Bachman, 1990 and McNamara, 1996). According to this view, the primary responsibility for validity is with the people who design assessments. 'If they put garbage in, no one further down the line can deliver anything other than garbage out' (Pollitt, 2009, p. 4).

Currently, we do not think that the comprehensive view of validity, with its focus on score/result interpretations and the intrinsic view, with its focus on the assessment process, are mutually exclusive. Rather, we believe that they are both necessary and complementary.

Both the comprehensive and the intrinsic models of validity assume that construct validity is the fundamental aspect of validity, and construct validation the basis of the validation process. Originally, the construct model was proposed by Cronbach & Meehl (1955) to be used 'whenever a test is to be interpreted as a measure of some attribute or quality which is not operationally defined' (ibid., p. 282), and 'for which there is no adequate criterion' (ibid., p. 299). Cronbach & Mehl defined constructs as 'some postulated attribute of people assumed to be reflected in test performance' (1955, p. 283). More recently, a definition of construct as 'the concept or characteristic that a test is designed to measure' (AERA, APA, & NCME, 1999, p. 5) has become widely cited. Bachman (1990, p. 255) cites Messick, who states:

> *A measure estimates how much of something an individual displays or possesses. The basic question [of construct validation] is, What is the nature of that something? (Messick, 1975, p. 957)*

In other words, what is the nature and definition of the relevant construct? In the work of (Messick, 1989), construct validation became synonymous with a comprehensive model that encompasses different kinds of validity evidence and other kinds of validity models that can all contribute to validation of assessments. Messick emphasises that:

> *…by making construct theories of the performance domain and of its key attributes more explicit … test construction and validation become more rational, and the supportive evidence sought becomes more attuned to the inferences made.* (Messick, 1989, p. 64)

The process of validation can be seen to involve 'developing a scientifically sound validity argument to support the intended interpretation of test scores and their relevance to the proposed use' (AERA, APA, & NCME, 1999, p. 5). This reflects the modern view of validation as an integrated process that draws on multiple sources of evidence (Cronbach, 1971) and essentially represents a process of hypothesis and theory testing using all of the appropriate experimental, statistical, and philosophical means that are normally used in scientific enquiry (Messick, 1989). It requires both professional judgement and empirical research (Bachman, 1990, p. 49).

Among a number of approaches to validation that have been proposed over time, the argument-based approach proposed by Kane (2006) (see also Kane, Crooks, & Cohen, 1999) has been widely accepted as the one that provides the most detailed practical guidance for validation, as well as guidance on prioritising validity evidence of different types, and is based on the current theory of validity (see, e.g., Chapelle, Enright, & Jamieson, 2010; Wools, Eggen, & Sanders, 2010; Shaw, Crisp, & Johnson, 2012) for recent implementation). It implements Messick's view of validation as scientific enquiry and proposes three basic methodological principles of validation, i.e. providing an explicit statement of the proposed interpretation

of the score, an extended analysis in validation, and consideration of alternate interpretations (Kane, 2006, p. 22).

This approach to validation was suggested by Ofqual in their invitation to tender document (Office of Qualifications and Examinations Regulation (Ofqual), 2011) and we considered it to be appropriate as the starting point for the development of the validation framework proposed in this study. We provide more details about this approach in the relevant section at p. 38.

This study's focus:

Although we agree with the abovementioned view that validity is a property of the assessment process, it is important to point out that the context of VQ work-based assessment, not very much of the assessment is designed, to the extent that certain views do not consider WBOA to be an assessment method at all (cf. the discussion in the previous section). Thus, it might not be easy to conceptualise, among other things, what 'intrinsic' might refer to in that context, or even who the 'test developer' might be. However, despite these difficulties, one of the points that this study will attempt to make is that it is important, for validation but also other purposes, to think of WBOA as an assessment method that should have an appropriate level of intrinsic validity, and to approach its validation from that point of view.

In this study, we are essentially focusing on the validation of a specific assessment method – that of WBOA – and its processes and properties. However, the interpretations of its outcomes contribute to forming a picture about a candidate's competence and ultimately contribute to the final judgement of whether a candidate deserves a particular qualification. Therefore, if WBOA properties, processes and outcomes are invalid ('garbage in'), the outcomes of the whole qualifications in which it is used as an assessment method might be deemed invalid ('garbage out'), especially where WBOA forms a large part of the assessment for that qualification or covers crucial elements of the qualification.

Report structure

This report begins with a Review section in which we describe and discuss the status and properties of WBOA in VQ assessment. Here we describe and discuss (the challenges about estimating) the preponderance of WBOA in VQ assessment and how WBOA fits into the wider space of VQ assessment. We then go on to discuss some of the key concepts and properties of WBOA that we consider relevant for its validation – highlighting its status as a performance assessment method, and issues around the relevant construct definition, operational procedures and quantification of observations that are used in this method. We also briefly discuss verification as the quality assurance process used for WBOA, and associated threats to validity. In the summary section at the end of the Review, we provide a definition for WBOA.

We then provide some background for the validation framework – we briefly describe the argument-based approach to validation and a generic validity argument for performance assessment, and we review some existing argument-based validation frameworks.

The information presented and discussed in these two major sections is then synthesised in our proposal for the interpretive argument for validation of WBOA (at pp. 49ff), which essentially represents the proposed framework for validation.

Following that, we present the results of our fieldwork, in which we attempted to evaluate our proposed framework and to deepen our understanding of how WBOA functions in practice by seeking the views of some relevant stakeholders about its key properties and assumptions. The collection of these views was guided by a previous version of the interpretive argument for WBOA.

The section on quantitative evidence of validity in WBOA (p. 87) presents our attempt to derive some statistical indicators that might provide evidence of validity (or lack thereof) in this assessment method.

The next section contains the final WBOA validation framework which includes the interpretive argument accompanied by the validity argument with some proposals for possible validation/evidence collection methods.

The report finishes with the Discussion, Conclusions and Recommendations sections.

Review of the status and properties of WBOA in VQ assessment: a description and discussion

Preponderance of WBOA in VQ assessment

One of the first steps that we undertook when embarking on this study was to estimate the preponderance of WBOA in VQ assessment. We quickly found that a category such as WBOA, observation, or similar did not exist on the databases that normally collate information about assessment methods used in different VQ units.

Initially, we explored the possibility of using Ofqual's Regulatory IT System (RITS) to establish the preponderance of WBOA in VQ assessment. RITS is a database of active and inactive qualifications and units accredited by Ofqual (see: http://www.ofqual.gov.uk/rits). It uses around ten categories to classify types of assessment:

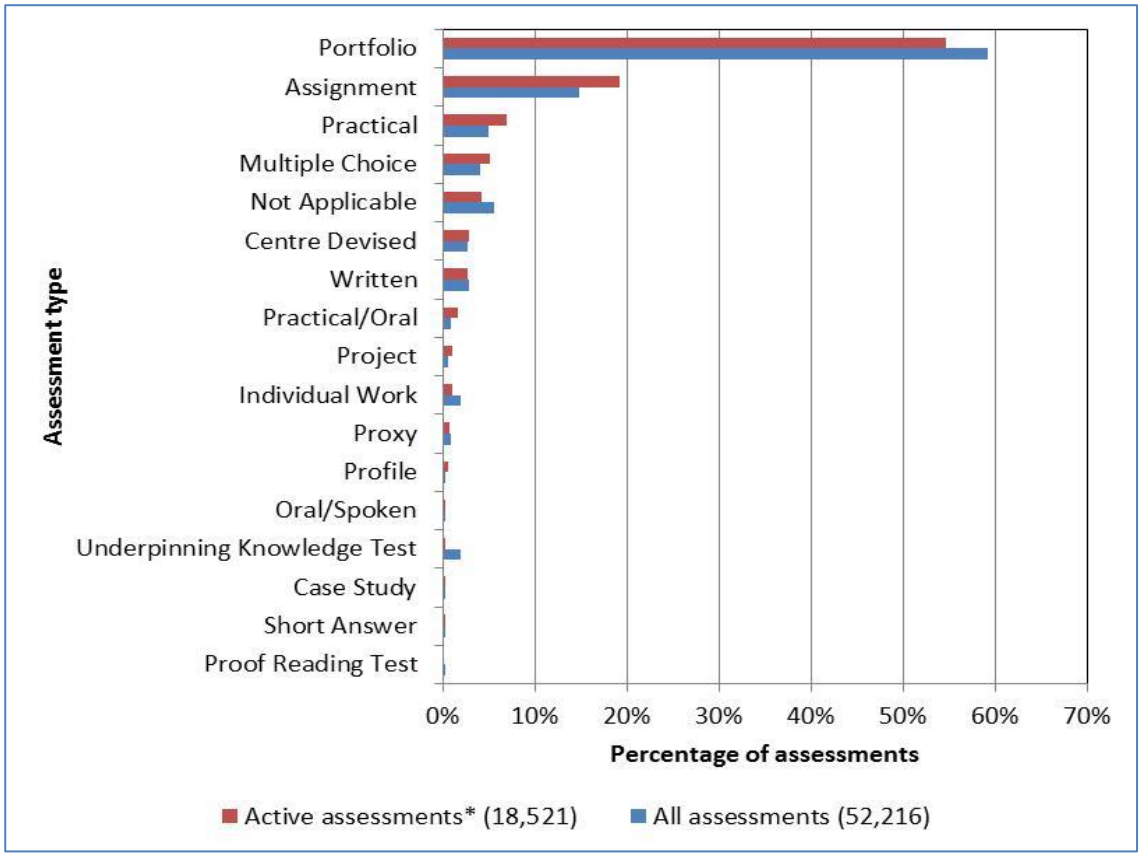
- Aural Examination
- Coursework
- e-assessment
- Multiple Choice Examination
- Oral Examination
- Portfolio of Evidence
- Practical Demonstration/Assignment
- Practical Examination
- Task based Controlled Assignment
- Written Examination

Apart from other ambiguities in this database (e.g., e-assessment being offered as an assessment method, when this could refer to a variety of different methods), none of the categories unambiguously relate to WBOA. Of the above categories, 'portfolio of evidence', 'practical demonstration/assignment', 'practical examination' and 'task based controlled assignment' are most likely to be using an observational method of assessment, however, this could not be confirmed. In addition, RITS allows multiple assessment types to be attached to each unit, since awarding organisations typically do not commit to a single assessment method for each unit at the point of accreditation (and sometimes multiple methods are indeed used), and there is missing information about assessment type, making it impossible to identify with any certainty the units using WBOA in this database.

City and Guilds' SAP database contains detailed information on active and inactive, accredited and non-accredited qualifications. The classification of assessment types in SAP differs slightly from RITS, but, again, does not include a category to unambiguously denote WBOA (or indeed, other methods or evidence types that are typically used in the workplace, e.g., professional discussion, witness statement, etc., which all typically contribute to a portfolio of evidence – see the next section and Appendix 1 for more details).

Figure 1 shows the breakdown of all City and Guilds assessments in the SAP system in November 2011 by assessment type. This illustrates that over half of all assessments are referred to as portfolio, although, a portfolio is a repository of information, rather than a method of assessment (Stone, 2012). As repositories of information, among other things portfolios contain records of results from other assessment methods such as multiple-choice (often underpinning knowledge) tests and short answer questions, and they tend to be widely used in work-based qualifications. It is, therefore, our best guess that portfolios also contain records of WBOA, and that the units that use portfolio of evidence presumably often also use WBOA. This was the case with the qualifications that featured in our fieldwork (City and Guilds Hairdressing, Electrical Installation and Plumbing), and also in a few others for which we inspected the units to find out which methods of assessment are typically used.

Figure 1: All and active City and Guilds assessments by type



* More than five candidates completed assessment in any one year over past three years

Source: City and Guilds SAP database, November 2011

This lack of clear differentiation in City and Guilds databases between various work-based assessment methods and types of evidence is partly related to the fact that differentiating between them is not necessary for standard operations of receiving results and for certification, for which these databases are mainly used, as was confirmed to us by the operational staff. In addition, having reviewed recommended assessment methods for different units in several qualifications, we found that the relevant assessment strategy (developed by the SSCs) often leaves it up to the assessor to decide whether WBOA will be used to collect evidence of competence, or one of other evidence-gathering methods such as expert/non-expert witness statements, professional discussion or simulation.[2]

Given the unavoidable vagaries of workplace situations and conditions, this flexibility might be useful in some situations. However, it might detract from the reliability and validity of assessment in the workplace, as there sometimes does not seem to be clear guidance (or agreement?), about the skills and tasks that should be assessed by a certain method, which should be supported by relevant arguments as to why a particular method might be superior to others in specific situations.

Given this optionality in choosing the methods for assessing 'work-based' units, the lack of clear categorisation in awarding organisation and regulator databases is perhaps unsurprising. However, we believe that it would be beneficial, and certainly helpful in validation and in research more generally, if electronic databases included information of this type, even if it is not used in everyday awarding organisation operations.

In the next section, we begin with the overview of key properties and concepts relevant to WBOA, in order to try and understand its status and use in VQ assessment, and to attempt to define it in standard measurement terms as far as this is possible.

Situating WBOA in the assessment process of competence-based vocational qualifications

The process by which it is established whether a candidate is (occupationally) competent and deserves a particular competence-based VQ, whether these are based on the legacy NVQ[3] or the current QCF[4] framework, is perhaps more akin to evaluation than measurement. By evaluation we mean 'the systematic

[2] There are also cases where the relevant assessment strategy clearly specifies which method of assessment should be used, and whether WBOA may be substituted by another method – e.g., in Hairdressing and Plumbing.

3 See Ofqual website for a brief description: http://www2.ofqual.gov.uk/qualifications-assessments/89-articles/18-nvqs

[4] A brief description of the QCF is provided by Stasz (2011, pp. 3-4). More information is available on the Ofqual website: http://www.ofqual.gov.uk/qualifications-assessments/89-articles/145-explaining-the-qualifications-and-credit-framework

gathering of information for the purpose of making decisions' (Weiss, 1972). In general, and also in the case of competence-based VQs, the information is not only quantitative (e.g., test scores), but can also include verbal and written descriptions and testimonials, performance profiles, as well as overall impressions, which can all contribute to providing important information for evaluating individuals (cf. Bachman, 1990).

Competence-based VQs are typically formed of compulsory and optional units at appropriate level, each of which is supposed to contribute to the assessment and gathering of evidence of different elements of occupational competence, expressed through Learning Outcomes (LOs) and Assessment Criteria (ACs). It is common practice for a variety of assessment methods (e.g., written tests of various kinds, professional discussion, WBOA) and other pieces/sources of information (witness statements, photos, videos, etc.) to be used across and within units to make a decision on whether the candidate is competent or not.[5] All this evidence is typically organised into a portfolio owned by the candidate (see e.g., Johnson, 2008; Wolf, 1998), which documents the progress through the qualification and serves as the basis for reaching the decision about a candidate's overall competence at the end of the qualification.

Suitably trained assessors are predominantly external but may also be work-based (e.g., supervisors, line managers, etc.), although in the original idea and proposals for NVQs the opposite was supposed to be the case (Jessup, 1991).[6] In several parts of this report as well as where appropriate in the framework, we highlighted where this might have implications for the validity of WBOA.

Different assessors may take part in the assessment process for the same candidate, depending on the assessment method or units in question, as well as other factors. Candidates are generally assessed when ready, which might be established by candidates themselves, or their assessor, tutor or employer. Following each assessment occasion, whether they meet the required criteria and pass or not, the candidates should be given relevant feedback. In situations when they do not meet the criteria, the summative assessment effectively becomes formative, and serves as the basis for feedback and a learning opportunity prior to the next assessment occasion. Once all the criteria have been met, and all the relevant units achieved, the candidate will gain certification for the qualification which indicates that they are competent in the relevant occupational domain. The assessment process is verified by Internal Quality

[5] See Appendix 2 for a summary and brief description of some assessment methods that might contribute to a portfolio of evidence of a candidate's competence – the list is not exhaustive.

[6] External assessors are also known as 'peripatetic assessors', or visiting assessors. They assess learners when there are insufficient numbers of qualified assessors in the learners' workplace, or where learners work in isolated, very small or dispersed settings; training agencies place and support learners in workplace settings; or where workload of work-based assessors could impact on assessment. They sometimes support work-based assessors, who directly observe learner performance, with planning, reviewing, completing documentation etc. (City & Guilds, 2011c). According to a recent survey of employers involved in apprenticeships, only one per cent of employers are involved in assessment (IFF Research and the Institute of Employment Research (IER), 2012) .

Assurers (IQAs), appointed by centres,[7] and Qualification Consultants (QCs),[8] appointed by awarding organisations – see section on Verification later in the report for more details.

As Bachman (1990, p. 22) points out, the probability of making the correct decision in evaluation is a function of both the ability of the decision maker and of the quality of the information upon which the decision is based. Thus, the information that is gathered should be reliable and relevant. WBOA is the assessment method that often provides key information that supports the evaluation of VQ candidates in terms of whether they are competent or not yet competent in a particular occupational domain. It involves candidates being observed by assessors while performing a range of relevant tasks that naturally occur in their workplace, as well as, where relevant, inspection of artefacts produced while performing the tasks.

WBOA was widely introduced in the assessment of NVQs in the nineties (see e.g., Jessup, 1991; Wolf, 1995). The key architect of the NVQ system, Gilbert Jessup argued that: 'Assessment of performance in the course of normal work offers the most natural form of evidence of competence and has several advantages, both technical and economic' (Jessup, 1991, p. 51). Although NVQs have recently been superseded by the QCF, WBOA has remained an important assessment method for those candidates that are already in the workplace[9] and taking qualifications where evidence of occupational competence is required for assessment (see next section for more details).

WBOA has a number of distinctive features compared to more traditional assessment procedures such as written tests, but also compared to other performance assessments (e.g., simulations of various activities), particularly with respect to the lack of standardisation of the tasks and assessment locations and conditions. It is 'grounded in the reality of the workplace with its own rules, norms, expectations and prohibitions' (Benett, 1993, p. 84). Some other distinctive features relate to broader issues such as the way competence, performance, and assessment outcomes are defined in VQs; how standardisation of assessors is implemented and ensured given their key role in the assessment process; the role of the candidates in the assessment process and their interaction with the assessor, who might or might not be a familiar person (e.g., manager or supervisor vs. external assessor); the issues with expressing some of WBOA's key properties in traditional psychometric terms; etc.

Overleaf we give two examples of possible WBOA situations that might occur in Hairdressing.

[7] A 'centre' is an organisation (such as a school, college, training provider or workplace) accountable to an awarding organisation for the assessment arrangements leading to a qualification (City & Guilds, 2011c).
[8] Until recently, IQAs and QCs used to be referred to as Internal Verifiers (IVs) and External Verifiers (EVs) respectively.
[9] *Workplace* can be defined as a place where people work, the majority of staff are already qualified and the main function is not training and assessment (City & Guilds, 2011b).

Candidate A

Candidate A is one of three full time apprentices at a busy city salon. One member of staff is the assessor. The manager acts as IQA.

Candidate A helps out all day around the salon observing the qualified hairdressers and chatting to clients. He gets training one afternoon a week on cutting techniques using artificial heads, and then on live clients (who are usually students from the local college) under supervision. The trainers give detailed explanations on processes and theories as they work, and the apprentices also attend college one evening a week to supplement their learning of knowledge and understanding.

Candidate A is to be assessed on hair cutting and has an appointment at a suitable time for him and his assessor. The client has long curly hair and this is not a type of cut he has been assessed on before, so this should provide all of the evidence necessary for this section of the unit. He has taken some tests at college to show the knowledge he has, and his assessor asks questions at appropriate times during the assessment to fill gaps.

The assessment goes well and he achieves all of the required assessment criteria for the cutting style, and fills all of the gaps for knowledge and understanding. His assessor will double check that all assessment criteria for the unit have been achieved and request certification for the unit. He only has two more units to complete for the qualification.

Candidate B

Candidate B is the only part time NVQ candidate employed part-time at a salon situated in a small town.

There is a visiting assessor who schedules visits 2-3 times a month to catch up and carry out assessment when appropriate.

Candidate B helps out around the salon between 10AM and 3PM (while her children are at school) supporting the stylist and senior stylist. She gets training on artificial heads when there are no customers booked and is able to start trying out the techniques under supervision on live clients who are often friends or local students.

Candidate B is to be assessed on hair cutting, and has arranged a time with her assessor who will be able to observe her from 10AM to 3PM on that day. Candidate B will be able to carry out the required cut, and can also demonstrate evidence for the salon reception unit. She is also prepared for professional discussion with her assessor where she expects to cover the majority of knowledge she needs to provide to supplement the evidence provided by the question paper she answered last month.

Unfortunately the client does not show up, and another regular client who is well known in the salon agrees to take their place. The replacement client wants a style which has already been observed, but where Candidate B was found not yet competent for the finishing aspect of the cut. Given that by this point Candidate B has mastered the finishing aspect, she is able to fill the gap left by the previous assessment, and will rearrange to be assessed for the remaining style. In addition she was able to fill all the remaining gaps for the knowledge and understanding through professional discussion. Candidate B hopes to complete the unit next time. She has a number of gaps to fill across the qualification and would like to get these filled as well as start some of the next unit.

These examples demonstrate several features of WBOA, for instance that assessment dates are scheduled for when the candidate's job is likely to provide suitable evidence, but the evidence that occurs can be affected by uncontrollable contextual issues (e.g., customer requests). The assessor observes the candidate at work, recording all of the ACs where sufficient evidence is observed that meets them. On each observation occasion all ACs that are evidenced can be signed off, from any unit, whether these were expected to arise on that assessment occasion or not. Similarly, there are many reasons why an AC may not be achieved on any specific occasion, leaving a gap in the record to be filled on another occasion. Also, if the candidate does not fully meet the demands of particular ACs in a specific practical task, the assessor may ask questions to elicit/test whether the candidate can meet the requirement, but has not been able to demonstrate it in practice.

It should be noted that the same qualifications (e.g., NVQs or their QCF equivalents) may sometimes be achieved either while working as an apprentice (or while being in paid or voluntary employment) in an actual workplace, or in a college, where training and assessment will be in a 'realistic working environment.'[10] In both cases, a certain proportion of assessment will be by observation, but only observational assessment in the former case would be classed as WBOA in a strict sense. Each of these situations presumably has its advantages and disadvantages both for training and assessment, and this will have implications on how comparable the inferences are that can be made about the candidates who obtained the same qualifications in these different settings.[11] Our focus in this study is the workplace setting, as the locus both of a significant part of training and of assessment.

In the next section, we discuss the status of WBOA in VQ assessment and more generally, as well as the associated key concepts, deriving a definition for WBOA. We argue that it would be beneficial to perceive and treat WBOA as a potentially rigorous assessment method, which would enable a better understanding of how and when it might be most effectively used to assess in the workplace, maximising the validity of inferences about candidates based on its outcomes.

[10] A *realistic working environment* (RWE) is defined as 'a model environment, having an acceptable level of appropriate equipment and operating continually to professional standards. It should provide the opportunity for candidate assessment under conditions approximating as closely as possible to the workplace, under the control of a qualified assessor' (City & Guilds, 2006a, p. 45).

[11] See a recent paper by Tim Oates (Oates, 2013), which highlights the need for clarity in the different 'tracks' in vocational education training and for clarifying the purpose of different qualifications.

The key concepts and properties of WBOA relevant to its validation

WBOA as a performance assessment method

WBOA is probably closest to the work sample type of performance test as defined by Fitzpatrick & Morrison (1971, p. 242). The work sample test normally employs an actual job situation for which the examinee is expected to have had the necessary training and experience, with many tests in commercial education, industrial arts, vehicle operator training, and similar settings being of this type.

According to these authors, a performance test more generally is one in which some criterion situation is simulated with more fidelity and comprehensiveness than in the usual paper-and-pencil test. Kane, Crooks, & Cohen (1999, p. 7) define performance assessments as those that involve a close similarity between the type of performance that is actually observed and the type of performance that is of interest. Performance assessments can cover process or products (Fitzpatrick & Morrison, 1971, p. 238), and this is also the case in WBOA.

Fitzpatrick & Morrison (ibid., p. 238) point out that there is no absolute distinction between performance test and other classes of test. Thus, although WBOA cannot be classed as a traditional standardised test, like other performance assessments it is still a measurement procedure that is supposed to help 'reach a conclusion about the candidate's ability to handle the demands of the criterion situation' in the workplace. In addition, it is also often used within qualifications that are assessed based on 'national' standards. As such, the conclusions based on this procedure should be replicable, reliable and valid as far as this is possible.

As already stated, WBOA is just one of a variety of methods and sources of information that provides evidence of competence for a candidate taking a competence-based vocational qualification. Nevertheless, as noted on p. 8, it is important to devote attention to its validation as the interpretations based on the results of WBOA can have important ramifications for the validity of the interpretation of overall qualification results.

There is a wide range of methods that can be classed as performance assessments (Lane & Stone, 2006; Fitzpatrick & Morrison, 1971; etc.). One way of classifying them is into task-centred and construct-centred approaches (Messick, 1994). The former are based on sampling tasks and performances from the relevant domain, with the focus on identifying the relevant types of performances and tasks that elicit them (often real-life tasks). Since every skill, no matter how task-specific, is relevant for task completion, the scoring/rating criteria that are developed tend to be task-specific (Messick, 1994). The real-life tasks and their associated criteria that are often employed are often assumed to ensure authenticity and validity in

these approaches (McNamara, 1996; Bachman, 1990; cf. Wolf, 1995). WBOA as implemented in English vocational qualifications could be classed as a task-centred approach.[12]

In contrast, the construct-centred approaches start from identifying the knowledge, skills and other attributes that enable both a given performance, and also a range of other performances engaging the same knowledge and skills. Here, the nature of the construct guides the selection and construction of relevant tasks as well as the development of construct-based scoring criteria (Messick, 1994) – see also construct-related discussion on p. 7.

There is a considerable literature that discusses performance assessment (though rarely workplace observation specifically) in terms of its origin and reasons for popularity, but particularly in terms of the implications of its (psychometric) properties to validity (e.g., Fitzpatrick & Morrison, 1971; Moss, 1992; Haertel, 1990; Linn, Baker, & Dunbar, 1991; etc.). Arguments that have been proposed in favour of performance assessments have clustered around its 'authenticity' and 'a greater relevance in determining the degree to which the examinee can actually perform the tasks of the criterion job or other situation' (Fitzpatrick & Morrison, 1971, p. 268). According to Frederiksen & Collins (1989), in performance assessment, the skills of interest are directly measured insofar as they are apparent in the performances or products that are elicited. Performance assessment thus can be seen as more direct than objective testing, which, though offering a number of practical advantages (including the potential to be used in large-scale assessment, high reliability, potential wide domain representation, etc.), tends to provide indirect and partial indicators of educational outcomes (Kane, Crooks, & Cohen, 1999, p. 5). This directness of performance assessment is also argued to ensure its fidelity (Fitzpatrick & Morrison, 1971), meaningfulness (Linn, Baker, & Dunbar, 1991) and authenticity (Wiggins, A true test: Toward more authentic and equitable assessment, 1989). Meaningfulness to students as well as assessors and teachers is considered likely to positively affect motivation and direct learning, thus leading to positive educational experience (Lane & Stone, 2006, p. 389). Directness and high fidelity arguably represent the main advantages of performance assessment over objective testing in certain contexts.

However, there are also criticisms of performance assessment. Here we highlight a few key concerns related to task-centred approaches, particularly those using real-life tasks, which we also see as potentially applicable to WBOA and relevant to consider in its validation.

[12] Note that the task-centred orientation is not necessarily an inherent feature of WBOA in general, but results from the interaction of the definition of competence in English VQs and the behaviourist approach to assessment (see section on Defining the construct of competence operationally at p. 24). However, since our framework is intended for validation of WBOA as currently used in English VQs, we consider the task-centred incarnation of WBOA.

A number of authors have argued that it is important to distinguish between what is 'genuine' and what is authentic in task-based real-life approaches. In the context of language testing, Widdowson (1978, cited in Pollitt & Ahmed, 2007, p. 202) argued that a reading text might be genuine without being authentic: 'Genuineness is a characteristic of the passage itself and is an absolute quality. Authenticity is a characteristic of the relationship between the passage and the reader and it has to do with appropriate response' (p. 80). Thus, authenticity can be argued to lie in 'the interaction between the test taker, the test task and the testing context' (Bachman, 1990, p. 322) and what this reveals about the relevant skills, competencies and/or abilities, rather than just the fact that the task, the criteria, and the context is the same as in real life. This is related to criticisms about task-centred approaches being particularly prone to construct-irrelevant variance, particularly with respect to scoring (Messick, 1994).

Another criticism has to do with the emphasis in these approaches on content-related and predictive validity (McNamara, 1996) as well as 'face validity' (Bachman, 1990), rather than construct validity. However, the validation approach based on construct validation that we assume here, and that has been widely recognised as the more useful one amongst measurement specialists, considers content and predictive validity as only contributing to a bigger validity picture. As Haertel (1992, cited in McNamara, 1996) points out, construct validity 'is relevant to performance measurement even if the intended test interpretations do not appear to involve psychological constructs', i.e. it is relevant in both task-centred and construct-centred approaches. Along the same lines, Messick argues that:

> *… performance assessments must be evaluated by the same validity criteria … as are other assessments. Indeed, such basic assessment issues such as validity, reliability, comparability, and fairness need to be uniformly addressed for all assessments because they are not just measurement principles, they are social values, that have meaning and force outside of measurement wherever evaluative judgements and decisions are made.* (Messick, 1994, p. 13)

In addition, problems with feasibility (especially in large-scale testing contexts) and problems with sampling and reliability have also been noted in relation to task-centred and other performance assessments by several authors (e.g., Messick, 1994; McNamara, 1996; Kane, Crooks, & Cohen, 1999; etc.). The issues of sampling are discussed in more detail in the section on the validity argument for performance assessment at pp. 40-42.

Traditionally, the development of tests, and, arguably, performance assessments too, is based on a logical sequence of procedures that links the putative ability, or construct, to the observed performance. Bachman (1990, p. 40) cites Thorndike and Hagen (1977), who state that this sequence includes three steps: (1) identifying and defining the construct theoretically; (2) defining the construct operationally, i.e. specifying the relevant testing procedure, and (3) establishing procedures for quantifying observations. It is our view

that as a performance assessment method, and a measurement procedure that is used as the basis for making decisions about individuals, WBOA should follow a recognised approach to its development and implementation as outlined above. This, in turn, should make it possible to approach its validation from construct definition, and allow for investigation of various operational facets that might have implications for its validity.

We use this sequence to structure the discussion in the rest of the review section, where these three steps of assessment development are discussed with respect to WBOA and its properties. In our discussion, we particularly focus on the theoretical and operational definition of the construct of competence, as this is central in competence-based VQs, has not been considered sufficiently with respect to WBOA, and should constitute an important aspect of validation efforts. Where relevant, we also highlight potential threats to validity associated with certain procedural and other properties of WBOA.

Defining the construct of competence theoretically

As noted earlier (p. 7), the basic question in construct validation is that of defining the construct that is being measured, in order for 'test construction and validation [to] become more rational' and to enable a clearer link between evidence that is required and the inferences that are made on the basis of measurement outcomes. In VQs, WBOA is typically used in those qualifications/units where the key construct is that of competence, together with the relevant criterion domain[13] for which this competence is relevant.

Jessup's definition of NVQs states that 'competence should incorporate specified standards in the ability to perform in a range of work-related activities, and the underpinning skills, knowledge and understanding required for performance in employment' (Jessup, 1991). The following paragraph provides further details about what it is that 'competence in a role' might require:

> *Anyone who has worked knows that there is far more to being successful in a job than carrying out the basic tasks competently. Jobs are seldom performed in isolation. One has to work with other people, often solving problems or completing tasks as members of a team. One has to relate to people at a variety of levels on social and organisational matters, in addition to the performance of the functions specific to one's occupation. One also has to manage one's own job and cope with unexpected events which fall outside the practices and*

[13] Criterion domain can be defined as the relevant domain of behaviour, knowledge, or skill in relation to which we need to establish the candidate's standing (McNamara, 2006, p. 33). McNamara highlights Messick's view that the criterion domain is itself a construct. Statements about the quality of performance in relation to the criterion domain refer to the criterion domain as a model of knowledge or performance, a generalised statement about what target performances are relevant, the nature of their demands, and what constitutes success in relation to them (ibid., p.34).

> *procedures of routine activities. […] It is often these less tangible aspects of competence rather than their technical skills that distinguish the successful and less successful employees.* (Jessup, 1991, p. 27)

A more recent definition reiterates the need to acknowledge that competence includes more than technical skills:

> *At the heart of an NVQ is the concept of occupational competence; the ability to perform to the standards required in employment across a range of circumstances and to meet changing demands. NVQs are first and foremost about what people can do. They go beyond technical skills to include planning, problem solving, dealing with unexpected occurrences, working with other people and applying the knowledge and understanding that underpins overall competence* (National Council for Vocational Qualifications, 1995, cited in Semta, 2009, p. 3).

This view of competence, although it originated in the NVQs, arguably also applies to competence-based QCF units. We are not aware of a different view being espoused since the inception of the QCF (Qualifications and Curriculum Development Agency (QCDA), 2010, p. 14). We therefore assume that this is roughly how competence might be defined in the relevant English VQs, and that this is the construct that WBOA is supposed to assess in different occupational settings. The discussion in this and the following section therefore attempts to summarise some main aspects in which this, or indeed any, definition of competence might present issues for individual assessment methods that might attempt to assess evidence of this construct, and thus for validity of qualifications.

The definitions of competence mentioned above suggest that, although the term competence is perhaps 'compelling in its common sense and rhetorical force' (Norris, 1991, p. 331), it refers to a complex concept, which defies clear-cut definition both within and across the fields in which it may be used. Whereas most authors discussing conceptualisation of competence seem to agree that knowledge and skills are its essential components, there is disagreement both about the other components that it encompasses, and about approaches to its assessment (see Appendix 2 for a table summarising common terms (characteristics) within a range of definitions of competence identified from the literature).

For instance, some definitions of competence within the field of vocational education and training (VET) internationally have notable differences compared to the English ones mentioned above (Brockmann, Clarke, & Winch, 2011). The EQF (European Qualifications Framework) emphasises responsibility and autonomy as important aspects of competence (European Commission, 2008, p. 11). In France and Germany a holistic approach is taken. In France there is an accepted definition of competence consisting of a combination of *savoir* (knowledge), *savoir-faire* (know-how or skills) and *savoir-être* (know-how-to-be or

behavioural/social aspects) (Le Deist & Winterton, 2005, p. 37; Gordon, et al., 2009, p. 36). In Germany, competence includes *Selbstkompetenz* (personal competence) which involves the individual moral and social responsibility in the workplace (Gordon, et al., 2009; Weigel, Mulder, & Collins, 2007, p. 57). Other authors, both within and outside the VET sector, also emphasise experience, maturity and autonomy and responsibility (e.g., (Eraut, 1994, pp. 167-168; OECD, 2005, p. 8). Attitudes have been regarded as one of the integrated elements of competence and some conceptualisations add motivational and social aspects (Gijbels, 2011, p. 382). Beliefs are also closely related to attitudes and motivation and so may also be included in the notion of competence (Weinert, 2001, cited in Braun, Hammad, & Hannover, 2011, p. 417; Bieri & Schuler, 2011, p. 405).

Furthermore, various authors tend to agree that competence involves integration of different elements within a context, irrespective of what exactly these elements might be (e.g., Tchibozo, 2011, p. 194; OECD, 2005, p. 4; European Commission, 2008, p. 11; Jessup, 1991, p. 26) and argue that people 'do not have competences independent of context' (Fischer, Bullock, Rotenberg, & Raya, 1993, p. 113, cited in Le Deist & Winterton, 2005, p. 30). The integration of elements means that the whole is more than the sum of its parts and in turn 'when we see the whole, we see its parts differently than when we see them in isolation' (Epstein & Hundert, 2002).

In the UK context, at the time when NVQs were being introduced, WBOA (alongside structured and varied training and work experience in the workplace) was proposed as a method that held a promise of enabling the assessor to see evidence of this broad notion of competence more directly, with the alternative being 'far more extensive and imaginative practical work, project work or other forms of simulation, in colleges and training centres' (Jessup, 1991, p. 27). Although this idea was accepted and WBOA became widely used, the issue with defining and agreeing what constitutes competence for specific professions and occupational roles remained, perhaps especially in the domain of the 'less tangible aspects'. In summarising the major critiques of the competence approach in England, Mulder et al. (2007, cited in Lucas, Spencer, & Claxton, 2012) suggest that the concept of competence is too often used to reduce what is valued down to an assessable ability to demonstrate skills and abilities successfully. Indeed, the question is also to what extent WBOA can always be used to assess these less tangible aspects of competence appropriately, or whether different assessment methods that might be specifically targeted for these might be needed (see discussion on pp. 25-27). This is another issue that should be addressed in validation.

It is clear from this discussion that defining competence in general is a complex endeavour, without a straightforward solution. To the extent that, for specific qualifications, the definition of competence represents the theoretical definition of the construct being assessed, the lack of an appropriate definition (or at least a generally accepted definition) could pose significant threats to validity. This is because the

theoretical definition of the construct affects its operational definition and decisions regarding how it is measured (Bachman, 1990, p. 43; Hagar, Gonczi, & Athanasou, 1994, p. 3). Indeed, the fact that the construct of competence is not easy to define and provide a theory or a model for perhaps explains why a task-centred approach is taken in WBOA, where the 'construct' appears to be a list of tasks that are supposed to be observed, rather than a model of the underlying abilities involved in executing these tasks.

In legacy NVQs as well as most publicly funded QCF qualifications, it is National Occupational Standards (NOS)[14] that provide 'definitions' of competence, or, more frequently, lists of tasks that need to be accomplished to demonstrate competence in individual occupations. In the next section we briefly describe how NOS are developed and what they involve. We discuss the issues surrounding their use as 'embodiment' of standards in specific occupations (Wolf, 1995) in the discussion of 'quantification' of WBOA observations on pp. 33-34.

National Occupational Standards – an example of how competence might be defined 'theoretically'

NOS attempt to define and specify the aspects and standards of competence relevant for individual occupations. Their development (with, in theory, a representative sample of employers as well as other stakeholders such as professional bodies, trade unions, awarding organisations and research organisations (Carroll & Boutall, 2010, p. 21) starts with an analysis of sector/occupation needs (occupational mapping) before a functional analysis to identify the functions people do, in terms of their purpose and outcome which are of value to an employer (ibid., p. 9). A key purpose is broken down into main functions which are then split into possible NOS (ibid., p. 24). Each standard describes a function or task within an occupational area. Each will specify elements (sub-functions or tasks), assessment criteria to be met, and the knowledge and understanding to be attained. Additionally, each standard may specify behaviours, values and skills (ibid., p. 53). NOS are intended to cover more than the technical essentials and include personal and meta-skills such as communication, organising work and making judgements (ibid., p. 8). Standards may include 'range statements' which 'indicate the range of application' (Jessup, 1991, p. 34). Without them, 'experience has shown that an element of competence is often open to different interpretations unless a more detailed specification is provided of what the element covers' (Jessup, 1991, p. 34).

However, there are a number of problems with using NOS as definitions of competence in specific occupational contexts. Research shows that NOS are recognised as being inconsistent between sectors, with employers reporting inconsistency in content and presentation of NOS, lack of relevance of NOS to their business, and difficulties they are having understanding the standards (UKCES, 2011). Similarly, (Eraut,

[14] Responsibility for the development and review of NOS was originally with the National Council for Vocational Qualifications (which changed into the Qualifications and Curriculum Authority) before this was handed over to the sector skills councils under the overview of what is now the UK Commission for Employment and Skills (UKCES).

Steadman, Trill, & Parkes, 1996) found considerable variations in the pattern of work between different contexts and stated that 'no set of occupational standards can be universally valid'. Other authors echo this view, by arguing that the idea, apparent in the conceptualisation of NVQs, that there is a consensual notion of competence for each role is, 'an heroic – and questionable – assumption' (Wolf, 2001, p. 3) and that 'there will be no absolute and objectively defined role in the first place' (Hodkinson, 1992, p. 32).

A mismatch between assessment content and the criterion domain has implications for content validity. The mismatch is more likely to be an issue when the criterion domain itself is not clearly defined (Benett, 1993), or, we would argue, where it is conceptualised as a list of tasks to be achieved, rather than as a model of crucial abilities involved in executing these tasks – which, as noted above, is the case with occupational competence in English VQs. NOS that are not fit for purpose in turn compromise content validity of WBOA as they may inadequately represent the criterion domain, which then the assessment claims to represent, and even if they represented the content domain, they may not be correctly understood and translated into assessment criteria. Thus, it is important to establish the degree of content validity, though, as noted earlier, even if content validity is deemed fine, further construct validation is still needed.

Apart from just defining the standards of competence for different occupations, NOS are also supposed to be closely aligned with the assessment process in that they attempt to promote a shared understanding of standards and thus consistency and comparability of standards (both for individual assessors and nationally) in assessing competence, by providing detailed specifications of assessment criteria and context of application. We discuss the validity issues and the implications of the 'standardising' function of NOS in pp. 33-34 below.

Defining the construct of competence operationally: a discussion of the rationale for using WBOA and its procedures

Irrespective of how the constructs of interest are theoretically defined, it is necessary to specify how they will be isolated and made observable for the purposes of measurement. The specific operations that are used for making the constructs observable will, however, reflect both the theoretical definition of the constructs and what is considered to be the context in which the relevant behaviour typically takes place (Bachman, 1990).

WBOA, as practised in competence-based vocational qualifications, rests on the behaviourist tradition, which infers competence from performance – of action, behaviour or outcome that can be demonstrated,

observed and assessed (Norris, 1991, p. 332)[15] in the real work setting. The context of real work setting corresponds to the definition of competence in these qualifications that is based on the analysis of actual job roles, functions, etc. in the NOS. Thus, as noted previously, WBOA in competence-based VQs operationalises the construct of competence in a task-centred way, with the existence of various aspects of competence being established through observing a candidate successfully executing a task, or their successful performance being confirmed by inspecting an artefact produced.

One problem with the assessment focus on easily observable actions, behaviours and outcomes, is that certain important aspects of competence may effectively be removed from the theoretical definition of the construct and from assessment focus because they cannot be easily observed and measured. For this reason, the behaviourist approach is sometimes considered too simplistic (Hyland, 1993, p. 59; Hodkinson, 1992, p. 34) since not everything that might be considered part of competence may be (easily) observable (e.g., attitudes and values). Hagar, Gonczi, & Athanasou (1994, p. 14) express this concern as follows: 'This poses an apparent dilemma for competency standards – omit attitudes and values and be invalid, or include them and be unworkable'.

This has potentially important validity implications, particularly in professions where less tangible aspects of competence such as attitudes and values might play an important part in carrying out the relevant occupational tasks (e.g., in customer services, social care, etc.) and thus in the relevant definition of competence. Validation efforts should thus establish whether these less tangible aspects of competence are included in relevant occupational standards, and if so, to what extent these are appropriately assessed through WBOA, whether more emphasis on them might be needed in WBOA, or whether different, more appropriate methods of assessment for those aspects might be warranted. Mismatch in this domain or a failure to acknowledge an existing mismatch in the proposed interpretation of assessment results might represent a potential threat to validity.

As noted earlier, a variety of different methods are used in the assessment of competence in competence-based VQs, with WBOA generally providing key evidence. No studies have explicitly investigated the advantages of assessment in the workplace, and of WBOA specifically, over other possible assessment settings and methods (e.g., simulations) before (or indeed, after) its introduction in the nineties. However, medical research literature presents findings of a number of studies that have demonstrated that doctors' abilities when assessed in a controlled environment do not dependably predict their actual day-to-day performance (e.g., Rethans, Sturmans, Drop, van der Vleuten, & Hobus, 1991; Kopelow, et al., 1992, cited in Crossley & Jolly, 2012). According to Crossley & Jolly (2012, p. 29), this

[15] It has been acknowledged that underpinning knowledge must be assessed too in order to get a complete indication of one's (occupational) competence (City & Guilds, 2011a).

represents an argument in favour of using work-based assessment in medical education (see also Norcini & Burch, 2007), and this would presumably be the case in NVQs and other vocational qualifications too. In healthcare, for instance, observational techniques in particular are thought to allow evaluation of core competencies such as patient care, interpersonal and communication skills and professionalism, which are arguably difficult to assess through other methods (Fromme, Karani, & Downing, 2009).

Based on Baillie & Rhind (2008), the figure overleaf attempts to summarise the key methods that are used in medical work-based assessment, both formatively and summatively, and maps this to what might be deemed equivalent methods in VQ workplace settings. This figure illustrates that there are a number of assessment methods that are used in the workplace setting to assess different aspects of medical competence. For instance, the DOPS method is designed to assess procedural skills, and this has implications for how the recording forms that are used in observations will be designed, and how observation itself is focused and carried out. On the other hand, the methods like Mini-CEX, with its focus on more integrated skills will have different rating scales associated with it, and observation will be focused differently.

In contrast, most of these methods map to a more or less generic method of WBOA in VQs, where to our knowledge there is no consistent practice of systematically amending and targeting the generic approach in order to get at different aspects of competence more directly.[16] Furthermore, as noted earlier, it is sometimes left up to the assessor (and occasionally circumstances) to decide whether WBOA will be used to collect evidence of competence, or alternatives such as expert/non-expert witness statements, professional discussion, report, questioning, etc.

Although there are clear advantages of assessing certain aspects of competence in the workplace, it would arguably be in the interest of increasing assessment reliability and validity of WBOA and VQ work-based assessment in general, to devise discreet methods that might provide complementary evidence, and each be more suitable than the others for making different aspects of competence observable for measurement (e.g., procedures vs. professionalism and integration in the workplace).

[16] A potentially useful approach to identifying the most appropriate assessment method (and also, the most appropriate construct/model of competence) might follow the mapping between vocational education outcomes and learning methods proposed in Lucas, Spencer, & Claxton (2012).

Figure 2: Comparison of work-based assessment methods in medicine and English VQs

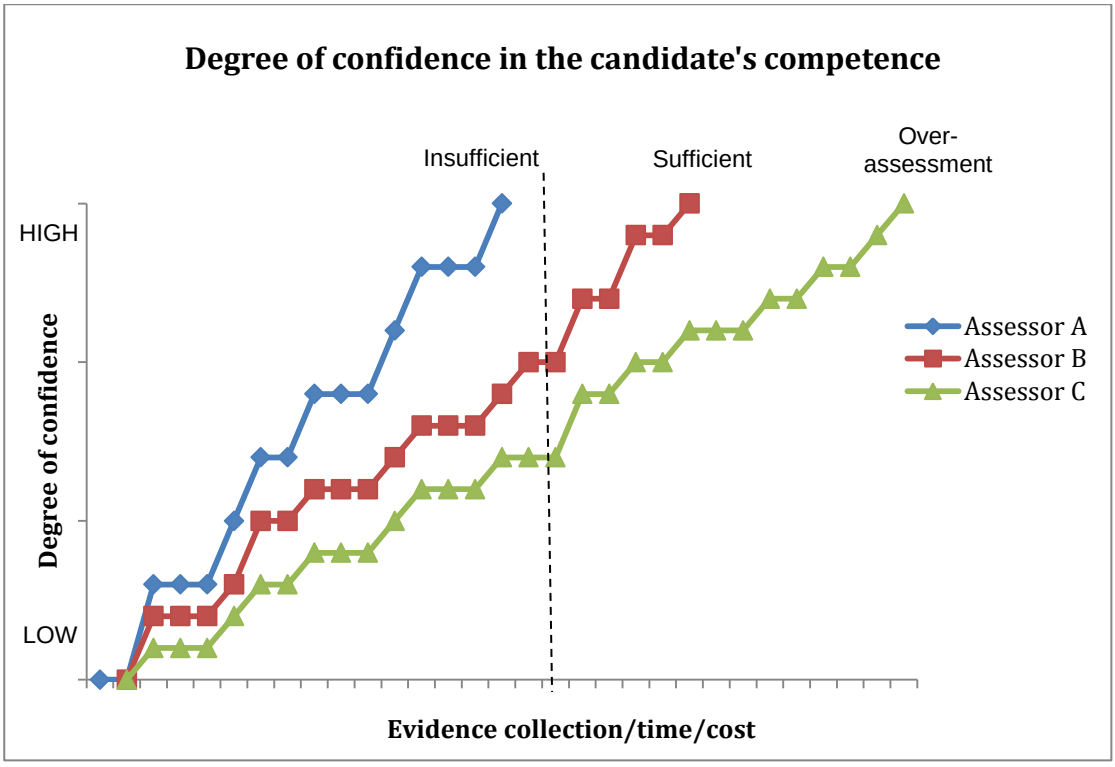
Workplace setting – assessment methods in post-graduate medical education				
Mini-clinical Evaluation Exercise (mini-CEX)	**Longitudinal Evaluation of Performance (LEP)**	**Directly Observed Procedural Skills (DOPS)**	**360° (Multi-source Feedback)**	**Case-based Discussion and Chart Stimulated Recall Oral Examination (CSR)**
• Primarily formative • Direct observation of a trainee by one examiner during a clinical encounter with a real patient in the normal work setting • Lasts 15 – 20 minutes • 4 to 6 mini-CEX exams per year in a variety of situations • Standardised tick box form used and performance rated for a list of skills as: at, above or below expectation **Skills assessed**: History taking, physical examination, problem solving, clinical reasoning, communication	• Primarily formative • Similar in format to the mini-CEX • Evaluations typically at weekly intervals and record trainee's progress towards the required standards • Does not require coverage of a specific list of cases – choice left to assessor/supervisor • A rating form used in observation **Skills assessed**: Technical skill, knowledge, clinical reasoning, professionalism, communication	• Contributes towards summative • Trainee observed and scored by an assessor while performing key routine practical procedures during normal clinical work • Typically six assessments per skill to be signed off as competent at that skill • A standard DOPS form used for scoring **Skills assessed:** Practical/technical skills	• Contributes towards summative • Feedback from staff who are more senior, more junior and peers; representatives of all groups in the clinician's daily working environment (not just co-professionals); and sometimes from patients and their families • A structured form or a questionnaire used **Skills assessed**: Communication, team working, professionalism	• Primarily formative • A formal discussion between a trainee and an assessor about a case that the trainee has managed and been directly responsible for • Structured rating form used **Skills assessed**: Application of knowledge, decision making, clinical judgement, professionalism
⇩	⇩	⇩	⇩	⇩
Workplace setting – assessment methods in competence-based VQs				
WBOA/Expert witness testimony	**WBOA/Expert witness testimony**	**WBOA/Expert witness testimony**	**WBOA/Non-expert witness testimony/Informal chat**	**Professional discussion**

The WBOA assessment process relies on assessors being trained and capable to carry out observations, which are supposed to be carried out in an unobtrusive way and without interruption. The competence of assessors in observational techniques could be an issue where assessors do not accurately observe relevant skills being demonstrated or errors being made by learners when performing a task (Norcini & Burch, 2007), or are unable to identify suitable positions and distances from which to observe learners performing tasks without interruption (Fitzpatrick & Morrison, 1971; Hauer, Holmboe, & Kogan, 2011). Even industry experts can be 'unsystematic and erratic' in their observations and different observers are likely to observe different aspects unless what they need to observe and note is specified in a checklist or some other form (Fitzpatrick & Morrison, 1971). These are all procedural issues that could threaten the validity of the conclusions based on WBOA, and should be investigated in its validation.

As discussed earlier, assessors need to observe enough evidence to make an accurate judgement about a learner's competence based on the set criteria but they also need to know when to stop looking for more evidence. The assessment design should enable trained assessors to collate enough relevant evidence to make such decisions and the number of observations (and length of task) is probably crucial in achieving this goal.

Bartram & Mitchell (1992) developed the chart in Figure 3, and an associated argument about sufficiency of evidence and assessors' confidence in judgement.

Figure 3: Degree of confidence in a candidate's competence



Source: Bartram & Mitchell (1992)

In Bartram & Mitchell's chart, three assessors observe performances/gather evidence of competence. As they gather evidence/observe performances, their confidence in their judgement of the candidate's competence increases: Assessor A is bold, she arrives at a confident judgement quickly, but her confidence may be misplaced. Assessor C is more cautious and requires a lot more evidence before he will confidently proclaim the candidate's competence. Assessor C's approach avoids 'false positive' classifications of competence, but is also costly, inefficient and potentially leads to 'false negatives', which may be unfair to the candidate. Assessor B is more cautious than assessor A, but more bullish than assessor C. The amount of evidence that she needs to observe to make a judgement is just right.

There is further debate around the types of evidence that are best to engender assessors' confidence that they have sufficient evidence to judge a candidate as competent, including the notion that certain observations or types of observations may be 'canonical'; affording a trained assessor 'definitive proof' of competence (Bartram & Mitchell, 1992, p. 17; see also Murphy, et al., 1995, pp. 26-27). Other distinctions are made between 'sufficient' and 'necessary' evidence; and the ability of an assessor to reasonably infer a candidate's underlying knowledge by observing a performance.

Reliability is likely to increase with the number of observations but it is challenging to determine minimum and/or maximum numbers to allow sufficient evidence to be gathered (Fromme, Karani, & Downing, 2009) while keeping the assessment workable and efficient.[17] In WBOA, candidates tend to be judged over an unknown or unlimited (or at least multiple) number of attempts (Harth & Hemker, 2012), although the relevant assessment strategy will provide guidance about the minimum number of observations. Thus, it will generally be left up to the assessor to determine how much or little is enough to reach a confident judgement, with threats to validity here of over- or under-assessment.

In WBOA, tasks supposed to address the same assessment criteria typically vary across candidates, employers and centres. Centres/assessors/candidates are allowed to select tasks for assessment as long as these are consistent with the required occupational and assessment standards. Variation in assessment outcomes due to the sampling of tasks or settings is a potential threat to reliability, generalisability and validity. In the context of medical education, Norcini (2005) talks about lack of control in patient complexity as a potential threat to validity, reliability, and fairness of observational assessment. This relates to the

[17] In the medical education context, Norcini (2005) states that determining the number of patients to ensure reliable results must be a priority. Norcini & Burch (2007) draw on a number of studies to identify the number of patient encounters needed to produce reliable results for different observational methods used in work-based clinical assessment:

- Approximately 12-14 encounters to achieve a 0.8 reliability coefficient for the mini-CEX (Norcini et al. 2003; Holmboe et al. 2003, cited in Norcini & Burch, 2007)
- Around eight encounters for a reliability coefficient of 0.8 or more in CEC (Hatala & Norman, 1999, cited in Norcini & Burch, 2007)
- Approximately seven encounters for a 0.7 reliability coefficient for the CWS (Turnbull et al. 2000, cited in Norcini & Burch, 2007)

variation among patients with the same condition caused by factors such as severity of illness, other health problems, and adherence to treatment (Norcini, 2005). Although real patients are used, complex strategies are applied to focus assessment on particular patients and conditions (Kessner et al., 1973, cited in Norcini, 2005) and to exclude patients with another severe illness or multiple conditions (Norcini, 2005). Thus, standardisation (and fairness) is improved, although possibly at the risk of reducing validity in a different way. The same issue applies to WBOA where a trainee chef or a hairdresser faces a difficult customer with extraordinary demands while being summatively assessed by WBOA. To our knowledge, there is no systematic attempt to control such variations in this context.[18]

Variation in learning and assessment environment is another important characteristic of WBOA with potentially significant ramifications for reliability and validity of assessment. For instance, Torrance et al.'s (2005) research found that the learning environment tended to vary between workplaces due to lack of resources. They gave the following example:

> *Small garages may not provide NVQ Level 3 opportunities to conduct diagnostic work with the latest computer technology. Equally, however, and somewhat ironically, well-resourced main dealers for leading car makers do not always provide NVQ Level 2 opportunities for basic repair.*

Fitzpatrick & Morrison (1971) make a similar point by stating that 'variations in equipment quality or state of repair may affect the quality of the product; a balky sewing machine may produce a faulty stitch even though the student has fed the material in straight'.

These issues relate to our previous discussion about the difference between a genuine (real-life) task, and a task that is authentic and allows us to see evidence of the construct that is being measured. If part of the construct is 'dealing with a difficult customer', or 'problem solving in a complex situation', then, arguably, every candidate needs to be assessed in the circumstances which will allow assessors to get evidence of that. Obviously, if dealing with complex customers and situations is not part of the construct, then issues of fairness arise for those candidates that happen to be assessed in such a situation. Of course, it is difficult to 'standardise' real-life situations, but perhaps it would be useful to at least have clear guidance about what assessors should do in such situations, and how they might 'weight' their decision to take into account the extra difficulty that some candidates might face in their assessment.

[18] Tasks involving team work also complicate matters. If a WBOA requires a team to perform a task (or set of tasks) or produce a product but with an individual within the team being assessed, then the assessor would need to make a judgement about the individual learner's contribution. Taking another medical example, a competent doctor's performance in a team may be hindered where other team members do not perform well and vice versa. In such a situation, certain procedures can be adopted e.g., including tasks where the candidate has more influence or control (Norcini, 2005) but this again may compromise validity.

It can be seen from the previous discussion that most of the threats to validity when it comes to implementing WBOA operationally have to do with the relationship of the method with the theoretical construct, and with the issues of standardisation (of assessment situations and tasks, as well as assessors – see next section) or lack thereof. Kane, Crooks, & Cohen (1999) note that the extent of standardisation in assessment on the one hand, and how complex and contextualised each performance task should be on the other, are two major concerns in the design of performance assessments in general that have a major impact on validity. All these issues are very important to consider both in implementing and validating WBOA. We discuss issues around standardisation of performance assessment further in the section on validity argument for performance assessment at p. 40.

WBOA procedures for 'quantifying observations'

Evidence of aspects of competence in WBOA is judged in terms of scope (range of situations in which the competence applies), quality of performance, and appropriate level of competence corresponding to the level of qualification, and is judged dichotomously (e.g., achieved/not yet achieved or competent/not yet competent[19]), often based on a list of ACs (Harth & Hemker, 2012). The assessors are expected to apply the relevant professional standards to assess performances, ensuring that all the relevant ACs have been met (see the examples of LOs and ACs in Table 1 and Table 2 and also in Appendix 3).

Table 1: Assessment criteria for LO 1 'Establish effective rapport with clients' in Unit G17 'Give clients a positive impression of yourself and your organisation' (L2 NVQ Diploma in Hairdressing (3008))

a) Meeting your salon's standards of appearance and behaviour b) Greeting your client respectfully and in a friendly manner c) Communicating with your client in a way that makes them feel valued and respected d) Identifying and confirming your client's expectations e) Treating your client courteously and helpfully at all times f) Keeping your client informed and reassured g) Adapting your behaviour to respond effectively to different client behaviour

[19] The notion of a standard or threshold of what is minimally acceptable is implicit within the concept of competence. This carries two connotations, either positive – an individual is competent – or more negative, that an individual is at a minimally satisfactory but not excellent level (Eraut, 1994, p. 166; Herling, 2000, p. 9). The former relates to a binary notion of competence while the latter may see it as a scale. The standard or threshold may relate to what is deemed as minimally efficient and or safe (Bradshaw, 1997, p. 350, cited in Redfern, Norman, Calman, Watson, & Murrells, 2002, p. 56). Carroll & Boutall (2010, p. 7) state that the NOS (upon which NVQs are based) 'describe best practice'.

Table 2: Assessment criteria for LO 3 'Be able to install plumbing and heating systems and components in the workplace' in Unit 345 'Install, commission, service and maintain domestic plumbing and heating systems' (L3 Diplomas in Domestic Plumbing and Heating (6189-31/32/33))

The learner can:

a) confirm that the incoming or outgoing main supplies meet the requirements of the system or component being installed
b) measure and mark out the position of the components to be installed:
    a. system pipework
    b. main system components
    c. system controls
c) make pipework and component fixings to the building fabric
d) position and fix pipework and components to the building fabric:
    a. copper
    b. plastics
e) connect pipework to system controls and main components:
    a. cold water systems
    b. hot water systems
    c. central heating systems
    d. sanitation systems
f) connect system pipework to incoming supplies or outgoing services
    a. existing system pipework and components
    b. cold water supply pipework
    c. below ground drainage pipework
g) carry out installation work, minimising the wastage of equipment and materials
h) take precautions to ensure that the system cannot be brought into operation before the installation work is fully completed

The lists of ACs in the examples above can be seen as 'checklists' of things that candidates are supposed to demonstrate competence in in order to be deemed to have met the relevant LO. However, these checklists are typically not accompanied by standardised global scales for judgements about specific ACs, which makes it difficult to perform reliability and other quantitative analyses (see the relevant discussion in the section on quantitative evidence of validity at p. 87).

The use of standardised rating scales is traditional in some forms of observational assessment in other contexts such as medical education or language testing, although research evidence regarding the use of standardised rating scales is controversial. Some research evidence shows that the type of an assessment tool and the relevant global scale can play an important role in achieving consistency in observational assessment (Fromme, Karani, & Downing, 2009). Hauer et al. (2011) state that existing medical assessment tools, which use numerical rating scales with benchmarks and behavioural descriptors can provide assessors with a framework 'that may promote greater rater accuracy and stringency and decrease rating biases'. Of course, these cannot guarantee consistency as evidenced in Noel et al. (1992), who found that structured forms listing specific skills were more accurate than open-ended forms but that they did not improve overall assessor accuracy where assessors' applied different standards, lacked experience and had insufficient training. Crossley & Jolly (2012) found that observers may agree with what they have observed

but often disagree with the interpretation of the rating scales i.e. distinguishing between scales categories and the competencies listed.

Thus, although rating scales evidently cannot solve the problem of assessor inconsistency and mismatch in standards, they might contribute to this at least to some extent, and also enable relevant analyses of assessor behaviour, in order to collect evidence of agreement levels, differences in severity, etc., thus contributing to the pool of evidence about the validity of WBOA. [20]

Currently in WBOA, instead of using rating scales, a shared understanding of standards, as well as consistency and comparability in assessment, is promoted by specifying assessment criteria in substantial detail as can be seen in the examples above. Wolf (1995, p. 28) observes that the actual process of recognising, and so assessing, competence in one's own vocational field is seen as unproblematic in VQ assessment as it should be based on knowing the NOS and how these relate to a specific workplace. However, evaluating complex performances – such as workplace performance, for instance – is never easy or unproblematic (Kane, 2006; Sadler, 1989).

Wolf (2001, p. 9) questions the usefulness of NOS and their 'precisely' defined standards for ensuring agreement regarding the interpretation and consistent application of standards by assessors in the field:

> *However 'precise' one becomes when one goes down this route, there is always call for yet more definition. Performance criteria might mean all sorts of things – so we added range. Range can be interpreted in all sorts of ways, we added more lists. At the end of this process, and in good faith, people can still be ascribing 'competence' to very different behaviour.* (Wolf, 2001, p. 8)

She argues that through this approach, 'an atomised, tick-list system is almost bound to result' (Wolf, 2001, p. 7). Yet, realistically, assessors 'operate with a complex, internalised, and holistic model – not a simple set of descriptors lifted from a printed set of performance indicators.' (Wolf, 2001, p. 9; cf. Eraut, Steadman, Trill, & Parkes, 1996).[21]

[20] Pollitt has argued that the measurement approach of adaptive comparative judgement (ACJ) allied to statistical techniques such as Thurstone Pairs or Rasch-based measures to derive judge, item and person parameters has important uses for judgement-based assessment (Pollitt, 2004, 2012) and might be superior and more valid than marking or rating. This appears sustainable from Pollitt's reported research. See the relevant discussion on p. 92.

[21] It is also considered essential within the competence-based approach (and thus in WBOA) that candidates understand and have internalised the assessment criteria and standards by which their work is being assessed to an extent that allows 'thorough preparation as well as accurate self-assessment and self-adjustment' (Wiggins, 1993). Torrance et al. (2005) state that transparency of learning outcomes and the clarity of assessment procedures, processes and criteria have significantly benefited learners in the learning and skills sector through extensive tutor and assessor support i.e. 'exam coaching and practice, drafting and redrafting of assignments, asking 'leading questions' during workplace observations, and identifying appropriate evidence to record in portfolios'. However, transparency, along with greater and widespread assessor support, can reduce the validity of outcomes achieved 'where assessment procedures and practices may come completely to dominate the learning experience, and 'criteria compliance' comes to replace learning' (Torrance, et al., 2005).

Obviously, if one is considering validity of assessment outcomes, it would be important to get some indication of what contributes to decisions about candidates that are based on this complex holistic model, if this is indeed used. For example, Colley & Jarvis (2007) found evidence of the content of assessment including not just formal judgements about apprentices' competence in terms of specified assessment criteria and underpinning knowledge, but also the subversion of such judgements on the basis of informal judgements about candidates' attitudes and dispositions and, consequently, of their personal worthiness. There was also indication of 'good blokes' being passed by their assessors, even though they had not necessarily demonstrated competence according to all the criteria in the officially required manner. Employers' expectations of 'the good bloke' were often related to their level of integration in the workplace, sometimes defined by gender, race and class, and this, perhaps unwittingly, permeated assessor judgement.

Torrance et al. commented thus on the tension between national and local, which could also be understood as the contrast between formal and informal standards:

> *The balance between complying with 'national standards' and interpreting them appropriately in situ needs to be re-examined … definitions of standards can never expunge local interpretation, and the evidence from this study and others (e.g., Fuller & Unwin, 2003, Stasz et al. 2004) is that local 'communities of practice' constitute the context in which all meaningful judgements about standards are made.* (Torrance, et al., 2005, p. 3)

It is thus necessary to appreciate the limitations of NOS or similar written standards in terms of how far they can go in ensuring a shared understanding of standards of competence among assessors. This also undermines the assumption that assessment that relies on such externally defined standards is unproblematic. Such assumptions might be more of a threat to validity than recognition that assessors are more likely to be somewhat inconsistent in application of performance standards. This in turn might lead to endeavours to develop effective mechanisms and procedures to establish and monitor consistency levels, and to mitigate inconsistencies (by perhaps promoting a shared understanding through discussions and communities of practice (Wenger, McDermott, & Snyder, 2002; Konrad, 1999, cited in Greatorex, 2005, p. 150; Walker, 2010), investing more in appropriate standardisation acitivities and verification, and/or developing meaningful global rating scales for those aspects of competence that are considered important and should be the focus of assessment).

Furthermore, if standards were indeed best understood at the level of (local) communities of practice, this might interact in important ways with whether assessors are work-based or visiting. In the former case, the assessors are arguably very directly part of a community of practice that 'holds' the industry and other relevant standards (though perhaps not the 'national' standard), which might lead to better consistency in

their judgements. In the case of visiting assessors (assessor being their primary role), the communities of practice with respect to a certain industry/sector might conceivably be weaker. See footnote 6 at p. 13, and the discussion about the implications this might have for generalisability of WBOA results at p. 42.

There are issues here, and potential threats to validity, with respect to both assessment methodology (nature of assessment tools, nature and effectiveness of standardisation, differences in whether assessors are work-based or visiting, etc.) and the definition of the construct. For instance, in Colley & Jarvis's (2007) 'good bloke' example (bottom of p. 34), perhaps attitudes and dispositions that make somebody a 'good bloke' are actually very relevant for assessors and should be explicitly assessed to the extent that this is possible, but are not explicitly included in the construct definition for various reasons, and are therefore not included in the assessment criteria.

Verification

Maintaining consistent standards in evaluating complex performances is not easy (Kane, 2006) as a great deal of research confirms (e.g., Linn et al., 1992 and Klein et al., 1995 – both cited in Kane, 2006; Swanson, Norman, & Linn, 1995; etc.). It is, therefore, necessary to implement a form of 'social moderation' (verification) (Kane, 2006), which can help build consensus across individual providers of assessments and qualifications, and is important for gaining public acceptance for qualifications and assessments. In English VQs, qualification delivery and assessment processes, including WBOA, are verified by IQAs, appointed by centres, and QCs,[22] appointed by awarding organisations.

Briefly, according to official guidance (see City & Guilds (2006b) for more details), IQAs focus on verifying the work of assessors associated with a specific centre including observing them while assessing to check on consistency and standards applied, they are involved in the assessor selection process, they support assessors in their development and guide assessors where appropriate, and ensure a level of standardisation. They also collaborate with QCs in providing information relevant for external quality assurance processes.

QCs have a wider role, ensuring that assessment and quality assurance undertaken within centres meets necessary requirements and are consistent between centres. Both internal and external verification procedures generally rely on risk-based sampling and could involve checking relevant centre documentation, checking candidate portfolios, observing assessors and/or IQAs conducting assessments, discussions with candidates, etc. QCs tend to rely on advice from IQAs when sampling assessors that they will observe during any particular verification visit. Currently, a City and Guilds QC will visit a low-risk centre once a year.

[22] Until recently, IQAs and QCs used to be ubiquitously referred to as Internal Verifiers (IVs) and External Verifiers (EVs) respectively.

There is some research evidence that verifiers do not adequately check assessor judgements and instead are more concerned with the related documentation (Eraut, Steadman, Trill, & Parkes, 1996). Based on this research, Eraut and colleagues recommended that verification, and the related training of assessors and verifiers, should focus on judgements and that attention should be given to aspects most important to distinguishing between 'just competent and not yet competent' (Eraut, Steadman, Trill, & Parkes, 1996, p. 68). This might help to address the potential threats to validity coming from a lack of standardisation and inconsistencies between assessors when assessing using WBOA.

Obviously, it is not just assessors, but also verifiers that are expected to evaluate complex performances. Therefore, unsurprisingly, research carried out by Torrance et al (2005) found evidence of inconsistent interpretation of the standards and practice by external verifiers themselves. This and the abovementioned problems could be caused by high turnover of external verifiers and by awarding organisations not monitoring external verifiers comprehensively (QCA, 1999c, cited in Greatorex, 2000). All these are possible threats to validity of WBOA, and would require attention in relevant validation studies.

A definition and summary

In this review, we have discussed various features of WBOA, situating them in the context of the method's use in competence-based VQ assessment, and in the wider context of performance assessment. Based on this review we propose the following definition of WBOA as it is currently used in competence-based VQs:

WBOA can be defined as a real-life task-centred performance assessment method, where an assessor observes a candidate performing practical tasks and other activities in the workplace, and/or inspects the artefacts produced, in order to make a dichotomous judgement (achieved/not yet achieved) against assessment criteria about aspects of a candidate's competence.

Importantly for the development of the validation framework, we have identified a number of aspects and potential threats to validity that would need to be considered in the validation of WBOA. These include:

- construct-related issues and the absence of appropriate focus on a theoretical definition of the construct in WBOA (abilities vs. tasks; easily observable vs. less tangible; is genuine always authentic; etc.)
- threats related to content validity – mismatch between assessment content and criterion domain
- questions of whether WBOA as currently implemented is always the most suitable method for assessing the relevant aspects of competence
- a variety of procedural issues related to observational techniques
- issues around assessor confidence and the potential for over- and under-assessment
- issues around standardisation or lack thereof in the domain of tasks – variation across employers and in relation to task complexity
- variation in learning and assessment environment
- issues related to quantification of observations or lack thereof, associated difficulties with performing reliability and other quantitative analyses, and issues around standardisation of assessors
- issues around the assumption that the WBOA process is unproblematic since external occupational standards ensure consistent judgements
- issues related to effectiveness of verification and standardisation of verifiers
- potential differences between work-based and visiting assessors with respect to a number of issues mentioned above.

These key threats have informed the development of our validation framework and have been incorporated into its relevant parts.

The argument-based approach to validation: a review

In this section we briefly describe the key aspects of the argument-based approach to validation proposed by Kane (2006) and summarise and discuss Kane, Crooks, & Cohen's (1999) proposal of how a validity argument for performance assessment might be developed. Finally, we review some previous implementations of the argument-based approach to validation in the frameworks developed for language assessment, traditional assessments in general qualifications, and for competence assessment. We point out the key 'lessons' we have learned from these frameworks and attempted to implement in our own.

The argument-based approach to validation

Kane (2006) adopted Cronbach's (1988) proposal that the validation of score interpretations and their uses be based on the logic of evaluation argument, developing what has become known as the argument-based approach or framework for validation. According to the argument-based approach, a clear statement of the proposed interpretation and use of the results of an assessment is necessary as the foundation for the validation process (cf. AERA, APA, & NCME, 1999). The next step involves development of an interpretive argument that provides an explicit statement of the inferences and assumptions leading from the observed performances to the conclusions and decisions based on these performances, i.e. score interpretations (Kane, 2006, p. 23). An interpretive argument essentially provides a framework for validation. Validity argument, in turn, provides an evaluation of the inferences and assumptions in the interpretive argument, that is, an evaluation of the appropriate evidence for and plausible alternative interpretations of the assumptions (Kane, 2006, p. 23).

Kane (2006, pp. 25-26) states that the usefulness of developing and using interpretive argument in validation is that it makes the reasoning inherent in the proposed interpretations and uses of scores explicit so that it can be better understood and evaluated, and be less likely to rely on implicit assumptions. To the extent that the interpretive argument is specified in some detail, gaps and inconsistencies are harder to ignore, and a lack of evidence for one or more steps in the argument may be more apparent. Of course, part of the challenge of developing the interpretive argument for validation is deciding which of the assumptions need to be evaluated and thus included in the validation framework, and which can be taken for granted and accepted without evidence.

Kane (ibid.) sees the development of the interpretive argument as going hand in hand with the development of the measurement procedure to which it applies, although it can be developed post-hoc too. Efforts to make the measurement procedure consistent with the proposed interpretations and uses provide support for the plausibility of the interpretive argument. Also, the inferences and assumptions in the interpretive argument would be evaluated to the extent possible during test development. Any

weaknesses identified in this process could lead to modification of the interpretive argument or the test. This process continues until the test developers are satisfied with the fit between the test and the interpretive argument. This development stage has a confirmationist bias – if a problem is identified, it is fixed if this is possible.

According to Kane (2006), at the point when the development process is considered complete, validation effort should adopt a more neutral or critical stance. During this appraisal stage of validation, the test is taken as a finished product and the assumptions and inferences specified in the interpretive argument, as well as the clarity and coherence of the interpretive argument are interrogated and critically evaluated using appropriate methods, and conclusions about their clarity, coherence, plausibility and justifiability or lack thereof are reached. In other words, backing or rebuttals are provided for the assumptions and inferences in the interpretive argument (Toulmin, 1958). Kane notes that the appraisal stage should also involve a search for hidden assumptions and investigations of possible alternative interpretations of the test scores. This is because the interpretive argument specified during the development phase may be incomplete in various ways (e.g., the criteria used for scoring performances may make a number of value judgements that are not spelled out in any detail).

One of the challenges of validation is creating a set of evidence that is comprehensive enough yet can be accumulated in a way that is practical enough to be implemented operationally in awarding organisations' procedures. Though validity can be seen as an evolving property and a matter of degree, and validation as essentially a continuing process (Messick, 1989, p. 13), decisions have to be made regarding how much evidence is enough for a particular purpose. Kane (2006, p. 29) observes that if it were necessary to support every inference and assumption with empirical studies conducted after the assessment procedures are developed, validation would be essentially interminable, because most interpretations involve a number of inferences each of which relies on multiple assumptions, and the evaluation of these assumptions will rely on other assumptions. Fortunately, many inferences and assumptions are sufficiently plausible a priori to be accepted without evidence unless there is some reason to doubt them in a particular case.

In general, several different kinds of evidence would be relevant to each inference and its supporting assumptions, including expert judgement, empirical studies, the results of previous research, and value judgements (Kane, 2006, p. 25). For highly questionable inferences and assumptions, Kane suggests that it is plausible to consider several parallel lines of evidence and judge their plausibility in terms of all of the evidence for and against it.

The empirical studies are perhaps the most cumbersome and expensive way of collecting validation evidence and Kane (2006, p. 26) cites Cronbach (1989, p. 165) who proposed four criteria for identifying the empirical studies to be pursued by the test evaluator:

1. Prior uncertainty: Is the issue genuinely in doubt?
2. Information yield: How much uncertainty will remain at the end of a feasible study?
3. Cost: How expensive is the investigation in time and dollars?
4. Leverage: How critical is the information for achieving consensus in the relevant audience?

An interpretive argument that has survived all reasonable challenges to its clarity, coherence and plausibility of its inferences and assumptions can be provisionally accepted, with the understanding that new evidence may undermine its credibility in the future (Kane, 2006, p. 27). Based on this, the relevant stakeholders could conclude that the assessment of interest and the interpretations assigned to the scores (as specified in the interpretive argument) are valid.

In the next section, we give an overview of the validity argument for performance assessment mainly based on the paper by Kane, Crooks, & Cohen (1999), as it touches on several important issues that are relevant for validation of WBOA as a type of performance assessment.

Validity argument for performance assessment

As discussed earlier on p. 18, the apparent fidelity and authenticity/genuineness of performance assessment do not automatically ensure that its proposed interpretation is always valid (Messick, 1994). Regardless of how direct the assessment appears, the skills and knowledge are not measured directly, but are inferred from performances and products (ibid., p. 21). As with any other assessment method, the plausibility of these inferences needs to be established in order to justify claims regarding the validity of score interpretations based on performance assessment. In addition, many other aspects of the assessment process can affect validity of score interpretations, e.g., assessment conditions, assessment criteria, assessor inconsistency, etc.

Any assessment method or test could be argued to involve score interpretation based on a sample of performance, and extrapolation to a wider domain of performance (or target domain, cf. Kane, Crooks, & Cohen, 1999). The interpretation is typically not limited to the specific performances included in the assessment, and extends to some general type of performance (e.g., a job role), of which the performances in the assessments are examples. Many authors have observed that limited sampling of relevant performances from a target domain, owing to issues of practicality, safety and fairness as well as the complexity and/or length of the performance tasks, poses the main challenge for the validity of performance assessment in particular (e.g., Kane, Crooks and Cohen, ibid.; Swanson, Norman, & Linn, 1995). Kane, Crooks, & Cohen (ibid.) refer to the subdomain for which it is plausible to consider the observed performances to be a random or representative sample as the *universe of generalisation*, and an

individual's expected score over the universe of generalisation as the *universe score* for the assessment procedure. This range of observations is often much narrower than the range in the target domain.

Expressed in these terms, the question is how to get from performance on a sample of tasks drawn from the universe of generalisation to valid conclusions about expected performance over the target domain. In their discussion of the issues surrounding validation of the interpretations of performance assessment scores, Kane, Crooks, & Cohen (ibid.) propose that the core of the interpretive argument contains three steps:

SCORING ⇨ GENERALISATION ⇨ EXTRAPOLATION

The authors discuss the assumptions and issues that might affect the appropriacy and plausibility of these inferences in some detail in their article. For instance, the appropriateness of the inference from a performance to an observed score rests on at least two assumptions: first, that the criteria used to score the performance are appropriate and have been applied as intended and, second, that the performance occurred under conditions compatible with the intended score interpretation. The more open-ended the tasks and the more complex the performances to be assessed, the harder it becomes to evaluate the quality of performances in a consistent and unambiguous way. This is partly because it becomes more difficult to anticipate the range of possible responses and to develop fair and explicit scoring criteria that can be applied to all responses (Swanson, Norman, & Linn, 1995). The second assumption requires that the observed performances occur under conditions consistent with an interpretation of the examinee's level of skill, i.e. that there are no improper impediments to their performance, that they are motivated to perform well, and that they do not enjoy any inappropriate advantage. The evidence supporting the credibility of these inferences and legitimacy of scoring and ruling out alternative assumptions generally involves a critical review of the scoring rubrics, the scoring procedures, and the procedures for administering the assessment.

The second inference, the generalisation from observed scores to the universe scores, assumes that observed scores are based on random (or representative) samples of observations from the universe of generalisation. Kane and colleagues see this inference as a statistical generalisation from the observed score based on actual performance over the universe of generalisation, which includes performances on tasks exchangeable with those in the assessment. The relevant evidence is collected in generalisability and reliability studies, which indicate consistency of scores across samples of observations.

Evidence on the generalisability and reliability of performance assessments has generally not been encouraging, particularly in terms of the sampling error associated with the use of a relatively small number of tasks (e.g., Brennan & Johnson, 1995, cited in Kane, Crooks, & Cohen, 1999; Swanson, Norman, & Linn, 1995). The variability associated with raters and occasions has been found to be modest, but the variability

associated with tasks, particularly the person-task interaction, has been found to be substantial. The latter finding is of particular concern as it suggests that the scores obtained on a sample of tasks cannot be safely generalized beyond that specific set of tasks. Consistency can be improved by increasing the number of independent observations in each sample of performance, but this is often not possible due to the required resources, complexity and length of individual performance tasks. Improvements in generalisability can also be achieved by standardising task characteristics and test administration procedures. Kane, Crooks, & Cohen (1999) suggest that standardisation of at least those conditions of observation that are fixed in the criterion/target domain should be ensured. In contrast, standardising conditions that are free to vary in the target domain is more problematic and can lead to loss of credibility in the inferences to the criterion domain and to trivialising the assessment rather than improving it (Schuwirth, et al., 2002).

Of course, attempts to increase generalisability by increasing the number of tasks, or introducing a large number of short tasks instead of fewer more comprehensive ones, might be a problem if in the process the sight is lost of what it is that we want to elicit and assess, and of the thinking and behaviour that are deemed to be important aspects of the construct (Alastair Pollitt, personal communication). It can be argued that the most important thing in assessment is to ensure that 'students' minds are doing the things we want them to show us they can do' (Pollitt & Ahmed, 2007) or that the interaction between the student and the task is appropriate for the construct under assessment (Bachman, 1990). 'What matters is not the number of tasks but their quality – a few that truly elicit all the thinking and behaviours that are the essence of competence in the real-world are better than many, each of which captures just a few with much irrelevant variance thrown in' (Alistair Pollitt, personal communication).

In this context, it is perhaps also relevant to point out possible different implications for generalisation that assessment by work-based assessors might have compared to assessment by visiting assessors (cf. footnote 6, p. 13). In some cases (e.g., in Hairdressing), work-based assessors might both train and assess candidates, and also work alongside them on daily basis, and this whole process might inform their judgement, to the extent that even before the formal occasion of summative assessment, the assessor could already be sure that the candidate would pass. In situations like this, generalisability might be less of an issue, as it is not just the summative, but also the formative assessment (and presumably various informal occasions on which the work-based assessor might see evidence of a candidate's competence) that contribute to the final judgement (cf. Messick, 1994, p. 15). This would, conceivably be more of a problem with visiting assessors, who see candidates much less frequently, and often rely on views of witnesses and others to inform their judgement.

In view of this discussion, we believe that it is important to bear the issues around generalisability in mind and include them in the framework (though perhaps not necessarily think about them in strictly statistical terms – see footnote 23), if nothing else then as a reminder that there needs to be an attempt in any kind

of assessment, and so in WBOA, to ensure that the conclusions based on assessment outcomes can go beyond those specific assessment occasions.

When it comes to extrapolation, everything hinges on the similarity between the universe of generalisation and the target domain. If no major differences are found (e.g., in the nature of performances captured in assessment vs. those in the target domain, etc.), the extrapolation is likely to be accepted. Kane and colleagues note that, in practice, the argument for extrapolation is likely to be a negative one in that a serious effort is made to identify differences between the universe of generalisation and the target domain that would be likely to invalidate the extrapolation.

High similarity between universe of generalisation and target domain may be achieved by using high-fidelity tasks in assessment, and this is often the basis for the validity claims in performance assessment. However, as discussed above, since these tasks are often time consuming and necessarily include only a small number of performances, the generalisability from this small sample to a broadly defined universe of generalisation may be quite undependable. Kane, Crooks and Cohen see this as a trade-off: extrapolation can be strengthened at the expense of generalization by making the assessment tasks as similar to those in the target domain as possible, or generalisation can be strengthened at the expense of extrapolation by employing larger numbers of tasks – or, possibly, by ensuring that the tasks are as effective as possible in making observable the abilities and skills that are essential for making appropriate decisions about candidates (see above). In either case, the goal should be to achieve relevance without sacrificing too much reliability/generalisability (Kane, Crooks, & Cohen, 1999, p. 12).[23] Indeed, it is their view (p. 15) that for performance assessments, special attention should be given to the generalisability of results over raters, tasks, occasions, etc., as this seems to be the weakest part of the interpretive argument for this assessment type. In contrast, extrapolation should be more of a focus for objective tests.

As will become apparent later in the report, we have used the SCORING – GENERALISATION – EXTRAPOLATION structure of the interpretive argument proposed in (Kane, Crooks, & Cohen, 1999) as the basis for our validation framework. However, the framework was also informed by several other sources, which we review in the following section.

Review of existing argument-based validation frameworks

In developing our validation framework, we considered several previous studies that proposed different argument-based validation frameworks, for different assessment contexts. We reviewed the following

[23] But see Moss (1994) for an alternative view. Also, (Linn, Baker, & Dunbar, 1991, p. 16) suggest that standardisation should not be primary as long as 'acceptable levels are achieved for the particular purpose of assessment.' Kane (2006) quotes Eisner (1991, p. 203), who points out that sampling assumptions are rarely satisfied, and therefore, 'inferences are made to larger populations, not because of impeccable statistical logic, but because it makes good sense to do so'.

studies: Crooks, Kane, & Cohen (1996); Wools, Eggen, & Sanders (2010); Chapelle, Enright, & Jamieson (2010) and Shaw, Crisp, & Johnson (2012). The core of the interpretive argument is similar across all of the frameworks proposed in these studies and involves scoring, generalisation and extrapolation inferences (cf. Kane, Crooks, & Cohen, 1999 and Kane, 2006)). However, there are some important respects in which they differ. Sometimes, the differences are due to different contexts for which these various frameworks were proposed, while sometimes they are a matter of emphasis, reflecting the authors' views of the relative importance of certain aspects of the assessment process.

Crooks, Kane, & Cohen (1996) proposed a framework for validation that follows the stages of the assessment process rather than being based on an interpretive argument consisting of inferences and assumptions. The assessment process is depicted as a chain of eight linked stages:

ADMINISTRATION ⇨ SCORING ⇨ AGGREGATION ⇨ GENERALISATION ⇨ EXTRAPOLATION ⇨ EVALUATION ⇨ DECISION ⇨ IMPACT

There are a number of threats to validity that are associated with each of these stages, and the authors suggest that evaluation of validity should involve careful consideration of these threats.

This framework references the argument-based approach of (Kane, 1992), but does not formalise it in terms of argument structure a la Toulmin; it is proposed to be complementary to Kane's approach, but also more practical in that it identifies specific flaws which can occur in the interpretation and use of assessment scores. The authors emphasise that the threats to validity that they identified in the article are not comprehensive, and could differ depending on assessment type and context.

Their chain metaphor is useful as it emphasises that validity is limited by the weakest link and is similar to the notion of the chain of inferences in the argument-based approach. Nevertheless, they acknowledge that different assessment purposes can imply different validity emphasis, and that the level of risk to validity associated with each link may vary depending on purpose. This has implications for validation studies, as depending on the assessment purpose, the emphasis in evidence collection may change. (Kane, Crooks, & Cohen, 1999) echo this view, and note that different assessment types, as well as assessment purposes, have different risks associated with different links (or inferences). For instance, for performance assessments the weakest part of the interpretive argument is likely to be generalisation, whereas extrapolation is more of a problem for objective tests.

Lessons learned:

We considered the authors' attempt to provide more concrete guidance for validation efforts by identifying potential specific threats to validity useful. We therefore took a similar approach, and associated the inferences and assumptions in the interpretive arguments with relevant threats to validity identified from the literature on WBOA (as well as some mentioned by Crooks, Kane, & Cohen (1996). These threats to validity are introduced into the framework as potential rebuttals for the assumptions that underlie the inferences in the interpretive argument.

In designing our framework, we felt that Crooks, Kane, & Cohen's (1996) model was useful in terms of its emphasis on separating out administration from scoring and generalisation. On the other hand, we thought that validity issues around evaluation and decision are actually integral to scoring in WBOA, as there is often no clear separation of scoring and decisions in this assessment method. Therefore, some of the threats to validity relevant to these two links were subsumed under the scoring inference in our framework.

Chapelle, Enright, & Jamieson (2010) evaluated the differences between Kane's argument-based approach to validation and that described in the Standards (AERA, APA, & NCME, 1999). In their article, they summarised the original attempts to use the approach proposed in the Standards to develop a validity argument for TOEFL[24], and explained why they chose the argument-based approach as more helpful instead. The article usefully summarises the interpretive and validity argument that was developed for TOEFL (based on Chapelle, Enright, & Jamieson, 2008), and here we briefly describe it, and point out the main aspects that were considered useful to inform our framework for WBOA. The interpretive argument in this framework consists of six inferences with accompanying warrants and assumptions. The inferences are:

DOMAIN DESCRIPTION ⇨ EVALUATION ⇨ GENERALISATION ⇨ EXPLANATION ⇨ EXTRAPOLATION ⇨ UTILISATION

These authors started from domain description following Kane's observation that

> *…if the test is intended to be interpreted as a measure of competence in some domain, then efforts to describe the domain carefully and to develop items that reflect the domain (in terms of content, cognitive level, and freedom from potential sources of systematic errors) tend to support the intended interpretation* (Kane, 2004, p. 141).

[24] Test of English as a Foreign Language™

They also introduced explanation between generalisation and extrapolation, following (Kane, 2001)'s observation that, though this inference is not necessary in general, it is useful in the cases where a theoretical construct is used in score interpretation, which is the case in language assessment. Their utilisation inference is made on the basis of assumptions that the meaning of the test scores is clearly interpretable by admissions officers, test takers, and teachers (i.e. relevant stakeholders), and that the test will have a positive influence on how English is taught.

In structuring their interpretive argument, the authors use the notion of warrant in addition to assumptions underlying the inferences. A warrant can be defined as a law, a generally held principle, rule of thumb, or established procedure (Chapelle, Enright, & Jamieson, 2010, p. 7). These rest on assumptions, for which, ultimately, backing or rebuttal needs to be provided as part of the validity argument.

Lessons learned:

With Chapelle, Enright, & Jamieson (2010), we also consider the domain description as important and have included it in our framework as the first inference. In addition, we follow these authors in using the term Utilisation to denote the final inference. However, we use it to refer to the contribution of information based on WBOA to qualification-level decisions about candidates given that WBOA contributes only part of the evidence of competence.

The authors' use of both warrants and assumptions to structure the interpretive argument was considered to overly complicate the framework, especially considering that the framework should ultimately be expressed in a way that will be suitable for operational use. Thus, only assumptions are used in our framework.

Shaw, Crisp, & Johnson (2012)'s framework, proposed for validation of traditional A-level examinations, also mainly draws on the thinking of Kane (2001) as well as partly on the framework developed by Chapelle, Enright, & Jamieson (2010). This framework involves a list of inferences to be justified as indicated by a number of linked validation questions. They propose that for each question various data should be gathered to provide 'evidence for validity' and to identify any 'threats to validity'. The inferences included in this framework are the following:

CONSTRUCT REPRESENTATION ⇨ SCORING ⇨ GENERALISATION ⇨ EXTRAPOLATION ⇨ DECISION-MAKING

The emphasis in designing this framework was on making it accessible for operational users. This perhaps accounts for the lack of detail in the way the interpretive argument is phrased (for instance, no explicit statements of specific assumptions underlying each inference, unlike in (Chapelle, Enright, & Jamieson, 2010)). However, these assumptions are implicit in the methods proposed for evidence collection, which are linked with the validation questions.

Lessons learned:

The focus on operational use and avoidance of technical language in this framework are consonant with our intentions of creating a framework for operational application. However, at this stage, we propose to retain the detailed specification of the interpretive argument, especially as the accompanying validation methods are proposed only tentatively and require further research.

Wools, Eggen, & Sanders (2010)'s validation framework was designed for competence-based assessments. In their article, they illustrate the application of their framework using the example of a competence-based driver assessment, which, in their view, has great resemblance with performance assessment in vocational education. This validation framework also adopts the argument-based approach to validation. The proposed interpretive argument consists of the following chain of inferences that are made to translate a performance on a task into a decision on someone's abilities or competence.

SCORING ⇨ TEST DOMAIN ⇨ COMPETENCE DOMAIN ⇨ PRACTICE DOMAIN ⇨ DECISION

Unlike in other frameworks described above, here the extrapolation inference includes two steps, i.e. extrapolation from test domain to a 'competence domain' and then from there to a 'practice domain'. The competence domain entails 'an operationalization of competence that is being measured'. The practice domain represents real-life situations the candidates may be confronted with in their future professional life. This framework closely follows Toulmin's structure of argument (Toulmin, 1958) and uses it as a model to present inferences within the interpretive argument.

Lessons learned:

The aspect of this framework that was of most interest for us is the distinction that is made between competence domain and practice domain. In our discussion of the notion of competence earlier in the report, we emphasised the problems around attempts to produce universally accepted and comprehensive definitions of competence for specific occupations, and how the competence domain (i.e. the criterion domain) is itself a construct. Therefore, it seems useful to have a distinction between the two domains reflected in the framework. However, we see the distinction that these authors are making as a variation of the distinction between Chapelle, Enright, & Jamieson's (2010) DOMAIN DESCRIPTION, i.e. competence domain, and EXTRAPOLATION, i.e. practice domain, which we adopted in our framework. We noted earlier that we consider the inference related to decision (about a candidate's competence) as integral to scoring in the context of WBOA.

Summary

The following points summarise the aspects of the reviewed frameworks that we considered useful/informative to include into ours:

- Pre-identified threats to validity used in the development of the interpretive argument (based on Crooks, Kane, & Cohen, 1996).
- Separate administration inference introduced (based on Crooks, Kane, & Cohen, 1996).
- Separate domain description inference introduced (based on Chapelle, Enright, & Jamieson, 2010)
- We use UTILISATION to denote the final inference. However, we use it to refer to the contribution of information based on WBOA to qualification-level decisions about candidates.
- Simplifying the structure of the interpretive argument to include only assumptions, without explicitly stating the warrants.
- The focus on operational use and avoidance of technical language in Shaw, Crisp, & Johnson (2012) are consonant with our intentions of creating a framework of WBOA for potentially wide operational application. However, we believe that retaining a detailed specification of the interpretive argument would be beneficial at this stage.

The next section presents the interpretive argument (or validation framework) for WBOA.

The interpretive argument for validation of WBOA

The final structure of the interpretive argument proposed here for WBOA contains the following inferences:

DOMAIN DESCRIPTION ⇨ ADMINISTRATION ⇨ JUDGEMENT ⇨ GENERALISATION ⇨ EXTRAPOLATION ⇨ UTILISATION

For each inference, we have identified several assumptions on which the interpretation of WBOA results rests. These assumptions would need to be evaluated in a validation exercise in order to establish their plausibility and thus validity of WBOA. To help with focusing validation efforts, we have attempted to identify some key threats to validity/possible rebuttals associated with each inference and assumption – these are presented alongside the relevant assumptions.

The interpretive argument for WBOA presented in Table 3 is the final version, honed after we carried out the fieldwork to evaluate the framework and to help us decide if the list of assumptions and threats to the validity of WBOA that we identified based on the literature review was fairly plausible and comprehensive.[25] The results of our fieldwork are presented as two separate major sections starting on pages 53 and 87 respectively. The fieldwork also gave us a chance to investigate the potential of some qualitative methods for collecting evidence of validity of WBOA. We have also explored the potential of providing quantitative evidence for validity of WBOA interpretations, which is presented in the section on Quantitative evidence of validity in WBOA at p. 87.

The complete framework, including suggestions for some potential methods that might be used in the validation of WBOA is presented on pp. 108-112.

Given that this validation framework is developed for a specific assessment method rather than a whole qualification, the assumptions related to the DOMAIN DESCRIPTION, EXTRAPOLATION and UTILISATION have a narrower scope than they would have if the framework was intended for validation of whole qualifications. For instance, the UTILISATION inference in our framework refers to the use of WBOA results for making overall qualification-level decisions, together with the results of other assessment methods used within a specific qualification. DOMAIN DESCRIPTION and EXTRAPOLATION should be understood to refer to a narrower target domain than they would in the case of whole qualifications, consistent with the scope of WBOA.

[25] We do not present the previous version for reasons of space.

Table 3: The interpretive argument for WBOA validation

INTERPRETIVE ARGUMENT		
INFERENCE	**ASSUMPTIONS UNDERLYING THE INFERENCE**	**THREATS TO VALIDITY/ POTENTIAL REBUTTALS**
1. DOMAIN DESCRIPTION/theoretical construct definition		
	1. NOS or comparable sets of standards addressed by WBOA representative of relevant job role and of the relevant definition of competence (criterion/target domain) – construct defined appropriately	– NOS problematic as descriptors of criterion domain/construct definition – Lack of consensus among stakeholders on what the important aspects of competence are
	2. LOs and ACs appropriately reflect the construct	– LOs and ACs do not represent the construct and the criterion domain appropriately
	3. LOs and ACs appropriate for assessment by WBOA	– Alternative assessment methods required – Simulation (or witness statement, report, other evidence types) used instead of WBOA for inappropriate reasons – or vice versa
	4. Appropriate assessment tasks/situations used for assessment	– No opportunity to collect evidence or demonstrate competence for certain AC and LO in the workplace setting – Cognitive/psychomotor/etc. demands of tasks/situations do not match LOs and ACs – Demands too variable across candidates and occasions
	5. Statistical characteristics of tasks appropriate	– Task difficulty too variable (e.g., situation/customer complexity) especially relevant where it is left up to the assessor or candidate to select which tasks will be observed for summative assessment)
2. ADMINISTRATION		
	1. Learners are sufficiently made aware of the properties of WBOA assessment procedures and of the criteria and standards applied in assessment	– Learners lack understanding of assessment procedures, criteria and standards – Learner behaviour adversely affected by observation process – Assessment anxiety (less likely if assessment is repeatable, as is often the case with WBOA); possibly higher if assessed by peripatetic rather than work-based assessor
	2. Task administration conditions are appropriate for providing evidence of targeted competence	– Inappropriate assessment conditions
	3. Assessor conduct appropriate and consistent with relevant guidance and best practice	– Inappropriate help by assessors – Personality clashes/bias
	4. Relevant verification procedures effective	– Verification ineffective or does not address administration issues

3. JUDGEMENT		
	1. Observation tools/checklists are appropriate and used appropriately for recording and judging evidence of relevant competence and promoting reliable judgement	– Design and wording of observation tools/checklists does not support reliable judgement/reliability analyses – Observation tools/checklists do not capture important qualities of task performance – Observation tools/checklists not used where recommended, or used inappropriately
	2. Assessors are sufficiently trained in observation techniques and carry out observation according to best practice	– Assessors not trained in observational techniques – Assessors fail to observe the performance appropriately, miss important aspects, intrusive etc.
	3. Assessors apply relevant criteria/standards and observation tools appropriately and consistently	– Assessor inconsistency or lack of understanding and application of criteria and standards (+ possible differences between work-based and visiting assessors) – Inappropriate weighting given to different aspects of performance – Positive or negative bias – Severity/leniency – difficult to check and adjust for differences here
	4. Assessors require appropriate amount of evidence in order to reach a decision	– Assessors over/under assess (+ possible differences between work-based and visiting assessors)
	5. Decisions based on WBOA appropriate	– Problems with classification consistency at WBOA level (too many false positives; false negatives perhaps not a major problem as new assessment opportunities are always available – problem: over-assessment) – Assessor decisions affected by funding pressures (+ possible differences between work-based and visiting assessors)
	6. Relevant verification procedures effective	– Verification procedures too infrequent (lack of investment) – Verification procedures focus on paperwork and procedural compliance rather than judgement and assessor standardisation

4. GENERALISATION		
	1. The sample of observations is representative of the universe of generalisation, OR assessment focuses on what is proven to be important to stakeholders (as indicated by the construct)	– Inappropriate sampling of tasks from assessed domain – Standardisation of tasks; conditions of assessment too variable; extreme tasks/situations (unless warranted by construct definition) – Task specificity issues
	2. The sample of observations is large enough to control random error/to enable confident judgement	– Too few observations – Too much variability in the minimum number of observations used – Assessors lack confidence in their decisions (+ possible differences between work-based and visiting assessors)
5. EXTRAPOLATION		
	1. There are no systematic errors that are likely to undermine the extrapolation at the level of WBOA (↺ DOMAIN DESCRIPTION)	– Parts of criterion domain that should be assessed by WBOA not assessed or given little weight (construct under-representation) – Conditions of assessment too constrained/task fidelity issues – Failure to make explicit any legitimate limitations of the assessment that should limit extrapolation
6. UTILISATION		
	1. WBOA results considered useful by relevant stakeholders for informing decisions about qualification-level competence	– Stakeholders not satisfied about contribution of WBOA results to overall decisions about candidate competence
	2. WBOA positively impacts the learning and instruction process	– Inappropriate assessment strategy dominates the learning and instruction process – negative washback – No feedback or inappropriate feedback due to lack of time in the workplace, lack of training/awareness of assessors in this respect, etc. – reduced chance of improving future performance

Evaluating the interpretive argument for WBOA (and collecting opinion evidence) through views of key practitioners

It is clear from the previous discussion, and from the complexity of the interpretive argument proposed based on reviewing the relevant literature, that there are several aspects of WBOA that might present both potential threats and boons for its validity and the validity of those qualifications where it is used to assess candidates. We therefore considered it important to try and to understand more thoroughly how WBOA functions in the field, by collecting the views of some of the key participants in the process. This also informed our recommendations as to which of the various aspects of the method need to be investigated more or less thoroughly as part of a validation process.

Given the time and resource constraints on this study, we decided to focus only on the views of assessors, IQAs and QCs as the key practitioners in charge of delivering and quality assuring WBOA. In order to collect their views, we designed a questionnaire and an interview schedule which addressed several key aspects of our validation framework. In addition, we saw this as an opportunity to test the questionnaire and interview methods in terms of how far they could be used as suitable opinion gathering methods for collecting validity evidence about WBOA.

Consulting WBOA practitioners using a questionnaire

About the questionnaire

Questionnaire items were written to test out assumptions and threats to validity identified in the validation framework and addressed the span of inferences within the framework. (Appendix 4 shows question texts, the parts of the framework that they addressed and a summary of the findings for all quantitative items.)

We have not conducted a formal analysis – for example of the internal consistency of the data set generated from this questionnaire. But in general, we observe results on all questions to be largely coherent and meaningful. This augurs well for this approach as part of a 'validation toolkit' for WBOA. The exceptions to this (see Table 6) are the sectors into which respondents categorised themselves, and several items (noted in Appendix 4 which had large numbers of 'blank' responses.

The former case (confusion over sector allocation) may speak to the extent to which data on VQs can be overly-complex and confusing to those of us who do things other than design database systems. This has been mentioned elsewhere in this report (see the Preponderance of WBOA in VQs section at p. 10). Presumably, this would be less of an issue were this instrument to be used 'for real' to validate a specific qualification; users and questionnaire designers would all know and agree which sector they came from.

Where there are large numbers of non-responses to an item, questionnaire designers should review the particular items concerned and seek to amend and ameliorate them.

On occasion, the question-to-framework mapping causes one to ponder; for example, questions 6 and 7 are classic validation questions. They address construct irrelevant variance and construct under-representation, respectively. However, matching these questions to the framework was not straightforward. Construct-irrelevant variance seemed to fit best in DOMAIN DESCRIPTION; however, its 'sibling'/converse construct under-representation could fit into extrapolation as well as DOMAIN DESCRIPTION. This may well be reasonable; however, it is also arguable that the very general categories in the validation framework may be hard to operationalise on occasion.

The questionnaire was administered online and the respondents were given two weeks to return the questionnaire. The questionnaire was communicated in three different ways:

1) to a selection of the City and Guilds Viewpoint panel[26]
2) by posting the link in the Consultant Update email[27]
3) by posting the link on the 15 most visited City and Guilds Smart Screen Tutor discussion boards.[28]

Achieved sample of respondents

Nearly 500 respondents took part in the survey. About 60 per cent of these respondents were IQAs, just over 20 per cent were QCs and the remainder were assessors. Some of the respondents performed more than one of these roles. In those cases, the questions they responded to were based on the most senior role. The respondents had substantial experience, 'more than five years' being the most endorsed category for each type of respondent. Details can be seen in Table 4.

Table 4: Number of years in role pivoted against role

Number of years	Assessor	IQA	QC	Total
Less than 1 year	4	22	5	**31**
1 to 2 years	5	55	19	**79**
3 to 5 years	26	12	4	**42**
More than 5 years	40	220	84	**344**
Total	**75**	**309**	**112**	**496**

The respondents were employed in various types of organisation. Amongst assessors and IQAs, the largest number of respondents worked for colleges and private training providers (PTPs) (only 5% of assessors and

[26] ViewPoint is a City and Guilds online stakeholder panel which currently has 1622 members who have opted in to be contacted for their views on a variety of topics, typically by responding to surveys. We sent the questionnaire to 877 members that belonged to the categories Assessor/ Verifier (282) and Tutor/ Trainer/ Teacher/ Lecturer (595).

[27] Consultant Update is a monthly email sent to all City and Guilds QCs. It contains major news items, as well as reminders and guidance on City and Guilds policies and systems.

[28] SmartScreen is a City and Guilds online portal for tutors, assessors and learners that provides unit-specific support materials and tutor forums to complement the delivery and learning experience. It is used by over 2000 City and Guilds centres.

11% of IQAs were work-based[29]). Amongst QCs the largest group was self-employed or freelance. This reflects the general working practice within that group.

Table 5: Employer type pivoted against role

Employer type	Assessor	IQA	QC	Total
College	48	176	32	**256**
Private training provider	14	76	23	**113**
Self-employed/freelance	4	10	49	**63**
Workplace	7	35	2	**44**
Other	2	12	6	**20**
Total	**75**	**309**	**112**	**496**

Respondents came from a range of occupational sectors as summarised in Table 6. Many respondents said that they worked in more than one sector, and there were a very large number of combinations of sectors.

Table 6 shows the most prevalent combinations of sectors that were indicated by respondents. The biggest groups of respondents stated that they worked in: Health, Public Services and Care; Construction, Planning and the Built Environment, and Education and Training (or those sectors in combination with some other sector).

Table 6: Main sectors cross-tabbed against role

Sector							Role			
Business, Administration and Law	Construction, Planning and the Built Environment	Education and Training	Engineering and Manufacturing Technologies	Health, Public Services and Care	Leisure, Travel and Tourism	Retail and Commercial Enterprise	Assessor	IQA	QC	Total
				Y			8	72	14	**94**
	Y						24	39	14	**77**
		Y					8	30	15	**53**
						Y	9	28	4	**41**
			Y				7	21	9	**37**
		Y		Y			2	13	2	**17**
		Y				Y	1	10	4	**15**
	Y		Y				1	10	1	**12**
					Y			6	3	**9**
	Y	Y					1	7	1	**9**
Y							2	5	1	**8**
Y		Y					1	5	2	**8**
		Y	Y					5	2	**7**
									Sub-total	**387**
									All other combinations	**109**
									Grand total	**496**

[29] Although this finding suggests that the proportion of work-based assessors and IQAs is higher (though still quite low compared to college and PTP based) than the results of the IER (2012) report indicate (cf. footnote 6), an appropriate sampling strategy is required to confirm this.

Main findings from the questionnaire

In this section, we present analyses of several example items, particularly where significant findings were observed. Full analysis of all quantitative and qualitative items was carried out, but space precludes the reporting of all item outcomes in this report (the quantitative results that are not reported in the main text are summarised in Appendix 4). Here, we attempt to exemplify a range of analysis types, and some of the key issues that we will refer to in the discussion at the end of this section.

Question 8 concerned the extent to which respondents had taken different types of action when confronted by an inappropriate or missing task (this question was related to the previous two questions asking respondents if WBOA assesses tasks that are rare in the workplace, and if it covers all the tasks that are required for the relevant role (see Appendix 4). The results for question eight are summarised in the table below.

Table 7: Action taken for inappropriate tasks cross-tabbed against role

Action taken						Role			Total
Informed relevant IQA/IV	Informed relevant QC/EV	Informed Awarding Organisation directly	Informed relevant SSC/SSB directly	Other	No action taken	Assessor	IQA	QC	
						41	179	70	**290**
					Y	14	31	5	**50**
Y						13	16	12	**41**
	Y						36	1	**37**
		Y				3	9	7	**19**
Y	Y					2	9	3	**14**
				Y		1	10		**11**
	Y	Y				1	8		**9**
Y	Y	Y					4	5	**9**
Y		Y						7	**7**
Y	Y	Y	Y				1	2	**3**
	Y	Y	Y				2		**2**
			Y				1		**1**
	Y		Y				1		**1**
	Y	Y		Y			1		**1**
Y				Y			1		**1**
					Total	**75**	**309**	**112**	**496**

This item had 290 missing responses (first data row of the table above) despite there being a 'I haven't taken any action' option. The three biggest combinations of responses were 'I haven't taken any action', 'I have informed the relevant IQA/IV' and 'I have informed the relevant QC/EV'. Relatively few respondents had contacted either an awarding organisation or an SSC directly.

Question 9 had three sub-items concerning respondents' perceptions of the feedback process when tasks supposed to be assessed by WBOA were inappropriate or missing. For exemplification we present results from sub-items 9b and 9c.

Sub-item 9b asked whether respondents felt that relevant people had 'heard' them when they reported a missing or inappropriate task. In Table 8 we show how many respondents from each role type responded in a particular way to each question type. Although there were large numbers of null responses to this sub-item, most respondents who gave a definite answer felt that relevant people did hear them when they escalated a concern.

In respect of item 9b the chi-squared statistic suggested that there was a significant difference in the responses of the three role types. Table 8 also shows adjusted residuals for each observed count – these numbers indicating which parts of the contingency table were most likely to have caused the chi-squared statistic to have returned a significant value (see 'Behaviour of assessors in centres' section – below at pp. 101ff). Adjusted residuals of greater than absolute two are shaded in the table to indicate those cells that are tending to make chi squared significant.

Table 8: Whether relevant people have heard me, cross-tabbed against role, with adjusted residuals shown

			I know that the relevant people have heard me						Total
			Strongly disagree	Tend to disagree	Neither agree nor disagree	Tend to agree	Strongly agree	*No response*	
Role	Assessor	Count	6	10	5	7	6	*41*	**75**
		Expected Count	3.2	4.2	5.9	10.6	7.1	*44.0*	**75.0**
		Adjusted Residual	1.8	3.1	-.4	-1.3	-.5	*-.8*	
	IQA	Count	11	13	30	47	27	*181*	**309**
		Expected Count	13.1	17.4	24.3	43.6	29.3	*181.3*	**309.0**
		Adjusted Residual	-1.0	-1.8	2.0	.9	-.7	*-.1*	
	QC	Count	<5	5	<5	16	14	*69*	**112**
		Expected Count	4.7	6.3	8.8	15.8	10.6	*65.7*	**112.0**
		Adjusted Residual	n<5	-.6	n<5	.1	1.2	*.7*	
Total		**Count**	**21**	**28**	**39**	**70**	**47**	***291***	**496**
		Expected Count	**21.0**	**28.0**	**39.0**	**70.0**	**47.0**	***291.0***	**496.0**

The two cells that have consequential residual values are: assessors who tend to disagree with the statement and IQAs who neither agree nor disagree. In the former case, the observed count was rather higher than one would have expected if that cell had been consistent with the overall pattern. In the latter case, rather more IQAs had neutral views than would have been expected given the overall pattern.

Sub-item 9c asked respondents whether relevant people took necessary action once informed about an inappropriate or missing task. The results for this sub-item are displayed in Table 9.

Table 9: Whether relevant people are perceived to take necessary action cross-tabbed against role

		I know that the relevant people take necessary action						Total
		Strongly disagree	Tend to disagree	Neither agree nor disagree	Tend to agree	Strongly agree	*No response*	
Role	Assessor	8	7	5	8	6	*41*	**75**
	IQA	14	20	31	37	26	*181*	**309**
	QC	5	4	5	18	11	*69*	**112**
Total		**27**	**31**	**41**	**63**	**43**	***291***	**496**

As with other parts of item 9, the most salient feature in this data set is the large number of missing responses. Following that, the majority of respondents agree with the statement to some extent. There was no significant difference between role types.

Question 11 asked respondents to describe the factors that they attended to when making judgements. Their responses to this item are summarised in Table 10.

Table 10: Action taken when forming judgement cross-tabbed against role

Action taken					Role			Total
Take note of important features of candidate's performance	Refer to assessment criteria first and then make an overall judgement	Form an overall judgement and then refer to the assessment criteria	Form an overall judgement without referring to assessment criteria	Other	Assessor	IQA	QC	
Y	Y				37	120	40	**197**
	Y				13	52	25	**90**
Y	Y	Y			6	42	8	**56**
Y		Y			4	38	13	**55**
					6	20	6	**32**
		Y			3	11	7	**21**
Y					4	3	4	**11**
	Y	Y				7		**7**
Y	Y			Y	1	3	3	**7**
Y	Y	Y	Y			5	1	**6**
Y			Y			2	1	**3**
Y		Y		Y		1	2	**3**
				Y	1		1	**2**
			Y			1		**1**
	Y		Y			1		**1**
Y				Y			1	**1**
Y		Y	Y			1		**1**
Y	Y		Y			1		**1**
Y	Y	Y	Y	Y		1		**1**
				Total	**75**	**309**	**112**	**496**

The four highest combinations of responses were (each with over 50 responses):

- 'Take note of important features of the candidate's performance' AND 'Refer to the assessment criteria first and then make an overall judgement'
- 'Refer to the assessment criteria first and then make an overall judgement' alone
- 'Take note of important features of the candidate's performance', 'Refer to the assessment criteria first and then make an overall judgement' AND 'Form an overall judgement and then refer to the assessment criteria'
- 'Take note of important features of the candidate's performance' AND 'Form an overall judgement and then refer to the assessment criteria'

Relatively few respondents from any role group admitted to forming holistic judgements without referring to the assessment criteria.

Question 14 asked all role groups whether they thought that the number of observations required in assessment strategies was appropriate. This issue belonged to the GENERALISTION inference in the framework. Findings for question 14 are expressed in Table 11.

Table 11: Perceptions of sufficiency of observations cross-tabbed against role

		In general, do you think that the number of observations required or suggested by the relevant assessment strategy is enough to make a confident judgement?				Total
		Yes, number of observations is about right	No, more observations are needed	No, fewer observations are needed	*No response*	
Role	Assessor	47	17	5	*6*	**75**
	IQA	211	55	13	*30*	**309**
	QC	81	16	4	*11*	**112**
Total		**339**	**88**	**22**	***47***	**496**

There was a large majority of the respondents that thought the number of observations was appropriate, although there was a sizeable minority that thought that more observations were needed. There were no significant differences between the role types in respect of this issue.

Those who stated that more or fewer observations were needed were asked to explain why the number of observations was not right. Respondents tended to state that more observations were required to gather sufficient evidence, confirm competence and ensure consistency of performance over time; particularly when units were technical, many criteria needed to be covered, and/or the range was vast. It was also suggested that one or two observations were often insufficient in achieving this and that one-off demonstrations could be misleading. An IQA stated:

> *Most assessors think 1-2 holistic observations suffice, however for continuous, valid assessment I feel a minimum of 3-4 observations are required. Too often assessment is driven by time constraints and not quality.*

A QC wrote that the assessment criteria specifying minimum observations tended to be interpreted as the actual number by assessors, IQAs and candidates. An IQA noted that the assessor should know when further observations would be beneficial.

In contrast, some of the respondents who favoured fewer observations claimed that: too many observations were required for some units; work was sometimes duplicated; and evidence could be gathered for more than one unit from one observation. It was suggested that observations should not be repeated if all the criteria were met and the evidence needed was gathered. Fewer observations were also favoured in situations where it was difficult to get naturally occurring tasks such as perms in hairdressing. In addition, travelling a vast distance to carry out an observation was seen to be time consuming and costly, and possibly an inconvenience to the employer, in which case the alternative of obtaining an expert witness statement from the candidate's supervisor was seen as preferable.

However, some said that having more or fewer observations would depend on aspects such as the qualification, the outcomes, the range to be covered, the task being observed, the complexity of learning within a unit, the learner's ability, and the experience of the assessor.

Question 16 was concerned with the issue of which standards assessors observed in WBOA. As such, it addressed issues within the JUDGEMENT part of the framework. Respondents gave a rating from 1 to 4 to show the importance to which they attached particular types of standards.

Table 12 is made up of two sub-tables. The top table is a raw count of responses (missing excluded) within each ordered category from 1 (most important) to 4 (least important). The bottom table multiplies the number of responses in each category by the category value. Then, the overall total score for each type of standard is divided by the number of respondents to derive a mean score for each type of standard.

Table 12: Importance attached to different types of standards – count of responses and mean score

Type of standard	How important are each of these standards to you when observing a task?				N
	Most important	Count of 'raw score'		Least important	
	1	2	3	4	
The qualification standards	284	106	32	12	**434**
Current sector standards	97	208	81	48	**434**
Your own standards	38	63	112	221	**434**
Your organisation's standards	15	57	209	153	**434**

Type of standard	Most important	'Adjusted score' (Raw score x rating)		Least important	Total	Mean
	1	2	3	4		
The qualification standards	284	212	96	48	**640**	**1.475**
Current sector standards	97	416	243	192	**948**	**2.184**
Your own standards	38	126	336	884	**1,384**	**3.189**
Your organisation's standards	15	114	627	612	**1,368**	**3.152**

The average values for each standard type, and the counts of the numbers of respondents answering in each importance category merit careful analysis. The mean value for the different standard types ascend from 'the qualification standards' through to 'your own standards' – showing the former to be most important and the latter least important. However, – taking 'your own standards' as an example – 101 respondents (nearly a quarter of those responding) rated their own standards as either 1 or 2. This suggests that respondents' own internal standards are an important guide to judgement for many assessors. In addition, nearly 35% of respondents did not give first priority to the qualification standards (cf. similar findings in (Eraut, Steadman, Trill, & Parkes, 1996)). This does not necessarily imply invalidity; we have seen in the review section (pp. 31ff) that judgement is an integrated, value-laden process which cannot be reduced merely to the explicit statements in qualification standards.

Question 19 was made up of two sub-items, both targeting the judgement section of the framework. Sub-item 19a addresses 'false positives' (those who passed, but should not have) and sub-item 19b addresses 'false negatives' (those who did not pass, but should have).

In the case of both sub-items, the chi-squared statistic had a value that was significant – at the 1 per cent as well as the 5 per cent levels. In the tables that follow, we present counts of the number of respondents endorsing particular options on a Likert scale and expected values (given overall proportions) and adjusted residuals to show those cells that are likeliest to have caused the chi-squared statistic value to be significant.

Table 13: Perceived preponderance of 'false positive' and 'false negative' errors in WBOA cross-tabbed against role, with adjusted residuals

			In your experience of work-based observational assessment, how often on a task do candidates PASS but SHOULD NOT.						Total
			Always	Often	Sometimes	Seldom	Never	*No response*	
Role	Assessor	Count	<5	5	7	23	25	*14*	**75**
		Expected Count	.9	3.9	13.5	22.8	23.7	*10.1*	**75.0**
		Adjusted Residual	n<5	.6	-2.1	.0	.3	*1.4*	
	IQA	Count	<5	14	47	89	114	*41*	**309**
		Expected Count	3.7	16.2	55.4	94.1	97.8	*41.7*	**309.0**
		Adjusted Residual	n<5	-.9	-2.0	-1.0	3.2	*-.2*	
	QC	Count	<5	7	35	39	18	*12*	**112**
		Expected Count	1.4	5.9	20.1	34.1	35.5	*15.1*	**112.0**
		Adjusted Residual	n<5	.5	4.2	1.1	-4.0	*-1.0*	
Total		**Count**	**6**	**26**	**89**	**151**	**157**	***67***	**496**
		Expected Count	**6.0**	**26.0**	**89.0**	**151.0**	**157.0**	***67.0***	**496.0**

			In your experience of work-based observational assessment, how often on a task candidates DO NOT PASS but SHOULD have.						Total
			Always	Often	Sometimes	Seldom	Never	No response	
Role	Assessor	Count	<5	<5	9	16	35	14	**75**
		Expected Count	n<5	.9	8.6	26.2	29.0	10.1	**75.0**
		Adjusted Residual	n<5	n<5	.1	-2.7	1.5	1.4	
	IQA	Count	<5	<5	31	101	133	41	**309**
		Expected Count	n<5	3.7	35.5	107.8	119.6	41.7	**309.0**
		Adjusted Residual	n<5	n<5	-1.3	-1.3	2.5	-.2	
	QC	Count	<5	<5	17	56	24	12	**112**
		Expected Count	n<5	1.4	12.9	39.1	43.4	15.1	**112.0**
		Adjusted Residual	n<5	n<5	1.4	3.8	-4.3	-1.0	
Total		**Count**	**<5**	**6**	**57**	**173**	**192**	**67**	**496**
		Expected Count	**<5**	**6.0**	**57.0**	**173.0**	**192.0**	**67.0**	**496.0**

Once again, these tables merit careful consideration. It is a good thing – for those seeking to confirm the validity of WBOA as a method – that large numbers of respondents believe that they have seldom or never witnessed mis-classifications (308 'false positives' and 365 'false negatives'). However, there remain 89 and 57 respondents who were prepared to admit witnessing the respective types of error sometimes. These substantial numbers may be a challenge to the validity of WBOA.

The comparative reactions of role types are also of interest. QCs seem relatively more likely to have seen candidates who pass but should not have sometimes, when compared to IQAs and assessors (cf. the patterns in Eraut, Steadman, Trill, & Parkes (1996). To some extent, this pattern is repeated in the case of false negative errors; there are fewer QCs endorsing the 'never' category than would be expected, given the overall pattern. Conversely, there are more IQAs saying 'never' than one would expect in the lower table.

Assessors, IQAs and QCs who answered 'Sometimes', 'Often' or 'Always' for either part of Question 19 were asked to give examples or common reasons for 'false negatives' and 'false positives'. The more common assessor/IQA-related reasons included:

- misinterpretation of assessment criteria/ standards or application of own standards
- lack of experience or training of assessors in industry and/or assessment
- level of consideration given to past performance/overall candidate ability in making decisions

However, it is worth noting that misinterpretation of the standards or assessment criteria may relate to poorly written assessment strategies and assessment criteria, which was also given by a few respondents as the main reason for candidates not getting the right results. Other assessor/IQA-related reasons included: inadequate IQA sampling; lack of standardisation between assessors and IVs; inadequate recording of information; assessor subjectivity or bias; and poor assessment planning, preparation, review and feedback to candidates.

Similarly, various aspects relating to candidates were identified as reasons for why they DO NOT pass when they SHOULD. One of the common reasons was the candidates not performing to required standards on the assessment day – could be due to lack of preparation, mistakes relating to health and safety (e.g., forgot to wear gloves), exceeding time set for tasks, personal issues or illness, lack of confidence, nervousness, or uneasiness about being observed. Other common reasons were: lack of evidence to prove competence; inconsistency of performance over time; candidate not suited for programme or job; and unfamiliarity of assessment criteria.

In comparison, a common reason given by respondents for candidates PASSING when they SHOULD NOT was pressure from learning providers, centres and management to pass candidates quickly in order to get high achievement rates and meet success targets as these are linked to funding. This was said to be particularly difficult to resist for new and inexperienced assessors.

Other reasons related to the work environment or circumstances during assessments: workplace is uncooperative (e.g., limited time allocation); there is a lack of opportunity, experience or range in workplace to carry out tasks on-the-job or during assessment; and situation is unsuitable (e.g., no clients or children do not want to take part in planned activity).

Question 25 asked the three role types how often they took part in standardisation activities. This item addressed assessor training and consistency, and was most closely linked to the JUDGEMENT part of the validation framework.

The chi-squared statistic was significant at the one per cent level, and hence the contingency table (Table 14 displays observed and expected counts, and adjusted residuals.

Table 14: Frequency of participation in standardisation cross-tabbed against role, with adjusted residuals shown

			How often do you take part in standardisation activities relevant to observation in the workplace? I take part in standardisation activity…							Total
			Every month	Every few months	Once a year	Once every few years	Never	*Not applicable*	*No response*	
Role	Assessor	Count	15	29	7	<5	<5	*6*	*15*	**75**
		Expected Count	14.2	34.6	8.3	2.6	2.3	*1.5*	*11.5*	**75.0**
		Adjusted Residual	.3	-1.4	-.5	n<5	n<5	*4.0*	*1.2*	
	IQA	Count	65	156	28	<5	7	*<5*	*47*	**309**
		Expected Count	58.6	142.7	34.3	10.6	9.3	*6.2*	*47.3*	**309.0**
		Adjusted Residual	1.5	2.5	-1.8	n<5	-1.3	*n<5*	*-.1*	
	QC	Count	14	44	20	12	6	*<5*	*14*	**112**
		Expected Count	21.2	51.7	12.4	3.8	3.4	*2.3*	*17.2*	**112.0**
		Adjusted Residual	-2.0	-1.7	2.6	4.8	1.6	*n<5*	*-.9*	
Total		**Count**	**94**	**229**	**55**	**17**	**15**	***10***	***76***	**496**
		Expected Count	**94.0**	**229.0**	**55.0**	**17.0**	**15.0**	***10.0***	***76.0***	**496.0**

These results show fairly frequent participation in standardisation. 323 out of the 410 respondents who gave a definite answer reported attending standardisation at least every few months. The adjusted residuals suggest that – in the main – it was QCs' different patterns of participation in standardisation from IQAs and assessors which was causing chi-squared to be significant. They were less likely to be in the most frequent category for standardisation, but more likely to experience standardisations a year or more apart. The lack of frequent standardisation in the case of QCs is something that might cause some concern, given that they are the ones that should be monitoring and ensuring standardisation and consistency of assessors and IQAs. This is further explored in the analysis of our interview data.

The analysis of the follow-up open question (question 26) asking all three groups of respondents to specify what standardisation activities they engaged in gave some useful insights into their practice, and was based on 345 responses provided.

The respondents indicated that meetings were amongst the most common methods of standardisation and mentioned the following: team meetings, standardisation meetings/events, moderation meetings, update meetings, network meetings, group and individual discussions, workshops and briefings. However, it was not always entirely clear if the difference between these related to content and coverage or just the terminology being used. These meetings mainly involved assessors, IQAs and/or QCs but might also include one or more of the following: sector skills councils, awarding organisations, employers, learners, their line managers, lecturers, centres and providers as appropriate. According to several respondents, these meetings helped in discussing and sharing good practice. The respondents mentioned a variety of aspects which were discussed as part of the standardisation process, including regulations, requirements, criteria

and their interpretation; assessment planning, methods and approaches; assessment materials; common issues faced by assessors, IQAs and QCs; sampling strategies; and candidate support.

In addition to meetings, it seemed to be common practice for assessors to review work/evidence, make a judgement against a set of standards or criteria, and compare decisions made with each other in a meeting. Although evidence from candidate portfolios tends to be used often for this exercise, other techniques are also used to ensure consistency between assessors i.e. example observations, simulations, scenarios, role plays, demonstrations, case studies, and video recordings of practical tasks.

Another common standardisation activity was observing others during WBOA: assessors observing other assessors (peer/group observations); IQAs observing assessors; and QCs observing IQAs. IQAs and QCs also stated using sampling of assessment decisions and candidate portfolios as well as candidate interviews as a way to ensure consistency between assessors.

In questions 28 and 29 at the end of the survey, the participants were asked to identify the main benefits/advantages and difficulties/constraints of assessing in the workplace using observation (353 responses were provided). Many assessors, IQAs and QCs simply stated that the main benefit or advantage of WBOA was that they could see work being carried out in the actual workplace in a realistic/natural situation in real time. It gave them the opportunity to see how candidates handle pressure, problems and challenges in the workplace as well as planned and unplanned activities, which gives a better indication of the candidate's competence. It provides assessors, IQAs and QCs with authentic evidence, a holistic picture over time, more control over what evidence is captured, evidence for multiple units, and enables candidates to demonstrate their true competence, practical experience and application of knowledge. Some also mentioned that it can be used to assess the candidates' behaviours and attitudes in the real work environment as well as their interaction with colleagues and clients. And when supplemented by questioning, observations can provide a lot of relevant evidence.

Respondents stated that WBOA and the workplace setting for training are advantageous and beneficial for the candidates because it allows them to build their confidence through on-the-job training, practice and feedback. Candidates feel more relaxed in their own environment i.e. the workplace. The method allows them to be observed when they are ready for assessment and the chance to be reassessed. It also gives them the opportunity to give feedback, as mentioned by a QC: 'To have an opportunity of speaking to the learners to ensure that they are comfortable with the process and to listen to any reasons why they are not.' Observational assessment in the workplace can accommodate candidates with special assessment needs and does not disadvantage those who struggle with writing and academic education.

Some IQAs and QCs stated that this was a good way to check consistency and competency of assessors in making judgements, confirm ability of assessors against relevant assessor qualification standards, and

ensure they worked according to the qualification requirements. A few also mentioned that WBOA was objective, valid, authentic and reliable.

Assessors, IQAs and QCs also identified several difficulties and constraints with using WBOA but one of the main reasons related to logistics i.e. organising the observational assessment and ensuring it goes to plan. Respondents stated that it was sometimes difficult to arrange observations at a mutually suitable time for candidates, assessors, IQAs, QCs and clients (where applicable) especially where candidates move to different locations or work different shifts (e.g., evenings or weekends). Furthermore, they mentioned that it was difficult to ensure that relevant work is always available on the assessment day and that a sufficient spread of activities (or the range) is covered to satisfy the criteria e.g., a chef asking the candidate to cook irrelevant dishes. Sometimes assessors are told that the candidates or clients cannot attend the assessment due to shift change, illness, workloads or emergencies at work but by that point the assessor has already made the journey.

According to survey respondents, aspects related to the work environment also cause issues in WBOA. Examples of issues included:

- disruption/distraction from noise, colleagues, and other work activities nearby
- lack of space to observe and listen e.g., some kitchens
- accessibility of workplace e.g., schools and prisons as well as time restrictions for visits
- limited scope to cover range of tasks, lack of clients/customers, and limited resources e.g., tools
- candidates moved to different site/room without notice
- workplace techniques used by candidates do not meet assessment criteria

Another common difficulty or constraint mentioned by survey respondents was the lack of participation and cooperation from employers because they see it as intrusive and a disruption. It was the respondents' opinion that employers (e.g., managers, supervisors or other staff) feel uncomfortable having observers present and may even feel threatened or intimidated. This sometimes means that employers do not: inform candidates about forthcoming tasks that are relevant, give candidates or assessors enough time to complete tasks, train candidates adequately, grant permission for observations, or give access to the observer. It was suggested that good communication from the start would help overcome some of these problems so that employers and managers are well informed about the process.

Furthermore, respondents also pointed out that there were serious confidentiality and sensitivity issues that needed to be carefully considered, for instance in health and social care (dealing with abuse, end of life and first aid). For such circumstances, observations may be inappropriate and/or consent may be needed from senior management, clients, customers, patients, and relatives.

In WBOA, subjectivity of assessors and inconsistency of practice/decision making between them could be an issue. Expert assessors, who are familiar with the standards, are needed to do the job but some complained about the cost, time (travelling, observing, completing paperwork and providing feedback) and volume of work (i.e. number of candidates/observations) relating to observations in the workplace. Others mentioned the pressure to pass candidates to meet funding requirements as a weakness. On the other hand, candidates may get nervous or stressed from being observed, especially if more than one person is observing, and this could impact on their performance on tasks e.g., may make mistakes.

There was also a view that WBOA is artificial or staged for assessment purposes and is not real practice as candidates may act differently in the absence of the assessor. It was also stated that standards dictate what must be observed, which takes the richness and spontaneity away from the observation and that assessment strategies had too many constraints i.e. service times were too rigid and the range was not sufficient. Most of these issues also came up in our interviews, and we return to them and discuss them in more detail in the next section.

Summary

This questionnaire analysis has given us a number of insights into how it might contribute to a validation process, and indeed the nature of validation itself. These are summarised below:[30]

- This form of data collection seems viable for validation purposes in the vocational setting – we managed to capture 500 responses in a reasonable timeframe – and to give reasonably meaningful and robust results.
- Mostly, respondents' perceptions supported the validity of the method. They rarely endorsed categories that implied invalidity. This may suggest bias in the responses, i.e. respondents were reflecting positively on their own practice, or a feeling that they had to give a particular type of answer because of who we were (an awarding organisation), though some of the responses were surprisingly candid despite this. But we can only speculate about such causes. We do know that we got broadly positive answers. A counter-action to such potentially fallaciously positive responses would be to collect answers from other stakeholder groups – candidates or employers, for example.
- There is a question of how to interpret response rates; for example if 20 per cent of respondents say they have 'sometimes' observed bad practice in an assessment method (and 80 per cent say they have not) is this evidence of invalidity? Arguments here become similar to those that have existed for years in reliability research and practice – it is relatively easy to derive a measure, much more difficult to decide on a cut-off point between 'acceptable' and 'unacceptable'.
- Although we recognise the issue mentioned above about deciding on cut-offs for validity, and are reassured that majority response rates provide evidence for validity, we believe that at present, given a thin evidence base on validity of WBOA, the presence of the 'dissenting minorities' should be taken as a serious indication of the need to carry out more validation research in these domains.
- Comparing the responses between role types gave some useful insights – sometimes significant differences were unsurprising – e.g., QCs being somewhat more confident of their ability to spot bad practice than IQAs, however other types of significant difference were harder to explain.
- Relatively few respondents from any role group admitted to forming holistic judgements without referring to the assessment criteria (JUDGEMENT, assumption 3).
- A large majority of the respondents thought the number of observations was appropriate, although there was a sizeable minority that thought that more observations were needed. There were no significant differences between the role types in respect of this issue. Respondents enumerated a number of factors that having more or fewer observations would depend on, such as the qualification,

[30] In this and the summaries of interview data (p. 65) and of the results of our quantitative analyses (p. 83), where a point that we are making, a point that the respondents are making or a statistical indicator relate to relevant aspects of the framework, we tried to note alongside it the inference and assumption number (e.g., DOMAIN DESCRIPTION, assumption 3) from our framework.

the outcomes, the range to be covered, the task being observed, the complexity of learning within a unit, the learner's ability, and experience of the assessor. (GENERALISATION, assumption 2; JUDGEMENT, assumption 4)

- With respect to assessor standards (JUDGEMENT, assumption 3), 'the qualification standards' were most important for a majority of respondents, and 'own standards' were least important. However, 101 respondents (nearly a quarter of those responding) rated their own standards as either first or second most important. This suggests that respondents' own internal standards are an important guide to judgement for many assessors. In addition, nearly 35% of respondents did not give first priority to the qualification standards (cf. similar findings in Eraut, Steadman, Trill, & Parkes, 1996). This does not necessarily imply invalidity; we have seen in the review section (pp. 33ff) that judgement is an integrated, value-laden process which cannot be reduced merely to explicit statements in qualification standards. However, these findings have important implications for standardisation and suggest that this aspect should certainly be addressed in validation.
- With respect to classification consistency (JUDGEMENT, assumption 5), large numbers of respondents believe that they have seldom or never witnessed misclassifications (308 'false positives' and 365 'false negatives'). However, there were 89 and 57 respondents who were prepared to admit witnessing the respective types of error sometimes. These substantial numbers may be a challenge to the validity of WBOA. (QCs seem relatively more likely to have seen candidates who pass but should not have sometimes, when compared to IQAs and assessors.) A common reason given by respondents for candidates PASSING when they SHOULD NOT was pressure from learning providers, centres and management to pass candidates quickly in order to get high achievement rates and meet success targets as these are linked to funding. This was said to be particularly difficult to resist for new and inexperienced assessors.
- In the domain of standardisation, consistency and verification (JUDGEMENT, assumptions 3 and 6), we found that QCs were less likely to be in the most frequent category for standardisation, but more likely to experience standardisation a year or more apart (this was confirmed in our interviews, see next section). The lack of frequent standardisation in the case of QCs is something that might cause some concern, given that they are the ones that should be monitoring and ensuring the standardisation and consistency of assessors and IQAs.
- Respondents were generally happy about the effectiveness of their standardisation activities, which commonly included activities such as double-observation, discussion of portfolios (and sometimes video recordings of practical tasks) and standards applied in judging various pieces of evidence, etc., which seemed relevant and potentially helpful in promoting consistency of standards in WBOA.

Consulting WBOA practitioners using interviews

As noted earlier, we also consulted assessors, IQAs and QCs in one-to-one interviews during the fieldwork stage of the project. Some of the interviews were carried out during our visits to sites where WBOA was taking place, and some were carried out over the phone.

In our centre visits and interviews we focused on three areas, namely Hairdressing, Electrical Installation (EI) and Plumbing. These were chosen based on a) likely availability of WBOA during September/October when the fieldwork was being conducted and b) absence of units where it might be difficult to arrange site visits due to confidentiality and other issues (which would have been the case had we chosen Health and Social Care, for instance).

In what follows, we summarise some main findings that emerged from our analysis of interview data and mention anecdotal evidence from our centre visits where appropriate.

About the interview schedules and respondents

The semi-structured interviews were carried out using an interview schedule designed to address several aspects of WBOA and the validation framework in slightly more depth than might have been possible to achieve using the questionnaire. An example of the QC interview schedule can be seen in Appendix 5.

We interviewed 18 people. Their roles (as well as whether peripatetic or work-based in the case of assessors) and sectors are detailed in Table 15. Some interviewees had more than one of the roles mentioned in the table. In those cases, they were asked to respond from the point of view of the more senior role.

Table 15: Roles and sectors of the interviewees

Role	Sector			Total
	Hairdressing	Electrical Installation	Plumbing	
QC	3	2	1	**6**
IQA	2	4	-	**6**
Assessor	2 peripatetic 2 work-based	1 peripatetic	1 peripatetic	**6**
Total	**9**	**7**	**2**	**18**

Main interview themes

In this section, we present and discuss the analysis of our interview data, which, we believe, further increases our understanding of how WBOA is used for assessment in the workplace and sheds some more light on which WBOA-related assumptions and threats to validity included in our validation framework are warranted and in need of further investigation and validation, and which might be of less concern.[31]

Advantages and disadvantages of using WBOA/implications for DOMAIN DESCRIPTION and assessment method choice

All respondents were very positive about using WBOA and did not see many disadvantages. A few enthusiastic comments included:

A. *The most effective assessment is direct observation, there's no two ways, no argument about it, it's just economical and practical to do everything that way, without a doubt the best, the most honest and true assessment is direct observation. (H, IQA, EI)*

B. *No major issues with this assessment model, it is the main form of assessment in the workplace, it is always a good form of assessment. (M, QC, HAIR)*

C. *It's endless - someone can tell you a story, but when you actually go out you see the candidate doing the work, not the physicality of doing the electrical installations but how they conduct themselves and interact with their employers, the other trades, primarily with the customer, you can see whether they, not* [so much] *have the hands on repetitive skills of being able to do the job, but whether they conduct themselves in an appropriate manner for the level of competency – so they have a correct communication style, take responsibility at that point. The level 3 … is not just can you do the job but are you aware of what you are doing, how you are doing it and why, with also providing a good, confident and competent ability... (I, IQA, EI)*

D. *More holistic approach than other methods, it offers a breadth across units. […] So you may think you're going out in a plumbing context to observe them fitting a bathroom suite, you pick up lots of other things there, you pick up communication, safety, decommissioning, commissioning. (J, QC, PLUMB)*

While it is reassuring that practitioners had a broadly positive view of WBOA, the above comments, as well as the ones we haven't quoted here, reveal a great deal more about the status of WBOA in VQ assessment and about practitioner views of the constructs that it can be used to assess.

[31] Throughout this section, the quotations are followed by the first initial of the respondent, the shorthand for their role (QC, IQA or A(sessor)) and sector (HAIR for Hairdressing, EI for Electrical Installation, and PLUMB for Plumbing).

For instance, despite the view expressed in the first quote above, it is often the case, in EI in particular, that evidence of competence is gathered in other ways, most often based on expert witness statements, often because, unfortunately, WBOA may not really be the 'most economical' method. This situation can sometimes present validity issues, particularly about confirming the 'authenticity' of candidate's work, but also about whether the alternative method chosen is really the best one for assessing the aspect of competence in question:

E. *Assessment strategy of the SSC requires about 80% of assessment by direct observation in the workplace. Most centres are quite fair with that and carry out as much assessment as they can like that, within the confines of the money they are allowed. (R, QC, EI)*

F. *I can't help but feel in a lot of instances when they* [candidates] *fill in their site diaries and so on and say to the workplace recorder 'I need you to sign this' and he just signs it. … So we are very reliant on the honesty and the integrity of the workplace recorder, because we are not 100% directly observing. And I also feel that that could be a bit of a weakness for the system, because we are not 100% sure about it. (H, IQA, EI)*

A view from the plumbing industry reveals a preference for a situation where the assessment strategy clearly states that certain things must be assessed by WBOA, and therefore removes the issues around confirming the authenticity of performance and suitability of the assessment method.

G. *…those who are working in the construction areas are envious of the plumbers because we have mandatory direct observations, there's criteria we must* [emphasis] *see. Whereas they're envious of us because they have to basically tick boxes and put their signature to a claim where in some cases they might not be 100% sure that somebody's got it. But they're presented the documentary evidence* [e.g., witness statements, reports], *they have to accept it. (J, QC, PLUMB)*

Most comments regarding the advantages and benefits of WBOA revolved around the contribution that seeing candidates in the workplace, with all its complexities and constraints, makes to assessors' insights into the candidates' competence. The comments often did not differentiate between advantages of assessing in the workplace (by any method) and the advantages of WBOA (which is just one of a number of methods that can be used to assess in the workplace). The comments below in particular highlight the value of having an opportunity to see a candidate in a work environment, rather than just in college:

H. *I find it quite illuminating really. You see somebody at college, in simulation or whatever, and when you meet them on site, really you meet a different person, and nine times out of ten it is very much to the positive. Whilst they are in college, they tend to expect things*

done for them and be led around. When you meet them on site, they are independent, they work unsupervised, they are quite impressive individuals really. (T, A, EI)

I. *Er, we get differences with attitude but not, not in ability. We get ones that are absolutely terrible in the centre and then when you get to the site you speak to employers and see them working, they're absolutely fantastic. You know, it's just a different environment which is nothing to do with the assessment criteria it's just the attitude* [of the learner]. *(J, QC, PLUMB)*

Several respondents said that they considered the ability to obtain the views of employers and colleagues while visiting the site as invaluable for helping them reach conclusions about a candidate's level of competence. Several of them said that observation in the workplace was less about observing candidates' practical skills, and more about seeing them in action in the real setting, seeing how they interact, how they form part of the team, what their standing was on the site, what their employers and colleagues thought about them (see also comments C and D above).

J. *'There is direct observation obviously, but I look at how long has he been on site, what he's been doing, he will then explain to me what he's been doing, he will show me doing that that day if you like. I also talk to the team. And I had one person this year, one employer that said he is not ready… I ask, 'Is he operating at L3 standards, is he meeting what you would class as an electrician', and one employer said 'No he's not.' (T, A, EI)*

K. *…some of them* [employers] *are very open to it, when we go out our assessors have an interview with the employers or just meet them on site while they are doing the site visit, have a little bit of a chat with them, take them to one side and say, how's he getting on. And they will tell you, they'll be very honest as to whether the candidate is worth the investment... (I, IQA, EI)*

This type of comment is indicative of the complexities around defining competence, as discussed in the first Review section. Thus, although assessors might be required to assess specific practical tasks and/or skills using WBOA, the actual construct that they are assessing involves more than that. This begs the question, as we discussed earlier, of whether the constructs that are embodied in the LOs and ACs which WBOA is supposed to address truly recognise and make explicit what assessors and other practitioners consider important aspects of competence.

This situation also suggests (as we have already argued, see e.g., pp. 25ff) that an approach of devising differently focused observation procedures (such as DOPS vs. LEP in medical assessment, see p. 27) might also be warranted in VQ workplace assessment. That is, designing two versions of WBOA, one of which might only focus on key procedural skills, while the other might focus on collecting evidence of

interpersonal skills, attitudes, professionalism, team work, and have suitably designed checklists and criteria to cover those aspects.

In addition, the necessity of asking employers and others for their opinion about candidates suggests that WBOA as currently used is not always sufficient for assessors (particularly peripatetic assessors) to be able to form a confident judgement about candidates (despite them having at their disposal other evidence types, e.g., witness statements, candidate reports, knowledge tests), and other forms of evidence are needed to reveal certain less tangible aspects of competence. Therefore, assessors sometimes resort to 'informal chats' to collect views of candidates' managers, clients and colleagues.

However, to our knowledge, informal chats are not recognised as an assessment method in VQs and it is essentially left to the assessors to use their own judgement when collecting these various views about their candidates. In contrast, as noted earlier, in medical education a method called 360° is used to collect the views of colleagues and others with whom candidates interact, using a structured form or a questionnaire, and making sure that a representative sample of colleagues and others is used for every candidate, rather than just the current boss. This might also reduce the likelihood of bias of the kind referred to in our Review section, where the 'good bloke' might get through based on just the boss's opinion. The comments that we heard clearly call for a similar approach in VQs.

Most respondents highlighted the issues around practical difficulties and time and funding limitations of meeting the range of situations/tasks on which candidates should be assessed using WBOA (cf. similar findings from our questionnaire). This in itself can lead to problems around the authenticity of candidate performance, but also of making sure that the appropriate tasks are used to assess appropriate skills and other aspects of competence. This is the problem, highlighted in the Review section, of assessment tasks being *genuine* – but not necessarily *authentic*, not enabling us to see the skill that we really want to see.

L. *Because it is mainly small to medium employers, it is sometimes difficult to make sure that when you book a visit and they get called away to something and have to do something else which you may have seen before and therefore … all the planning is up in the air and with observation being rigidly enforced in the NVQ structure it makes it very difficult for flexibility. … Particularly towards the end when candidates are really busy with normal work and you need to catch them to demonstrate some outstanding ACs. Might be better to have the option of a 'skills test' so you could fill in the gaps (J, QC, PLUMB)*

M. *…it is unlikely that each time you go on a visit you will cover everything that you might be expected to cover – but I've rarely seen in portfolios that assessors actually say that they need to revisit, usually all is covered… (J, QC, PLUMB)*

Obviously, in relation to comment M, the relevant guidance for WBOA does require assessors to revisit on another occasion, and to make sure that they are satisfied with the quality and extent of evidence of competence provided before they make a decision about a candidate. However, practical difficulties and other limitations seem to be somewhat of an incentive for bad practice and a threat to validity that needs to be addressed either through verification, or through combining WBOA with other methods, so that essential evidence of skills and abilities can be collected.

Other difficulties with WBOA mentioned involved issues with accessing certain sites (e.g., Ministry of Defence, London Underground), the necessity of being very tactful with employers inasmuch as poor performance of their candidates might suggest that the employer did not teach them well or that the employer's practice is questionable and does not conform to regulations, etc. In Hairdressing, where a lot of assessors are work-based rather than visiting, there are issues about whether they can set aside enough time for assessment and training (although normally a portion of their working hours is dedicated to this), and about 'getting them out of the salon' to attend standardisation and other activities.

The comments and issues summarised in this section confirmed to us the necessity of dedicating a part of our framework to clarifying the construct-related issues (DOMAIN DESCRIPTION), but also that the status of WBOA is unclear as to what it can and can't, should and shouldn't do, and that it might be beneficial to invest some effort into targeting WBOA procedures to suitably (and possibly more efficiently) assess different aspects of competence.

Competence definition/domain description/washback

We explored the issues around competence and domain definition through several other questions relating to the necessity of using WBOA to assess tasks that rarely happen in the workplace, and also related to possible lack of assessment focus on tasks/skills/aspects of competence that are considered important (and might often arise in the workplace).

Hairdressers gave several examples of tasks rarely occurring in practice that need to be assessed by WBOA, for instance, cap highlights, perming (though in an optional unit), and payment by cheque in the reception unit. Some highlighted the issues around potential skill loss due to the interaction of changes in fashion and the necessity of assessing the skills on paying clients in the workplace. This becomes difficult when for example, in Hairdressing, no clients want perms, yet perms are a compulsory unit for assessment. The solution for this problem in Hairdressing was for perming to become an optional unit, rather than allowing for at least some aspects of it to be assessed in simulation. However, some of our respondents were worried that this strategy might remove some very useful skills involved in perming from the profession, if candidates increasingly fail to choose this unit for assessment due to practical difficulties. This illustrates why it is necessary to reach an agreement at sector level regarding what is more important – training

candidates in a particular skill and assessing this even if it is in simulation (in order to retain the skill in the sector), or treating this as an optional unit in order to retain the realism of only assessing on paying clients (but with the danger of losing an arguably useful skill from the sector).

The respondents from the EI and Plumbing sector had a few similar examples, e.g., working with lead becoming an optional unit in Plumbing. In EI, fault-finding and inspection and testing are difficult to assess by WBOA as they either occur rarely (faults), or, in the case of inspection and testing, are done by separate teams in the workplace, and unless the candidate gets on that team, it is difficult to get experience and assessment opportunities in that domain. This raised questions of the need to assess these tasks and associated skills in simulated environments rather than necessarily in the workplace.

In terms of tasks/situations/skills that are often present in the workplace and considered important by some practitioners, but are not assessed, putting up long hair, and old-fashioned ways of setting were mentioned as either not assessed, or not assessed (and practised) enough in Hairdressing. In EI, practical tasks on motors were mentioned in this context.

These issues further highlight the importance, and difficulty, of defining what is important and what is not in a particular sector/role, but also the importance of combining different assessment methods and settings to ensure that important skills and aspects of competence are assessed appropriately, rather than strictly sticking with certain assessment procedures (or orthodoxies) and settings where this is inappropriate. These are important messages that relate to (negative) washback – effects that assessment might have on instruction and learning – which are important in the validation of assessment methods, and are addressed in the UTILISATION part of our framework.

Confidence in judgements/standardisation of tasks/generalisation

Several interview questions probed our respondents' views about how confident they were in their judgements, and about any practices that might be employed in WBOA to standardise the tasks that are assessed to some extent. These (together with the questions about the number of observations in the questionnaire) have implications for the thorny issue of generalisation in WBOA and in performance assessment generally.

Comments N and O below represent, in our view, two typical ways in which assessors might achieve confidence in their judgements when using WBOA. The comment N was given by a work-based Hairdressing assessor, who both trains and assesses the candidates, and who felt that this whole process informed their judgement, so that even before the formal occasion of summative assessment, the assessor could already be sure that the candidate would pass (which might present less of a threat to generalisability, see discussion on p. 42).

N. *Yeah, because we do the training, and once they're at a level, we practise and practise. And once their timings are up and they are capable of it, and […] if you've been teaching highlights for example and they've done lots of different highlights and they're in their timings and the highlights are good, then you'd go 'right, next one we'll do assessment'. So, I would be happy, the candidate is happy and it's normally planned. And I wouldn't put them forward for something if I didn't think that they were happy and confident, and I was, to do it. (L, A, HAIR)*

However, work-based assessors are not that common in all vocational sectors. For instance, in EI and Plumbing, it is more common for candidates to be assessed by external visiting assessors. The comment O below (and also comments J and K above) shows that information provided by candidates' employers and colleagues is particularly important in such circumstances for raising assessors' confidence in their decisions. Again, we would argue that in these situations it would be beneficial for both reliability and validity to ensure that the methods for collecting this important feedback about candidates are as rigorous as possible (along the lines of medical 360°), rather than just relying on informal chats and unrepresentative samples for getting the feedback (see also previous discussion).

O. *Exactly why I talk to the managers. Because, they all see the guys on a daily basis. That's why I think it is important to talk to the people that this guy, the trainee, is working with. (H, IQA, EI)*

We saw in the Review section that in medical education a great deal of care is taken to ensure that candidates assessed in the workplace are assessed in situations that are as far as possible comparable (e.g., in terms of the complexity of the condition that the patients on which they are assessed might have). We were interested to find out if issues like this (extreme or unrepresentative tasks/situations/complex client demands/etc.) affected the use of WBOA, whether there were any attempts to standardise tasks in this respect, and whether our respondents perceived that the problems of controlling these complexities in the workplace setting posed any issues of fairness.

All our interviewees agreed that this was not normally seen as a problem and confirmed that there was no explicit attempt to standardise assessment in this respect. They agreed that they would not stop a task because it was too extreme to be representative, except for health and safety reasons, although candidates themselves could request for an assessment to be stopped if they could not deal with the situation/task/customer. The following comments illustrate this:

P. *Because I do tell them, it's hard luck you're gonna get clients like that … (L, A, HAIR)*

Q. *No, he's doing whatever he's doing that day. If it's units that are within the criteria for assessment, then it's hard luck if some candidates end up being assessed in complicated situations. (H, IQA, EI)*

R. *They are gonna have difficult clients in their everyday columns[32], they are used to dealing with difficult clients, it is not normally a problem …. All it's doing it's actually training them very well for running a column. (M, QC, HAIR)*

This is a controversial issue and it highlights the tension between reliability/standardisation and validity in WBOA. In Hairdressing, part of the range directly addresses this issue and to some extent makes it part of the construct in that some ACs require evidence of dealing with 'unconventional customers'. Where this is not the case it might be worth considering to what extent dealing with difficult situations is important for the relevant notion of competence, and if this is considered relevant, attempting to come to a more formal agreement on how to make these sorts of situations part of assessment without putting the fairness of the assessment at risk. This is particularly relevant in light of the following comment, which shows that, if candidates want to avoid being assessed in difficult situations, they can often do that. To the extent that being assessed in difficult situations should be part of the construct, this presents a problem for authenticity and validity.

S. *…the candidates are quite savvy and make sure assessments happen in predictable situations, they are not going to pick anything difficult out. It is difficult to request to see a particular task because we can't dictate to employers what and when they are doing it. If you are not confident on something and really want to see something you can call them into the centre and put them through a simulation, but this rarely happens. (H, IQA, EI)*

Issues of assessor consistency, standards and standardisation, and spotting bad practice

The IQAs and QCs we interviewed were generally positive about the quality of assessors, though they did point out that this was partly dependent on assessor experience, and therefore less experienced assessors would be quality-assured more often as they would be 'higher risk'. For instance:

T. *Sometimes* [they over- or under-assess]*, but mainly newly-qualified, they are classed as high risk (B, QC, HAIR)*

Discussion about whether there are problems with assessors leading/providing inappropriate help to candidates during WBOA gave rise to mixed comments, suggesting that this issue might be worth following up in validation:

[32] In hairdressing jargon, a column is a list of appointments that each of the hairdressers in a salon has.

U. *Inappropriate guidance maybe happens with inexperienced assessors, but don't see it later (I, IQA, EI)*

V. *Er, no I can't say I've encountered that, I think they're all professional enough, er yeah, there's nobody going out with malice or forethought. (J, QC, PLUMB)*

W. *I think, sometimes when it's their candidate, they tend to be more helpful. If they taught them (the candidate), they shouldn't be doing the assessment but a lot of them do because they do not have anyone else to do it. (I, QC, EI)*

X. *There can be a bit of a nod and a wink at times in terms of should you be doing this, but generally no…. It's very rare.(R, QC, EI)*

Y. *Sometimes there can be a fine line between coaching and assessing. (M, QC, HAIR)*

QCs and IQAs tended to recognise that there were differences in standards between assessors, particularly between different providers, but, again, generally did not present this as a major problem:

Z. *There is a difference [in standards] but that's more on experience than anything else. The more experienced assessors, in general, do the best job. The ones who have just started and are still feeling their way, perhaps need more guidance. (R, QC, EI)*

AA. *Majority of time assessors are consistent but do get the odd one. IQAs I've come across get things done and deal with any issues quickly and are generally consistent. (IM, QC, EI)*

BB. *They do the best to work to the standards and we're all only human at the end of the day how you interpret those standards. (B, QC, HAIR)*

CC. *They are quite level yes [in severity], I think across the board are quite fair, most of them.(R, QC, EI)*

When we asked assessors about standards and how similar these were 'in-house' and nationally, they mainly thought that in-house standards were similar, but that nationally there were bound to be differences. They also said that opportunities for standardisation were fairly frequent in-house, but rare at regional or national level.

DD. *… I would like to think that my standard of what they should be doing would be the same as the assessor in here … that we would agree,[…] Cos we do train quite closely, and work quite closely, so as a salon we've got a standard that we like them to meet, as well as the actual standards that are set of the NVQs. (L, A, HAIR)*

> EE. *Within the college, pretty good, I'd say very good. It's the team of three of us. J and I will very often work the same area at the same time, doing assessments, so you are listening. Plus we literally do do the standardisation meetings … and occasionally we do change things, 'yeah, I accept that, I can see where you are coming from…'*
>
> *There's less evidence of standardisation nationally … I suspect there will be regional variation. (T, A, EI)*

We were pleasantly surprised to have found our respondents generally fairly happy about assessor practices and standards, and about ways of improving these and resolving issues in the case of less experienced assessors. With respect to any disagreement in decisions between assessors and IQAs/QCs, we were generally told that disagreements were rare, or not of much concern as they could be settled with discussion.

These findings were also supported by the corresponding findings from the questionnaire. In addition, the questionnaire responses suggested that a variety of standardisation activities were used to promote consistency of standards with respect to WBOA. We believe, however, that these issues still deserve a place in our framework, and that it would be useful to invest some research and funding to ensure that suitable methods are developed for periodically checking the consistency and reliability of assessor judgement independently for validation purposes.

This is because not all comments about agreement levels were positive (cf. the questionnaire-related discussion about dissenting minorities). In addition, we did observe a few instances of disagreement between assessors and IQAs/QCs during our visits (in three out of five visits), with one of these instances involving an assessor and an IQA who work closely together and take turns in performing the IQA role for each other. This suggested to us that perhaps the situations where there was disagreement were more prevalent than our interviewees suggested.[33]

Verification

We were interested in IQA and QC views of the effectiveness of their verification/quality assurance procedures, as well as the standardisation of their practices nationally.

[33] While talking to our interviewees and other practitioners we met during our centre visits, we suggested the idea of making videos of candidate performances and using these videos for standardisation and validation studies. We found most of our respondents quite supportive of this idea, given certain limitations, and some were even willing to provide facilities and opportunities for making such videos.
We believe that this would be a very useful exercise and if pilots proved to be viable in practice (even to a limited extent), this could benefit the vocational sector in a number of ways. Banks of videoed performances could be re-used for standardisation purposes. Multiple-marking studies could be carried out periodically to estimate marker reliability and consistency, the results of which could again be used in standardisation, but also in verification and validation. We believe that investment in this area would be beneficial for both AOs and the regulator in the long term.

The IQAs we talked to tended to approve of current quality assurance procedures and of their own level of standardisation. However, they were less happy with the level of standardisation of QCs, saying that they were often inconsistent in their requirements.

With respect to QC quality assurance procedures that should discover inconsistent or problematic assessors across different centres regionally, several respondents pointed out that this is not always easy, especially when carried out remotely, and that it was necessary to consult IQAs and candidates. The latter were considered an important source of information regarding assessor-candidate personality clashes, discrimination, and assessor severity.

> *FF. That's gonna very much be down to your sampling. To get inconsistency in an assessor you've got to obviously see a number of different units or the same unit on different candidates, etc. You also rely on IV feedback as well, so you have a look to see if they pick anything like that up, so if there are any inconsistencies then ask to see some more candidates from that assessor. (M, QC, HAIR)*

Several QCs emphasised that the verification process was as effective as it could be given the limited time allowed and low frequency of visits (once a year for low-risk centres):

> *GG. You just have to get a feeling that from what I can see this is correct. Beyond that I can't go. (R, QC, EI)*

> *HH. And also we're only in for that one day and it's a window on that day of what you see so you can only sample so much so if there's a lot, if it's a big centre and there's a lot going on you can only see a little bit. New assessors can come and go and there would be issues and problems […] and you would never know. […] But it's whatever you sample of what you see, it's all you can take it on. (B, QC, HAIR)*

> *II. … it's just to get a snapshot, obviously if I'm going to visit on a few occasions I'll pick different assessors and so on and staff just to see that there is consistency in interpretation. (J, QC, PLUMB)*

As might be expected, the QCs also thought that there were differences between them in terms of standards and were generally not happy with the amount of standardisation available (cf. the corresponding findings from the Questionnaire).

> *JJ.* [Standardisation is] *effective for those who attend. But there are differences between QCs, some consider themselves to be in a policing rather than supporting role. (J, QC, PLUMB)*

KK. I am not happy, no. We have probably one briefing day per year. Half of that is taken up with very generic matters. And as we share the day with plumbers, part of the day is taken up with plumber's work. Then we split apart and if we get to standardisation I am lucky … (R, QC, EI)

LL. I don't know, all I can say is I work to the standards and try to interpret the standards to the best of my ability. (B, QC, HAIR)

Although it is admittedly not easy to standardise people who are making complex judgements, this is arguably more likely for smaller groups such as groups of QCs for individual sectors. However, sufficient investment is needed, but probably also a recognition that the role of QCs and their quality assuring procedures are crucial for both ensuring validity of qualifications, and for collecting first hand validation evidence.

Extrapolation

As noted earlier, we believe that in the case of validation of individual assessment methods, the results of which only contribute to conclusions about candidates at qualification level, the EXTRAPOLATION inference needs to be considered for a narrower target domain, although perhaps it is generally more appropriate to consider extrapolation at qualification level. We asked several questions in the interviews relevant to EXTRAPOLATION, partly relating to how qualification-level results might be interpreted, but also about whether candidates who had been assessed by WBOA and passed, are then able to carry out all the relevant tasks and have all the relevant skills for the workplace.

With respect to qualification-level interpretations, we expressed this through questions about whether candidates were 'job-ready' once they qualified (or what they were ready for), how much supervision they might require, and how easy it might be for them to transfer between different workplaces once qualified in their domain.

The hairdressers tended to agree that the L2 NVQ did not prepare them to work as a stylist that would accept 'any client that walks in the door', they could work as a 'junior stylist', but would really need the L3 qualification to be able to engage in most normal workplace activities, although in some salons even qualified L3 candidates might be treated as 'improvers' or would have a 'graduate' status, and work at lower prices for a while. The respondents did not think that transferring between salons would be a problem, though they did recognise that different salons worked according to their own specific standards, and that in those cases some extra training might be required.

The EI and Plumbing respondents also tended to agree that L3 candidates were job and sector ready, that the qualification had reasonably good coverage, although, given how wide these sectors are, they thought

it was unlikely that the qualified candidates would always be able to carry out every possible task that they might encounter in the workplace. But they did not see a problem with transferring between workplaces, even if these were specialising in particular areas, as additional training could be provided where needed. Some relevant comments are below:

MM. Yes, without any doubt. Otherwise I wouldn't sign them off, because there is too much at stake, he could do himself some damage, damage to other people. I have to be happy, the people that he's working with have to be happy, his boss has to be happy. There's a whole host of hoops to jump through … (T, A, EI)

NN. [L3] Oh definitely they should be able to… …certainly into the workplace, especially if they've carried out an apprenticeship, they'll be employable and be able to work on clients. But even on level 3, you know, you still now and then need supervision or need to check with somebody. We all need that however long you've been in industry, you know from time to time and that's not a bad thing. Should have no problem transferring between salons… (B, QC, HAIR)

OO. … with the NVQ which covers virtually everything you can think of, that person, if they have legitimately covered everything, should be able to transfer from one company to another without any problem. And most electricians do, it's a very fluid workforce. It would only be if they were to go into a company with some particular specialisms that they would need additional training. (R, QC, EI)

With respect to the possibility that qualified electricians or plumbers might not be able to do certain things in the workplace, this is what our respondents had to say:

PP. … because the range of work as an electrician is very broad, there will be things which they will be doing less regularly once they are qualified, for instance they might work as a domestic electrician and then struggle with some commercial sites, and would need to pick that up again if they were to get employment in such a place – so sometimes it can happen that even though they are qualified, they are not able to do certain things, or might struggle with them… (I, IQA, EI)

QQ. Occasionally. […] It could also be because they have only done a particular task you are observing them doing once before to pass the qualifications so it's only the second time round. But this is not a major issue, does crop up occasionally. (B, IQA, PLUMB)

RR. ...it's such a variety that you do in plumbing that you cannot be trained on everything, you might know the basic concepts… there's also skills fade, it can happen. (J, QC, PLUMB)

Judging by all these observations, it seems that extrapolation to relevant job/sector domain for the three sectors that we looked at, and in particular for L3 Apprenticeships was not problematic, although some coverage will always be missing. Slightly more worryingly for WBOA, some respondents (cf. QQ above) mentioned the possibility that certain tasks might be observed only once for assessment purposes, so it would not be surprising that in the future they might not be very confident with that – though the comment did not suggest that this was a common or a major problem. The issues around skills fade or obsolescence were also recognised as relevant, and this would have to be taken into account in validation, as no assessment method could prevent this fact of life.

Funding pressures

Although in our original drafts of the interpretive argument for WBOA we did not take into account threats to validity related to funding pressures,[34] following our fieldwork we decided to include these because a number of our respondents referred to them as an explanation for instances of bad practice in WBOA, in particular in relation to 'pushing people through qualifications'. Although these funding pressures are not specific to WBOA, it appears that WBOA is not immune to them.

SS. *I've actually spoken with somebody recently who was about to finish with a provider, and gone to a salon for a job, and by the student's own admittance she can't do the work that she has been told that she can do and been assessed. She can't do it. …* [What's their motivation to do that?] *… Probably hitting their targets, getting the amount of students through that they have taken on.' (L, A, HAIR)*

TT. *There are funding pressures to complete, but this can partly be managed if assessors are well organised and 'keep rattling* [candidates'] *cages' to make sure that there is not a backlog of things that they end up having to be assessed on when they are supposed to be near completion. (T, A, EI)*

UU. *You can't cut out the funding issues altogether. So, sometimes you just can't stop copying or help by assessors – e.g., if someone is 98% through but can't be bothered to complete and someone just tells him 'write this down'. Then it's a case of prove that he didn't do it. It's signed off at the bottom. (R, QC, EI)*

We have included funding-related threats to validity in the JUDGEMENT part of the framework.

[34] Currently providers can claim funding contribution for eligible learners on achievement of government funded qualifications.

Summary

We believe that, in general, we have managed to obtain useful information about the functioning of WBOA in practice and the practitioners' views about aspects of its validity through structured one-to-one interviews. We think that this method for collecting validation evidence would be a useful one, particularly if a broader sample of stakeholders was consulted (in particular candidates, employers, and representatives of relevant SSCs).

In terms of the main themes that emerged from the interviews, and the implications this had for the WBOA validation framework, we highlight the following:

- There was overall support for the value of the workplace assessment setting in general and for WBOA in particular, as indispensable for getting appropriate insight and a holistic picture of candidates' competence.
- However, several aspects of what our interviewees told us confirmed to us that a systematic effort is needed to better understand and validate WBOA, and indeed to carefully think about whether it is always used in the best and most efficient way possible.
- We found that there were certain theoretical and operational construct definition issues, both in terms of domain coverage (which is possibly of less concern), and about reaching a level of agreement as to which less tangible aspects of competence as well as cross-task skills, abilities, processes, and attributes are important for each role and sector. (DOMAIN DESCRIPTION, assumptions 1 and 2)
- Agreement would also be required about the extent to which the less tangible aspects of competence should be assessed, are assessable with WBOA in its current form, and whether they might be better assessed by making them explicit in construct definition, and targeting WBOA better (or using other methods) to make sure they were assessed validly. (DOMAIN DESCRIPTION, assumption 3)
- We highlighted possible negative washback problems if assessment strategies are implemented too rigidly, and included this in our framework as a possible threat to validity of WBOA (and other workplace methods).
- We also encountered the possibility that candidates may sometimes be assessed on tasks that may be genuine but may not be truly authentic, i.e. not enabling the assessors to see the skills/aspects of competence that they really want to see. This was sometimes caused by the logistical and practical problems of assessing in a real workplace setting. However, this might also occasionally be caused by the fact that candidates could choose the tasks and situations they would be assessed on (within the required range), and so might attempt to avoid complex settings/situations, even though this might help assessors get a deeper insight into their competence. We believe that these are complex problems that deserve to be investigated in validation as part of both DOMAIN DESCRIPTION (assumptions 4 and 5), and GENERALISATION (assumption 1), but also addressed in general research

and assessment practice, which would seek to develop informed best practice related to these issues.

- The previous point is related to more general issues around task standardisation, and the necessity to at least recognise that there are potential problems there and that best practice would be to develop a consensus on how to deal with these issues (e.g., variable task/situation complexity across candidates); also, is task/situation complexity part of the construct?
- We found that the seriousness of the threats to validity related to the GENERALISATION (assumption 2) inference, in particular with respect to the number of observations and assessor confidence, could not be generalised across all settings for WBOA. In situations when a skilled work-based assessor is involved both in training and assessment of the candidate (e.g., in Hairdressing), we believe that threats to generalisation are less serious since an assessor's judgement and confidence are formed and reinforced both during training and formative and summative assessment. In the case of visiting assessors, the threats are bigger, and we pointed out that these might be alleviated by developing standardised methods, complementary to WBOA, for collecting additional relevant information about candidates.
- We found our respondents to be generally fairly happy about assessor practice and standards, and about ways of improving these and resolving issues in the case of less experienced assessors – although existence of differences was acknowledged, as well as some instances of bad practice. We believe, however, that these issues still deserve a place in our framework (ADMINISTRATION, JUDGMEMENT), as not all comments were positive. We also believe that it would be useful to develop suitable methods for checking the consistency and reliability of assessor judgement independently for validation purposes. (JUDGEMENT, assumptions 3 and 5)
- The issues of consistency were also present in the case of IQAs and QCs. The latter, in particular, were quite critical of the amount of time and resources devoted to standardisation. We also believe that a fairly low frequency of QC visits might pose some threats to validity of WBOA and more generally. (JUDGEMENT, assumption 6)
- With respect to EXTRAPOLATION, our respondents' views largely supported validity at qualification level in this domain. The issues identified were partly also those of construct definition and DOMAIN DESCRIPTION, and partly referenced problems with GENERALISATION (small number of observations), that might in turn affect extrapolation.
- Finally, following our interviews (see also related questionnaire findings), we decided to add some threats to validity related to funding pressures to our framework. (in JUDGEMENT, against assumption 5)

Quantitative evidence of validity in WBOA

In its bid to Ofqual, City and Guilds promised to:

- develop suitable measurement indices to show the quality of overall observed assessments and elements of assessments
- discuss how such indices could be integrated with scores from other types of assessments (e.g., tests, assignments, etc.)
- conduct analyses to demonstrate the functioning of such indices, using either existing data from City and Guilds qualifications or data simulated to fit with the structure of real City and Guilds qualifications.

In a sense, the assertion that we will do quantitative analysis to generate evidence of the validity (or invalidity) of WBOA is an unremarkable statement. Social sciences typically employ both qualitative and quantitative methods; and assessment research – particularly – is a highly quantitative discipline. Following that line of reasoning, doing quantitative analysis of outcomes in WBOA is just a standard activity, and therefore this section is redundant.

However, it is our view that a substantive section discussing quantitative approaches to generating validity evidence in the case of WBOA in VQs is necessary. Little quantitative research has been conducted into WBOA in VQs, and when we have discussed this with colleagues, some have doubted that quantitative approaches are valuable at all, or that – even if they might be valuable – they can be meaningfully applied.

However, although we do perceive some substantial challenges to doing quantitative analysis in the context of WBOA, we believe that such challenges need to be set out explicitly and resolutions to them need to be found. Perhaps foremost among such challenges is that VQ assessment generally, and WBOA particularly, requires the researcher to interrogate many of the assumptions that underlie quantitative methods. In this section we attempt to do this by considering the nature of quantitative methods that are typically used to show the quality of whole assessments and elements thereof. We do this in a thoroughgoing manner and address several particular issues that apply to quantitative methods across the validation framework that developed (e.g., ‘task difficulty’ in DOMAIN DESCRIPION, ‘assessor severity/inconsistency’ and ‘over/under assessment’ in JUDGEMENT, number of observations in GENERALISATION, etc.).

Common statistical indicators of the quality of assessments

Assessment researchers derive statistics and indices to show the quality of whole tests and/or elements thereof, and these are used to inform test development and monitor test quality. Such indicators could include:

- A statistic or index to show the difficulty of a question or part thereof. These may vary – for instance, Classical Test Theory (CTT) might provide a facility value, whilst Item Response Theory (IRT) might provide an item difficulty index (more technically, a location parameter on an ability – difficulty or 'trait' scale). The characteristics and/or limitations of such indicators are often at issue; facility values are said to be meaningful only in terms of the cohort of candidates that responded to the questions, whereas IRT difficulty values are claimed to provide 'sample independent measurement' – estimates of person ability that are independent of the particular questions to which those persons responded, and estimates of item difficulty that can be interpreted beyond the particular cohort of responding candidates.
- An indicator (such as a statistic or a curve) of how a question performs across a group of candidates. In CTT, discrimination indices are based on various forms of item – whole correlations. Test developers might interpret such indices as showing how well the item distinguishes between generally strong and weak candidates in terms of what is being assessed. Similarly, IRT models provide item characteristic curves (ICCs), which show how steeply the probability of a candidate responding correctly to a particular item ascends as candidate ability ascends.
- Various 'flavours' of reliability index[35] and/or analogous curves. CTT provides several indices of internal consistency; all modelling in various ways the stability and coherence of the data set generated from a test. IRT might provide a Fisher information function – a curve that can show the overall magnitude of measurement information from a particular assessment data set, and the location of the peak of precision on the ability—difficulty continuum.
- Other indices may help an assessment developer or evaluator to understand how well various assessors or raters are performing. Fairly simple statistics and tables may be provided to show percentage agreement rates between markers, and more complex statistics can be produced, such as Cohen's kappa, which show the percentage agreement discounting agreement by chance.
- There are also misclassification indices of various sorts. These might include loss indices. Threshold loss indices show the proportion of test takers that would be categorised into a different category were a measurement procedure to be repeated. Squared error loss indices would show misclassification rates, but 'weight' them so that more gross misclassifications (for instance candidates misclassified when further from a cut score) 'count for' more within the index.

[35] By reliability, we refer to the consistency of outcomes that would be observed from an assessment process were it to be repeated. High reliability means that broadly the same outcomes would arise.

- There are wide-ranging approaches such as generalisability theory (g-theory), which is essentially an application of experimental design techniques. G-theory allows a researcher to derive 'reliability-like' indices, but also to show how much different facets (things like questions, persons, raters, occasions, etc.) contributed to the variance in a data set generated from a given test. Further, g-theory practitioners can perform D-studies (or optimisation studies), which allow them to model measurement properties (especially variance components) for different numbers of questions, markers, etc.

This is clearly a partial listing of all the statistical techniques that could be applied to assessment data in order to understand the properties of assessments. Further, we have grouped such techniques in a particular way above; others might do this differently. However, despite such reservations, researchers and test developers can pick and mix from such quantitative techniques and gain useful information. Our clear contention is that researchers and developers of WBOA ought to be able to analyse their data using some of the above-mentioned techniques, and thus gain understanding of the functioning of their assessments.

Why this might be a good idea

In this section we outline the main arguments for investigating and providing quantitative evidence of validity.

- *Assessment research is fundamentally a quantitative discipline.*

Thus, eschewing quantitative techniques amounts to not using many of the main methods that an assessment researcher might use. It does not mean that quantitative approaches should predominate, but the sensitive application of quantitative techniques will surely increase the amount and types of information available to WBOA developers and researchers.

- *Many of the questions that a validity researcher might be interested in have quantitative answers.*

Did this assessor apply the same standard when coming to judgements as that assessor? How reliable/internally consistent are the results from that WBOA? What proportion of candidates was potentially misclassified in that WBOA? All such questions have quantitative answers; thus, it makes sense for us to find appropriate ways to employ quantitative analysis techniques.

- *Quantitative aspects tend to be the weakest link in the validity chain for WBOA.*

Without prejudice to our final validation framework, it seems reasonable to say that WBOA is strong on areas such as: fidelity, authenticity, and integration with instruction. However, in contrast, it may be weak on: task-to-domain generalisation, internal consistency and inter-rater reliability. If this is true, investigation of these issues necessitates the use of quantitative techniques.

- *Critics of VQs think that absence of evidence of quality amounts to absence of quality.*

VQs in general, and in particular internally-assessed elements such as WBOA, have been characterised as lacking 'robustness' by politicians and criticised as being inferior to external assessment by some media commentators. In such a political context, evidence is often conceived of in terms of statistics and indices. As such, proponents of high quality VQs need to make their case in terms that critics and sceptics will recognise. This means deriving statistics and indices to show how well assessments are functioning.

- *The VQ sector needs to understand much more about the functioning of their assessments, and elements thereof.*

Quite simply, quantitative methods provide types of information that can supplement the insights garnered from qualitative investigations.

Why this might be a bad idea

- *In seeking to better understand quantitative facets of validity, one might actually 'damage' other facets.*

As Schuwirth et al. (2002, p. 927) state, one should not evaluate a pointillist painting by counting how many dots there are. One should not lose the overall sense of making a qualitative judgement about an assessment. Quantitative evidence can provide useful information to inform validity reasoning, but it should not become an end in itself.

In a sense this is just a re-wording of the old chestnut that reliability should not drive out validity. This implies that quantitative approaches should be applied in a sensitive manner; it does not imply eschewing quantitative approaches entirely.

- *Much of the VQ assessment system was set up without considering how one would evaluate it quantitatively.*

To some extent this is speculation, but in the reading we have done around the setting up of national initiatives such as NVQ and QCF we perceive that quantitative considerations received scant consideration. This may be a slightly negative construction of a decision to prioritise considerations such fidelity and authenticity over psychometric concerns, which is common in much of performance assessment, and particularly competence-based assessment, but nonetheless it makes it harder to apply statistical techniques to data that have not been designed to be analysed in that way.

- *There has been little quantitative research into VQs, esp. WBOA.*

This perhaps reflects the design priorities, alluded to above. There has been (too) little research on VQs generally, but this is particularly true of work using statistical methods. This is arguably both a strength and a weakness; research has to be done 'from first principles', but any new research is a substantial contribution to the field.

- *We are not responding to existing demand from assessment developers.*

At least when we have discussed this with a small number of colleagues in our own institution, they do not seem to see the worth or applicability of quantitative indices for WBOA. They are comfortable using facility values and reliability indices to understand how their items and tests function, but are more sceptical about analogous indices for WBOA.

This may be for several reasons: perhaps, developing statistical indices for WBOA is just a poor fit – a bad idea. Also, though, this might reflect an inadequacy in the VQ sector generally; although such qualifications can be based on highly valid assessment, the sector is not good at communicating such quality in terms that politicians and other influential commentators will recognise.

- *Data recording can inhibit analysis.*

There are legitimate and illegitimate facets of this: when Harth & Hemker (2012) collected data for their reliability project, they found that centres used disparate forms for recording assessment outcomes. These were difficult to interpret, and required substantial interpretation by trained researchers, thus raising concerns about practicality and inter-rater reliability. Such concerns can potentially be mitigated as e-portfolios become more widely used.

However, there are also more legitimate senses in which data recording practices can inhibit analysis; i.e. things that are not a 'mistake' per se, but which still make analysis difficult. In a WBOA, one observation may provide evidence for the achievement of several LOs or AC (hair cutting *and* colouring; health and safety LOs as well as substantive LOs from the subject concerned). This is a legitimate way to increase assessment efficiency, and is relevant to real workplace conditions, but it does make it harder to do quantitative analysis. This is perhaps one of the types of difficulties that any quantitative analysis would need to overcome; we would not suggest that one observation can only provide evidence for one LO.

Quantitative analysis might be hard, even if it is a reasonable thing to do

In the previous section, we have discussed some of the possible objections to quantitative analysis of WBOA data, but we nonetheless believe that none of these objections disqualify the use of quantitative techniques. Notwithstanding this, quantitative analysis is still challenging. In this section, we set out some of the features of WBOA and the data generated by them that appear to 'defy' quantitative analysis.

- *'Scores' on observations are not additive in the same way as test scores.*

In a test, a person that scores more marks is generally better than another candidate who scores fewer marks. There are caveats to this which flow from confidence intervals due to measurement error or perhaps grading issues, but by and large the statement holds true.

The same statement is not true in any straightforward sense for WBOA. Typically, a VQ unit will specify the minimum number of times that a candidate needs to be observed in a WBOA. The better candidates will probably be observed on the minimum number of occasions only; weaker, more borderline candidates will probably need more watching before being deemed to be competent. However, one cannot assume that candidates that had more observations are weaker; it may just be that a trainee hairdresser needed to be observed more because the cut she was being assessed on did not occur frequently in a (her) workplace. There is no suggestion that she cannot do that type of cut.

- *Between-candidate variance tends to be low – a problem for conventional internal consistency reliability measures.*

Boyle & Rahman (2012) and Johnson, Johnson, Miller, & Boyle (2012) analysed VQ tests (multiple-choice tests and assignments containing short-answer questions). They both found that score variance in such tests was low, and concluded that this made interpreting reliability coefficients difficult. It may be that the value of a reliability coefficient was low, even if error variance was quite low.

This phenomenon – common in criterion-referenced tests – is likely to be even more noticeable in competence-based judgements in WBOA, thus making it difficult to interpret reliability values on any coefficients that have total variance as a term in their denominator.

- *There is often no artefact – just an observed performance.*

Pollitt has argued that the measurement approach of adaptive comparative judgement (ACJ) allied to statistical techniques such as Thurstone Pairs or Rasch-based measures to derive judge, item and person parameters has important uses for judgement-based assessment (Pollitt, 2004, 2012). This appears sustainable from Pollitt's reported research; however, one may doubt the applicability of ACJ to WBOA on the grounds that often WBOA does not produce an artefact for multiple judges to view, which Pollitt's methods would require. However, videos of workplace performances (or simulations), or some suitable artefacts, could be used for ACJ, even if this was to a limited extent operationally, or in standardisation, and with relevant caveats, to at least provide some indication of the extent of assessor agreement and severity. We believe that this would be a useful avenue to explore in future research.

- *The concept of 'fail' doesn't really exist, and there are no (or misleadingly few) zeroes in the data.*[36]

This is a problem for maximum likelihood estimation approaches to estimating item and person parameters in IRT, because such approaches involve finding the place where the curve for the probability of a correct response crosses the curve for the probability of an incorrect response, so as to find a location on the ability – difficulty scale for person ability and item difficulty.

[36] In other words, even if a 'fail' does happen, it is typically not recorded in the relevant databases or portfolios.

- *Data are in sparse and very unbalanced matrices.*

This makes it rather difficult to do g-theory using operational data from WBOA in VQs. Some g-theory software (e.g., EduG[37]) can function with variance components derived from other methods for deriving ANOVA variance components; some of which can work in the case of sparse, unbalanced and non-normal data. Nonetheless, this is a novel and potentially controversial approach. Johnson, Johnson, Miller, & Boyle (2012) were able to do g-theory on data from an item-banked test, because (effectively) all candidates in that test answered the same number of questions. But this is far from the case for WBOA; candidates are observed different numbers of times, and there are large parts of any data matrix that are unpopulated. This is unfortunate; g-theory seems like a highly powerful approach that could have many uses for WBOA. However, any approach that we can envisage operationalizing is likely to be at best non-standard, and possibly of dubious provenance.

- *Properties (such as spread, skew, kurtosis, etc.) of data from WBOA differ considerably from those of other qualification components.*

He (2012) reviewed literature and conducted simulation studies to illustrate techniques for deriving composite measures of reliability for qualifications made up of multiple components. Johnson, Johnson, Miller, & Boyle (2012) used multi-variate g-theory to provide a composite dependability measure for a test made up of two parts. However, Bramley & Dhawan (2012) reported difficulties in applying such techniques to academic qualifications in England. They found that the availability and disparate properties of data between A-level and GCSE components prevented the derivation of credible composite reliability indices. We would surmise that Bramley & Dhawan's (2012) problems would apply 'in spades' when one tried to combine scores from WBOA with those from other components in a VQ.

So, what did we do?

In accordance with our promise to our sponsors, and the drift of argument from this section, we proposed to develop some quantitative indicators of assessment quality, and apply them to WBOA. These indicators purport to show the quality of whole assessments and aspects thereof. They have been developed by analogy to indicators used in existing assessments (mostly tests) and are similar to: facility values, assessor consistency indicators, and reliability indices.

The indices were linked to the relevant inferences in the validation framework. The indicators we developed are 'prototype only', examples of the sort of analyses/indicators that could be used for a

[37] EduG is a specialist statistical software application used to perform generalisability analysis based on the Analysis of Variance (ANOVA) and Generalisability Theory (G-Theory) (Swiss Society for Research in Education Working Group (SSREWG), 2010).

full-scale validation of a WBOA. The indicators were derived from two data sources, which we discuss in detail in the next section.

Data

The quantitative data analysis conducted in this project was based on two data sets. These were:

1) Data file developed for the Harth & Hemker's (2012) study
2) Data collected from Learning Assistant, City and Guilds' own e-portfolio system

These sources essentially contained operational data. This applies to the ways in which the data were structured, even if they had been especially collected. We wanted to investigate how to analyse such 'naturally occurring' data, even though they can be structured in unhelpful ways (many of the challenges to analysis set out in the preceding sections apply particularly to operational data although such issues do also apply to an extent in the case of specially designed experiments). We believed that it was necessary to extract as much information as possible from data sources that were generated in the regular conduct of operational assessment; not to do so seemed wasteful. Also, whilst researchers can set up experiments to have controlled features, they do also have weaknesses. Subjects may behave differently in the experimental situation than they do in 'real life' and also experiments are inherently expensive, slow and resource intensive. One could only ever do few of them and small ones at that (cf. Johnson, Johnson, Miller, & Boyle, 2012).

In the sections that follow we describe our data sources in more detail.

The Harth & Hemker (2012) data file

City and Guilds researched the reliability of portfolio-based assessment as part of Ofqual's reliability programme. That study occupied most of 2010, the data (the portfolio records of assessor and IQA judgements for LOs in several qualifications) being collected in the first half of that year. Data were in respect of three qualifications:

- 2356: Electro-technical services (Level 3) (NQF)
- 3008: Cert/Dipl Hairdressing (Levels 1,2,3) (QCF)
- 3014: NVQ Hairdressing (Levels 1,2,3) (NQF)

Portfolio evidence was entered into a comprehensive spreadsheet. This document was designed in a 'long format' so that each assessment decision occupied a distinct row in the table. (Harth & Hemker, 2012) reported that the construction of this data table required both considerable time and interpretation by trained researchers to allocate judgements entered on portfolio documents to particular categories in the spreadsheet. Somewhat surprisingly, it appeared that many assessment centres' portfolios were locally

devised, and were therefore idiosyncratically structured. This is despite the awarding organisation providing standard logbooks, which are used by many centres.

The development of this data file had been a substantial exercise, and we therefore believed that it would be useful to re-analyse it for the current project. However, this is with the caveat that this file was not designed for the analyses that we conducted. This is a caution that readers should bear in mind when interpreting findings.

The data file we worked from had over 21,000 rows with an assessment decision in each. The recorded assessment decisions were divided between the three qualifications listed above. The following table, which pivots the assessment type against the three qualifications, suggests that observation was the major assessment type in Harth & Hemker's (2012) data file, which supported our decision to re-analyse it.

Table 16: Number of assessment decisions of different types in qualifications

Assessment type	Qualification number			Total
	2356	3008	3014	
Observation	3,887	1,662	8,001	13,550
Written multiple choice questions		1,845	819	2,664
Oral questioning multiple choice		25	43	68
Oral open ended questions	824	166	155	1,145
Witness testimony / comment	99		2	101
Records of past activity / supplementary / product evidence	35			35
Professional discussion	548			548
Assignment		5	53	58
Review of portfolio evidence / direct observation of performance (interim)	2,179		5	2,184
Documentary / written evidence (e.g., risk assessment document, variation order, technical information / query, or other relevant paper work produced at the time of assessment)	26			26
Candidate self-assessment / reflective account	121			121
Written open ended short answer questions	174	35	111	320
IV checklist of assessor		1		1
IV questions to confirm candidate's understanding		9		9
Observed product	200			200
Missing	71		5	76
Total	**8,164**	**3,748**	**9,194**	**21,106**

Source: Harth & Hemker, 2012

Further understanding of the data in this file can be gained by 'drilling down' to analyse only those assessment decisions coded as 'observation'. Table 17 takes the number of assessment decisions coded as 'observation', and pivots the actual decision against the qualifications studied in Harth & Hemker (2012).

Table 17: Numbers of assessments in judgement categories for observational assessments

Major judgement category	More granular judgement category	Qualification number			Total
		2356	3008	3014	
Not yet achieved	Not meeting the required standard/ referred assessment	34	48	31	113
Achieved	Meeting the required standard/ NVQ competent	3,580	1,459	6,922	11,961
Achieved	With comments/ action points	12		48	60
Not yet achieved, ready for assessment	Ready for summative assessment	1	3	54	58
Not yet achieved	Could not carry out the task due to external circumstances (e.g., client allergic to hair colouring product, lack of opportunity, n/a)	1		68	69
Not yet achieved	Needing further development	1		447	448
Not filled in	Box empty/ no records available	200	140	348	688
Not yet achieved	Not evidenced/ insufficient evidence to judge – action planned	24	1	44	69
Not yet achieved	Not evidenced – still to be assessed	1	9	38	48
Total		**3,854**	**1,660**	**8,000**	**13,514**

This table illustrates some of the issues around the structure of data gathered from observational assessment. The table confirms that a large majority of observations result in a designation of 'achieved'. However, this does have various 'flavours'; many candidates were deemed to have achieved unreservedly, whilst others achieved, but received some form of feedback or commentary. This table also points at the differing reasons for candidates not achieving – in some cases because they had not met the standard, or in rather more for a variety of reasons, some of which might imply the candidate had not attained the standard. The table also highlights the substantial amounts of missing data on the portfolio forms.

In further analysis of Harth & Hemker's (2012) data file, we concentrated on the 3008 Hairdressing qualification, and two units (GH10 – Style and finish hair – and GH11 – Set and dress hair), which coincided with the data we requested from Learning Assistant, and with our centre visits. We present these analyses in the Results of quantitative analyses section below.

E-portfolio data

The portfolio evidence gathering for Harth & Hemker's (2012) project was time consuming, and the manual data entry a potential source of error. Therefore, we prevailed upon our colleagues to extract some data from Learning Assistant, City and Guilds' own e-portfolio system.

Our colleagues provided data from the following units that are part of the 3008 Hairdressing qualification:

- GH8 – Shampoo, condition and treat the hair and scalp
- GH10 – Style and finish hair
- GH11 – Set and dress hair

The data were from five centres and in the following quantities, and were provided in July 2012.

Table 18: Summary of 3008 data downloaded from the e-portfolio Learning Assistant

Centre number	Units			Total
	GH8	GH10	GH11	
1	589	569	461	**1,619**
2	140			**140**
3	26	337	198	**561**
4	30	8		**38**
5	448	543		**991**
Total	**1,233**	**1,457**	**659**	**3,349**

Downloading these data was certainly more time- and cost-effective than collecting paper portfolios from around the country. However, even these data illustrate some of the vicissitudes of data held on VQs. Firstly, it can be seen that the data are rather unevenly distributed across the five centres; for unit GH11 there is only data from two centres for instance. Also, the e-portfolio data captured candidates who were in mid-course, as well as those who had completed. Thus, the data were best regarded as a 'snap-shot' of performance at a particular point in time, rather than a definitive basis upon which to evaluate measurement properties of the 3008 qualification.

Notwithstanding such concerns, however, these data did provide a basis for some indices that could be derived to show the measurement properties of observational assessment outcomes.

Results of quantitative analyses

In this section we summarise the results from our quantitative analyses. As hinted earlier, these analyses are designed to illustrate the types of analyses that could be conducted to increase understanding of the reliability and validity of WBOA. This work is exploratory, it does not amount to a full validation of any inference within the validation framework. Example analyses are based on analogies to common techniques employed in analysing test data (see Common statistical indicators of the quality of assessments section at p. 88), and are organised according to fit with a particular section of the validation framework.

Scoring/quasi-facility values

Table 19 sets out an analogy to facility values for the units in the 3008 Hairdressing qualification. This analysis was conducted on Harth & Hemker's (2012) reliability study data, and pertains to assessments that were coded as the 'observation' type and where 10 or more observational assessment decisions were recorded for each particular unit.

Table 19: Quasi-'facility values' for units in 3008 qualification (Harth & Hemker, 2012)

Units	Not yet achieved	Achieved	Not yet achieved, ready for assessment	Not filled in	Not yet achieved	Not yet achieved	*N*
	Not meeting the required standard/ referred assessment	Meeting the required standard/ NVQ competent	Ready for summative assessment	Box empty/ no records available	Not evidenced/ insufficient evidence to judge - action planned	Not evidenced - Still to be assessed	
G1	0.00	0.82	0.00	0.12	0.06	0.00	**17**
G17	0.00	0.85	0.00	0.15	0.00	0.00	**84**
G18	0.00	1.00	0.00	0.00	0.00	0.00	**25**
G2	0.00	0.95	0.00	0.05	0.00	0.00	**167**
G20	0.00	0.89	0.00	0.10	0.00	0.01	**183**
G21	0.00	1.00	0.00	0.00	0.00	0.00	**53**
G3	0.00	0.85	0.00	0.11	0.00	0.05	**66**
G4	0.00	0.85	0.00	0.15	0.00	0.00	**26**
G7	0.04	0.96	0.00	0.00	0.00	0.00	**81**
GH1	0.04	0.80	0.01	0.14	0.00	0.00	**70**
GH10	0.05	0.80	0.00	0.14	0.00	0.01	**182**
GH11	***0.30***	***0.70***	0.00	0.00	0.00	0.00	**46**
GH12	0.06	0.92	0.00	0.02	0.00	0.00	**53**
GH2	***0.10***	***0.74***	0.03	0.14	0.00	0.00	**73**
GH22	0.00	1.00	0.00	0.00	0.00	0.00	**32**
GH3	0.00	0.89	0.00	0.07	0.00	0.04	**55**
GH4	0.00	0.88	0.00	0.13	0.00	0.00	**32**
GH8	0.02	0.89	0.00	0.09	0.00	0.00	**314**
GH9	0.02	0.90	0.00	0.08	0.00	0.00	**51**
H32	0.00	1.00	0.00	0.00	0.00	0.00	**21**
Mean	**0.03**	**0.88**	**0.00**	**0.07**	**0.00**	**0.01**	**1,631**

Amongst all the 'unit facility values', only two seem to diverge from the previously noted tendency that nearly all candidates are deemed to have achieved the standard when being observed. These units have been highlighted in the table, and are G2 (Assist with reception duties) and GH11 (Set and dress hair); the latter being one of the units that was studied in further depth in this project (see below). The column 'box empty/ no records available' is the highest value after 'meeting the standard', showing the difficulties the researchers encountered in interpreting incomplete records.

Generalisation/Fisher information function

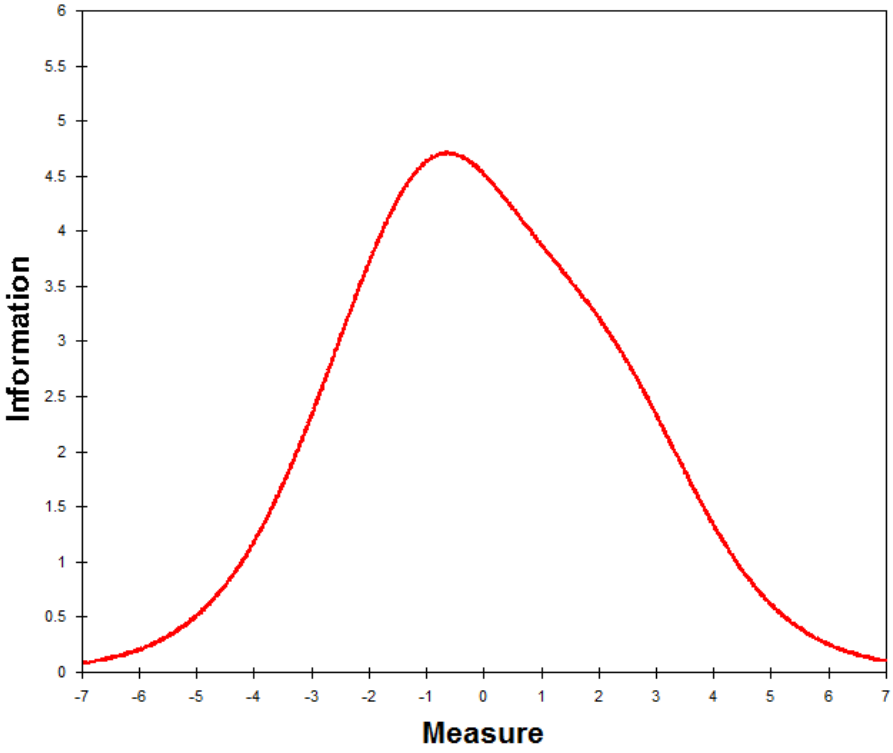
(Harth & Hemker, 2012) applied CTT techniques to derive reliability indices for the case of portfolio assessment, and (Johnson, Johnson, Miller, & Boyle, 2012) conducted a thorough g-theory analysis of a VQ written test, including the derivation of various types of dependability coefficients.

Rather than repeat either of those approaches to quantifying the generalisation of scores from a particular measurement procedure, in this example we derived a Fisher information function for observational assessment in the 3008 Hairdressing qualification (based on Harth & Hemker's data).

All assessment decisions coded as 'observation' for 3008 (including those with fewer than 10 observations) were formatted and loaded into the Winsteps Rasch model IRT software (Linacre, 2009). Judgements of 'competent/meeting the NVQ standard' were coded as '1' or 'correct' and all other outcomes were coded as '0' or 'incorrect'.

Overall infit and outfit statistics suggested that the data were 'productive for measurement' or (more colloquially) 'fit the Rasch model'. Figure 4 displays a 'Test (Fisher) Information Function' derived from Rasch analysis of that data set:

Figure 4: Test Information Function for observational assessments on 3008 qualification



The curve suggests that the observational assessment is providing a reasonable magnitude of measurement information, but that more information is being provided towards the top of the measure continuum (the 'bulge' on the red line).

This finding exemplifies one of the paradoxes inherent in the judgemental data. The Winsteps input file was derived from Harth & Hemker's (2012) reliability analysis spreadsheet and hence each assessment decision was coded as a separate row. Thus, a candidate doing multiple observed performances and coded as 'meeting the standard' each time would be 'assumed' by the Rasch model to be a strong student; as they would be repeatedly 'answering questions correctly'. However, our intuition is that if an assessor feels that a candidate needs repeated observation, it might well imply that that candidate is a little weak.

Scoring/why were candidates observed more than the minimum number of occasions?

To mitigate the 'fallacy' inherent in the 'naïve' IRT analysis above, we trialled a different approach using e-portfolio data. All candidates in unit GH10 have to demonstrate their competence of each LO on at least three occasions. Therefore, we selected the assessments of candidates who had been observed four or more times (on the grounds that this would mean that they had 'failed', i.e. did not demonstrate at least one LO at least once). Then, we noted which LO (or LOs) was (were) absent for each observation. Finally, we counted which LO(s) were absent for each observational occasion. The results are shown in Table 20.

Table 20: Table counting which LOs were absent during successive observations in the GH10 unit

No. of occasions LO was absent during the Nth observation	Outcome number and name					Totals
	1	2	3	4	5	
	Maintain effective and safe methods of working when styling and finishing hair	Blow dry hair into shape	Finger dry hair into shape	Finish hair	Provide aftercare advice	
first	22	36	107	29	26	**220**
second	22	40	112	24	26	**224**
third	34	42	124	35	34	**269**
fourth	31	49	114	35	33	**262**
fifth	30	44	100	32	31	**237**
six	22	32	78	23	20	**175**
seventh	15	23	60	17	16	**131**
eighth	7	15	42	9	8	**81**
ninth	4	6	32	5	6	**53**
tenth	6	7	23	7	8	**51**
eleventh	2	4	7	3	4	**20**
twelfth	0	2	3	1	2	**8**
thirteenth	0	0	3	1	1	**5**
fourteenth	0	1	2	1	1	**5**
fifteenth	0	0	1	0	0	**1**
Total	**195**	**301**	**808**	**222**	**216**	**1,742**

Several (tentative) conclusions can be drawn from this table. Firstly, it is surprising that some candidates appear to be observed up to 15 times. This does appear to be a substantial assessment burden. Equally, LO 3 'Finger dry hair into shape' is overwhelmingly the most frequently absent LO. This may give us a hint that this part of the assessment is difficult for candidates. However, our intelligence from hairdressing experts suggests that, rather, finger drying hair into shape is not something that customers frequently ask for. Thus, this finding says little about the difficulty of that LO. However, it might have implications for those engaged in updating standards; for example this LO may be omitted in a revised version and other cuts or techniques that are frequently requested may be included; or the assessment may be done on non-paying clients (cf. the discussion regarding negative washback on p. 75).

Behaviour of assessors in centres

As well as the difficulty of parts of assessments, a key issue for validating observational assessment concerns the way that judges use rating scales. To prototype a possible method of investigating this phenomenon, once again Learning Assistant data for unit GH10 was used.

We took assessors' ratings of candidates in the e-portfolio data. These were allocated into the following categories:

0 = No tick[38]

1 = Met (Candidate or Assessor on behalf of candidate)

2 = Met (Assessor)

3 = Question/query (Assessor)

4 = Met (Verifier)

A contingency table was set up showing how frequently assessors allocated candidates' observed performances into one of the five categories listed above. That table was analysed using the Pearson chi-squared test. That analysis sought to establish the extent to which ratings were independent of assessors. The null hypothesis was that there was no relationship between rating patterns and assessor. The alternative hypothesis was that there was a significant relationship between rating behaviour and assessor; for instance, that certain assessors were using the rating scale in an identifiably different way to the overall trend.

When the chi-squared statistic for the contingency table was found to be significant, follow-up analysis was conducted. The 'adjusted residual' for values in particular cells was calculated. The 'adjusted residual' is a standardised measure of the distance between an expected measure in a cell and the observed value. A value of greater than 2 or less than -2 on the adjusted residual tends to indicate that that particular cell is contributing significantly to the rejection of the null hypothesis. It was surmised that scrutinising contingency tables for patterns in adjusted residuals might give investigators insights about why the chi-squared statistic was significant for particular LOs.

In fact, on an initial run of 'assessors-by-rating' contingency table, it was found that many cells had observed values of less than five. This runs the risk of reducing the power of the calculated statistic (increasing the chance of a 'false negative error'; that is retaining the null hypothesis when it ought to be rejected). Therefore, assessors were grouped into their centres – thus allowing us to investigate rating behaviour between centres, rather than individual assessors.

[38] The Learning Assistant data does not distinguish between zero (for incorrect) and some other symbol for 'omitted' or 'not achieved'.

Table 21: chi-squared values and significance for assessor-by-rating category tables

LO number	LO name	Chi-squared value	significance (p value)	sig at p<0.05
1	**Maintain effective and safe methods of working when styling and finishing hair**	30.888	0.0003	YES
2	**Blow dry hair into shape**	18.518	0.0296	YES
3	**Finger dry hair into shape**	6.4434	0.6949	NO
4	**Finish hair**	27.8405	0.0010	YES
5	**Provide aftercare advice**	23.9445	0.0044	YES

The chi-squared values are significant for LOs 1, 2, 4 and 5, but not for LO3. This is consistent with the finding (see Table 20 and surrounding discussion that patterns of ratings for LO3 appear rather different to patterns for the other LOs.

By way of illustration[39], cell values and standardised residuals for LO1 are shown in Table 22 overleaf. Adjusted residuals with an absolute value of greater than two are emphasised by shading in the table.

Table 22: Contingency table showing adjusted residuals for LO 1

			Assessor/verifier rating for LO 1				Total
			0	**1**	**2**	**4**	
Centre number	**1**	Count	19	<5	71	<5	**91**
		Expected Count	25.0	n<5	64.4	n<5	**91.0**
		Adjusted Residual	-1.7	n<5	1.8	n<5	
	3	Count	51	<5	90	<5	**143**
		Expected Count	39.4	n<5	101.1	n<5	**143.0**
		Adjusted Residual	**3.1**	n<5	**-2.9**	n<5	
	4	Count	<5	<5	<5	<5	**5**
		Expected Count	1.4	n<5	3.5	n<5	**5.0**
		Adjusted Residual	n<5	n<5	n<5	n<5	
	5	Count	5	<5	41	<5	**48**
		Expected Count	13.2	n<5	34.0	n<5	**48.0**
		Adjusted Residual	**-2.9**	n<5	**2.5**	n<5	
Total		**Count**	**79**	**<5**	**203**	**<5**	**287**
		Expected Count	**79.0**	**<5**	**203.0**	**<5**	**287.0**

Centre 3 appears to be making a significant contribution to the chi-squared value in LO1. This may indicate that that centre is using the assessment categories aberrantly; however, the high residual values may just be an artefact of centre 3 conducting the most assessments in this data set. Perhaps more worthy of note is that centre 5 seems to have consistently high[40] residual values (higher than centre 1, which has more assessments in the data). This may suggest that a validator, external verifier, etc. should focus on that centre to establish if assessments were being carried out properly.

Our evaluation of the potential of this analysis technique should outline some of its possible strengths and weaknesses. Firstly it is a strength that we have applied a widely-used non-parametric statistic to an

[39] The patterns of adjusted residuals for the other LOs were broadly similar to those for LO1.
[40] They are high for LOs 2, 4 and 5 as well.

e-portfolio data set. It is necessary to use such 'generic statistics' to get an understanding of one's data sets as an initial step. It is possible to go on to use more specialised, derived indices of test and assessor characteristics. But the danger of leaping straight to such indices is that one loses sight of what the raw data actually mean. Further, given the highly skewed nature of data from observational assessments, it is right to use a non-parametric approach.

There are several limitations to the proposed approach, however. Firstly, there are issues with the source data. Like the data that Harth gathered for the previous Ofqual project, the Learning Assistant data are organised into a range of somewhat idiosyncratic categories; it would be better if we could distinguish 'failed' from 'not attempted' and also we need to consider some of the other 'not-quite passed' categories. Clearly, users in Harth's sampled portfolios and in the Learning Assistant system feel the need for such categories, but a standardised set of categories should be agreed upon.

The chi-squared and adjusted residual values do not comment directly on rater severity. If an observed cell value in column '2' i.e. LO met is above its expected value and the residual is significant, one may assume that that centre/assessor, etc. is lenient. But these data contain no linking design across assessors or centres (such entities may be awarding more passes because their candidates are genuinely stronger). It is possible that verification decisions could be used as a way of 'anchoring' decisions between assessors and/or centres. Such anchored data could be analysed using an IRT approach to derive sample independent measurement – for example using the Facets software.

Generalisation given different lengths of assessments

The final analysis using existing assessment data investigated the generalisability of results of assessments made up of different numbers of observations. Candidates must demonstrate acceptable performance of all five LOs in the GH10 unit at least three times in order to pass the unit. However, we have already seen (Table 20.

Table 20, above) that many candidates were observed many more than three times. Is this an effective measurement process, ensuring robust decisions? Or do the repeated observations (especially those that essentially seek to confirm one 'missing' LO, or one missing part of the range) constitute over-assessment and inefficiency?

Our intention was to use a relevant statistical technique to describe the reliability of outcomes from the extant measurement procedure, and then conduct follow-up analysis to impute potential reliabilities for assessment procedures involving more or fewer observations. To carry out this procedure, we investigated the applicability of two techniques; 1. Using the Spearman-Brown prophecy formula from within Classical Test Theory (CTT) (Haertel E. H., 2006, p. 77), and 2. a d-study from within the generalisability theory tradition. In fact, we were not able to make either of these techniques provide meaningful output. In this

section, we describe our attempts, and point to the limitations in data that prevented us from modelling this feature of the measurement procedure.

First of all, we calculated the Cronbach's alpha reliability index for the observed assessments of the GH10 unit. Cronbach's alpha – like all approaches that model variance within and between items – requires a 'fully-crossed' data matrix to function. That is, within the data set, each candidate should respond once and only once to each item. Further, each candidate should respond to the same total number of items. The observational data from Learning Assistant does not possess these properties. Candidates are assessed against the same LO repeatedly and they are also assessed on different numbers of occasions.

To mitigate this weakness of the data set, we derived a summary score for each candidate. That was a 'p-score' or an average; that is to say, the number of times that a candidate had been credited as having achieved the LO, divided by the number of times they were observed. This arrangement – whilst bringing some problems – had the benefit of providing a fully crossed data matrix between candidates and LOs.

Cronbach's alpha was calculated for 29 candidates who had completed the GH10 unit. The value of alpha, taking the 29 completed candidates and the five LOs as items, was 0.881. This was a high value; broadly in line with high reliability values found by Harth & Hemker (2012). As well as calculating the Cronbach's alpha summary statistic, we output an ANOVA table. This is displayed below:

Table 23: ANOVA table accompanying reliability analysis of GH10 data

		Sum of Squares	df	Mean Square	F	Sig
Between People		135.926	28	4.854		
Within People	**Between items**	611.224	4	152.806	264.957	.000
	Residual	64.593	112	.577		
	Total	675.817	116	5.826		
Total		811.743	144	5.637		
Grand Mean = 5.3405						

The most informative aspect of the ANOVA table is the ratio of squares (sums of squares and/or mean squares) – rather than the summary F-statistic. Perusing the sums of squares column, we may see that the between persons variance is small – unsurprising in a VQ in which the aim is to certify that individuals are beyond a threshold, rather than to spread those people along a measurement scale. However, also, the variance between items is relatively high compared to the residual. This is consistent with the high reliability values returned in the Cronbach's alpha index.

The Spearman-Brown prophecy formula can be applied to a classical test theory reliability index to simulate the reliability of a test of longer or shorter length than that that was actually analysed. However, in the current case, 'items' were effectively LOs; thus Spearman-Brown could simulate the reliability of a measurement procedure with more or fewer LOs, but not with more or fewer observations.

In our investigations we organised the data in different ways in order to attempt to apply the Spearman-Brown formula. Firstly, we analysed candidates' total scores using the Kuder-Richardson 21 reliability formula (Haertel E. H., 2006). However, the results of that analysis were improbably low. This was because of low variance amongst the total scores (which were in any case central tendencies). Also, we established a matrix, treating LOs as 'persons' and 'numbers of observations' as 'items'. However, in contrast to the previous attempt, this procedure produced implausibly high reliability values.

Not being able to model the impact of different numbers of observations using CTT, we turned to generalisability theory. That sophisticated modelling approach allows one greater freedom than CTT to model a range of differentiation and instrumentation facets. Further a d-study within this tradition has considerable flexibility to model the impact of changes on any facet within the measurement design.

We were able to derive a generalisability index for a simple crossed design of the facets 'persons' and 'items' (LOs); in essence this gives equivalent information to the Cronbach's alpha index reported above. We also attempted to express a measurement design that would represent the data as it was structured in the output from the e-portfolio. Unfortunately, we have been unable to produce a viable model for the three facets 'candidates', 'Items (LOs)' and 'numbers of observations'. We believe this is because the facet 'observations' is nested both within the differentiation and the instrumentation facets – a design that is not permissible within g-theory.

Summary

In this section we have demonstrated that it is possible to derive some meaningful and potentially useful indices/approaches to provide quantitative evidence of the validity of WBOA. These include:

- An approach to understanding something akin to item facility at the level of LOs in WBOA (DOMAIN DESCRIPTION, assumption 5)
- An investigation of absent learning outcomes, which might say something about their relative difficulty, or whether their frequency of occurrence in practice justifies insisting on their assessment in entirely 'genuine' circumstances (DOMAIN DESCRIPTION, assumption 4; UTILIZATION, assumption 2)
- Using chi-squared as an example of how to say something about the severity or leniency of assessors (JUDGEMENT, assumption 3)
- Deriving a value of Cronbach's alpha reliability index for the observed assessments of a unit of assessment (a replication of Harth & Hemker's (2012) work) (GENERALISATION, assumption 1)
- We believe that these and similar approaches could become more thoroughly researched and perhaps applied in practice if e-portfolios were to become more widely used, as data collection would be more efficient.
- Furthermore, e-portfolios could enable awarding organisations to standardise assessment practice

almost ‘covertly’ (Boyle, 2009, p. 33). The e-portfolio can be updated by its owner and centres can be ‘led’ to provide data to the e-portfolio owner’s specification. Quantitative and other analysis of this data would become even more straightforward if this specification and the design of e-portfolios were guided by sound assessment principles and a pre-conception as to what purpose the data collected there could be used for.

The final WBOA validation framework with suggestions for possible evidence collection methods

In this section we list methods that might be appropriate for evidence collection, against relevant inferences and assumptions in the interpretive argument.

Table 24: Validation framework, including suggested methods for evidence collection

INTERPRETIVE ARGUMENT			VALIDITY ARGUMENT
INFERENCE	**ASSUMPTIONS UNDERLYING THE INFERENCE**	**THREATS TO VALIDITY/REBUTTALS**	**EVIDENCE COLLECTION METHODS**
1. DOMAIN DESCRIPTION/theoretical construct definition			
	1. NOS or comparable sets of standards addressed by WBOA representative of relevant job role and of the relevant definition of competence (criterion/target domain) – construct defined appropriately	– NOS problematic as descriptors of criterion domain/construct definition – Lack of consensus among stakeholders on what the important aspects of competence are	– Examining what aspects of certain roles/sectors are considered important – e.g., Kelly's Repertory Grid technique, stakeholder questionnaires and interviews – and compare to what currently stands for the construct – Outside of direct influence of AOs. Relevant SSCs responsible, so investigation of their procedures, consultation processes with stakeholders etc. would also be necessary
	2. LOs and ACs appropriately reflect the construct	– LOs and ACs do not represent the construct and the criterion domain appropriately	– Establish what the construct should be (see above) and then cross-check if this is appropriately expressed through LOs and ACs
	3. LOs and ACs appropriate for assessment by WBOA	– Alternative assessment methods required – Simulation (or witness statement, report, other evidence types) used instead of real workplace tasks for inappropriate reasons – or vice versa	– Research and understand strengths and weaknesses of different assessment methods that could be used in the workplace in order to be able to judge the appropriateness of WBOA – produce a comprehensive inventory which could be reused in validation
	4. Appropriate assessment tasks/situations used for assessment	– No opportunity to collect evidence or demonstrate competence for certain ACs and LOs in the workplace setting – Cognitive/psychomotor/etc. demands of tasks/situations do not match LOs and ACs – Demands too variable across candidates and occasions	– Establish whether QA procedures cover issues like this – Questionnaires/interviews with relevant practitioners to establish how widespread and likely threats like this are in individual qualifications – Counting 'absent LOs' (cf. the relevant section on p. 100)

	5. Statistical characteristics of tasks appropriate	– Task difficulty too variable (e.g., situation/customer complexity); especially relevant where it is left up to the assessor or candidate to select which tasks will be observed		– Establish whether QA procedures cover issues like this – Interviews with practitioners to investigate whether variability/situational complexity is part of the construct – Statistical indicators – e.g., 'quasi facility values' (Table 19, at p. 98)
2. ADMINISTRATION				
	1. Learners are sufficiently made aware of properties of WBOA assessment procedures and of the criteria and standards applied in assessment	– Learners lack understanding of assessment procedures, criteria and standards – Learner behaviour adversely affected by observation process – Assessment anxiety (less likely if assessment is repeatable, as is often the case with WBOA); possibly higher if assessed by peripatetic rather than work-based assessor		– Check AO documentation – availability, content and ease of access – Check whether QA procedures deal with these threats + interview/questionnaire QA personnel – Questionnaires/interviews with candidates to investigate their views of these three threats – Compare experiences of candidates assessed by work-based vs. external assessors (interview/questionnaire)
	2. Task administration conditions are appropriate for providing evidence of targeted competence	– Inappropriate assessment conditions		– Check whether QA procedures deal with this threat + interview/questionnaire QA personnel
	3. Assessor conduct appropriate and consistent with relevant guidance and best practice	– Inappropriate help by assessors – Personality clashes/bias		– Check guidance to assessors – Check whether QA procedures deal with this threat – Interview/questionnaire with IQAs/QCs to establish prevalence and seriousness of this in individual qualifications
	4. Relevant verification procedures effective	– Verification ineffective or does not address administration issues		– All of the above checks should give an indication of the relevance and effectiveness of verification in this domain

3. JUDGEMENT			
	1. Observation tools/checklists are appropriate and used appropriately for recording and judging evidence of relevant competence and promoting reliable judgement	– Design and wording of observation tools/checklists does not support reliable judgement/reliability analyses – Observation tools/checklists do not capture important qualities of task performance – Observation tools/checklists not used where recommended, or used inappropriately	– Investigate checklists used – compare with best practice in similar assessment contexts – Questionnaires/interviews with practitioners to establish whether checklists embody the relevant constructs/what is important
	2. Assessors are sufficiently trained in observation techniques and carry out observation according to best practice	– Assessors not trained in observational techniques – Assessors fail to observe the performance appropriately, miss important aspects, intrusive etc.	– Check training procedures and what they involve – Interview IQAs/QC to get an indication of the prevalence of problems in observational techniques
	3. Assessors apply relevant criteria/standards and observation tools appropriately and consistently	– Assessor inconsistency or lack of understanding and application of criteria and standards (+ possible differences between work-based and visiting assessors) – Inappropriate weighting given to different aspects of performance – Positive or negative bias – Severity/leniency	– Interviews/questionnaire with assessors re their understanding and application of standards – Empirical multiple-observation study (e.g., using videos of performance, compare standard approach with comparative judgement techniques) – IQA/QC agreement with assessors – use available data (e.g., in e-portfolios where/if available), or design empirical study – Statistical indicators – chi squared and adjusted residuals – see Table 21 and Table 22
	4. Assessors require appropriate amount of evidence in order to reach a decision	– Assessors over/under assess (+ possible differences between work-based and visiting assessors)	– Check whether QA procedures deal with this threat + interview/questionnaire QA personnel – Statistical indicators – count of number of occasions particular LO absent (see Table 20 at p. 100) and 'Goldilocks' graph of bullish, cautious and 'just right' assessors at p. 28

	5. Decisions based on WBOA appropriate	– Problems with classification consistency at WBOA level (too many false positives; false negatives perhaps not a major problem as new assessment opportunities are always available – problem: over-assessment) – Assessor decisions affected by funding pressures (+ possible differences between work-based and visiting assessors)	– Interviews/questionnaires with relevant practitioners (assessors, IQAs, QCs) to estimate seriousness of these threats in individual qualifications – Empirical assessor/IQA double marking study could also give an indication of problems in classification consistency – Statistical indicators – cf. range of reliability indices developed by (Harth & Hemker, 2012)
	6. Relevant verification procedures effective	– Verification procedures too infrequent (lack of investment) – Verification procedures focus on paperwork and procedural compliance rather than judgement and assessor standardisation	– Compare frequency of QC visits with notional best practice – Investigation of QA procedures and documentation – Questionnaire/interviews with relevant practitioners could give an indication of the seriousness of these threats
4. GENERALISATION			
	1. The sample of observations is representative of the universe of generalisation, OR assessment focuses on what is proven to be important to stakeholders (as indicated by the construct)	– Inappropriate sampling of tasks from assessed domain – Standardisation of tasks; conditions of assessment too variable; extreme tasks/situations (unless warranted by construct definition) – Task specificity issues	– Consult practitioners regarding sampling and what is important – Reliability studies; D-studies – Identify aspects that might be standardised but are not and vice versa
	2. The sample of observations is large enough to control random error/to enable confident judgement	– Too few observations – Too much variability in the minimum number of observations used – Assessors lack confidence in their decisions (+ possible differences between work-based and visiting assessors)	– Consult practitioners regarding the appropriateness of the number of observations used – Investigate whether there is a notion of best practice with respect to these issues or attempts to establish it – Focus on sectors where external (rather than work-based) assessors are in the majority – possibly a higher threat

5. EXTRAPOLATION			
	1. There are no systematic errors that are likely to undermine the extrapolation at the level of WBOA (↺ DOMAIN DESCRIPTION)	– Parts of criterion domain that should be assessed by WBOA not assessed or given little weight (construct under-representation) – Conditions of assessment too constrained/task fidelity issues – Failure to make explicit any legitimate limitations of the assessment that should limit extrapolation	– Cross-reference with findings relevant to DOMAIN DESCRIPTION – Consult relevant stakeholders (including past candidates) for opinion of the extent to which they consider the assessment situations to be different from what actually happens in the workplace; also regarding transferability of knowledge – Check qualification documentation to establish whether relevant limitations to interpretation are spelled out in sufficient detail to be clear to stakeholders (especially candidates and employers)
6. UTILISATION			
	1. WBOA results considered useful by relevant stakeholders for informing decisions about qualification-level competence	– Stakeholders not satisfied/confident about the contribution of WBOA results to overall decisions about candidate competence	– Consult stakeholders – especially those who make decisions at qualification level (if different from those who decide at WBOA level)(questionnaire/interviews)
	2. WBOA positively impacts the learning and instruction process	– Inappropriate assessment strategy dominates the learning and instruction process – negative washback – No feedback or inappropriate feedback due to lack of time in the workplace, lack of training/awareness of assessors in this respect, etc. – reduced chance of improving future performance	– Consult stakeholders regarding washback – Investigate quality and effects of feedback (consult QA personnel, candidates, assessors); compare with best practice in comparable contexts – Counting ‘absent LOs’ (cf. the relevant section on p. 100)

Discussion

In this section, we highlight some of the key points made in this report and discuss some of their further implications. We highlight the limitations of this research where appropriate throughout the discussion rather than in a separate section.

Building a framework from first principles

The Alpha-Plus validity researchers (AlphaPlus Consultancy Ltd, forthcoming) accepted a validation framework laid down by the sponsor and qualifications regulator, Ofqual (Opposs & He, 2011) and applied it in validation of VQ qualifications. However, our fellow VQ researchers feel it necessary to state that:

> *… 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.* (AlphaPlus Consultancy Ltd, forthcoming, p. 36)

In contrast, in the relevant section of our report (see Review of existing argument-based validation frameworks at p. 43) we reviewed several approaches to developing argument-based validation frameworks, and proposed strengths and weaknesses of each. We also extensively reviewed and discussed the properties of WBOA that require validation treatment that is different from what might be the case in standardised tests (and also highlighted where there should not be any difference, but pointing out that any assessment method, including WBOA, should be subject to similar validation criteria at a general level). Thereafter, we derived and then evaluated a framework applicable to WBOA. We suggest that this – along with the long review sections culminating in the interpretive argument concerning WBOA – is the type of laborious scything of undergrowth that we refer to in our introduction. However, we also suggest that this exercise is worthwhile in itself – both in that it provides a sound theoretical basis upon which to understand WBOA specifically, and insofar as it provides insight into the practice of validation generally – both in vocational and general assessment.

There remain some matters to be resolved with the validation approach we have suggested. Firstly, we would acknowledge that validation using our framework remains a complex activity. If the Ofqual framework represents a ‘complex and multi-dimensional puzzle’ (AlphaPlus Consultancy Ltd, forthcoming, p. 14), this applies to the City and Guilds WBOA validation framework equally. Secondly, there were occasions during questionnaire analysis in which it was difficult to see where perfectly reasonable questionnaire items (for example items concerning construct-irrelevant variance and construct

under-representation) fitted into the framework. The problem here may be that validation frameworks as instruments inevitably are – at the highest level – defined in a very general way. (The top level categories in our framework amounted to a single word.) The danger is that such categories then are rather trite and lack meaning; for example, if construct under-representation might plausibly fit into EXTRAPOLATION or DOMAIN DESCRIPTION, then can these broad categories really have much meaning?

When interpreting some of our analyses we came up against an old problem in a new guise. Reliability investigators have known for many years that it is difficult to establish a precise cut-off between acceptable and unacceptable values on reliability indices (Baird, Beguin, Black, Pollitt, & Stanley, 2012, p. 783). However, in interpreting questionnaire and interview outcomes, we found that there were occasions when substantial majorities of respondents seemed to back the validity of WBOA, but some non-negligible minority questioned the validity. In such circumstance, the issue is how to interpret the existence of that minority. Should we just ignore minority views and accept 'the tyranny of majority opinion'? Or should we pay important heed to honestly held, possibly consequential dissident opinion? This is clearly a matter of degree and particular circumstance, but it is one of the difficult questions for those employing validation frameworks.

Finally, there are criticisms of the argument-based approach to validation referring to it being atomistic, losing sight of the bigger picture, there being difficulties in prioritising likely threats to validity, and difficulties in deciding which assumptions might be taken for granted and which might need serious attempt at validation, etc. We believe that these are reasonable objections, and that further research should focus on these problems. We also believe that prioritisation of threats to validity in a method and/or qualification, requires combined efforts and insight of assessment specialists and the relevant stakeholders, and thus suggest that the stakeholders are included in the future research addressing the abovementioned problems. Ultimately, a problem parallel to 'how much reliability is enough?' is also 'how much validity is enough?', and this is something that is bound to arise in any validation efforts, what with validity arguably being a matter of degree.

We are conscious of the size and level of detail of the framework, and appreciate that it might be difficult to implement operationally very frequently. On the other hand, we believe that time could be saved by targeting validation (and verification) efforts better (in which, we believe, our framework could be helpful) and avoiding validation studies that are little more than mere compliance exercises that require the collation of large quantities of disparate documents for not-always-obvious reasons. Although we have just stated above that prioritisation of threats to validity and steps in validation should be an empirical issue, our recommendations (see p. 121) could be taken to reflect our current view of what the priorities should be, based on relevant literature reviews and limited evidence from the field.

The definition and preponderance of WBOA

When proposing this research to Ofqual, we took considerable care before arriving on the term 'work-based observational assessment'. We took the view that this was a type of assessment that was common in VQs, yet which had not been widely studied. Rather, other approaches – which tended to be more straightforward to conceptualise – such as external assessment have been studied much more widely.

In studying WBOA in depth, we have invested some effort into clarifying its status in VQ assessment and into defining it in the first place. This has not been a straightforward exercise, given the tradition within the competence-based assessment approach (perhaps, in particular in work-based assessment) of neglecting the role and effects of the procedures for eliciting performances from candidates. We have argued that this is not a helpful attitude, and that it is worthwhile devoting attention to these issues, understanding the properties, advantages and disadvantages of different assessment methods for use in different contexts, and targeting the use of methods to make sure that they allow us to assess what we are interested in, while ensuring fairness and consistency as far as this is possible.

We have reviewed relevant literature about performance assessment, both general, and those specific to the context of medical and language assessment, and came to a conclusion that WBOA might be best defined as a performance assessment method. Given the way it is used in VQ assessment, we would also class it as a task-centred method, although we suggest that its validity might benefit from assuming a more construct-centred approach, and attempting to develop coherent models of competence for individual sectors that would not be based exclusively on lists of tasks and situations specific to individual roles (cf. for instance Lucas, Spencer, & Claxton, 2012) who have attempted to develop models of competence that might be helpful for designing and choosing appropriate teaching methods).[41]

We also argued that treating WBOA as an assessment method that conforms to some basic principles of assessment design would be beneficial as this would emphasise the importance of thinking carefully about the theoretical construct of assessment, the best way of making this construct observable for assessment, and the most reliable way feasible of measuring and quantifying the observations. Discussions with our interviewees have shown that a number of them were (perhaps subconsciously) aware that the construct they were assessing was wider than what the relevant LOs and ACs would suggest, and that they were using the best ways available to them of recognising this in their assessment. We would argue that, although some individuals might be willing and able to do this, not all of them will be as conscientious, and

[41] This is especially relevant in situations, common in VQ assessment, where formative and summative aspects of assessment are closely linked, and sometimes even indistinguishable, and where feedback from assessment is an important source of learning for candidates. In these situations, we would argue with Cronbach that, 'for understanding poor performance, for remedial purposes, for improving teaching methods, and for carving out more functional domains, process constructs are needed' (Cronbach, 1989, p. 22).

so in the interest of assessment validity and fairness, a concerted effort to pin down as far as possible all the relevant construct nuances and the best ways of making them observable is warranted.

In order to make sure that different aspects of the construct of competence are made observable in the most appropriate way, we have also suggested that different versions of WBOA might be designed (in a similar way to medical assessment), each of which would be focusing on different aspects. For instance, one version might be used mainly for assessing procedural skills, accompanied by relevant checklists; another version, with carefully designed rating scales that would do justice to the less tangible aspects of the construct, might be more useful for holistic assessment of a range of less tangible aspects of competence/integration in the workplace/professionalism/team-work and interactions, etc. that might be relevant for a role/sector. Furthermore, some versions might be more useful for formative and some for summative assessment. We recognise that this would not be a simple or straightforward project, but we believe that it would benefit the validity of WBOA and VQ assessment in general.

In relation to this, we have endeavoured to emphasise the importance of understanding, and appreciating, the difference between a *genuine* task and an *authentic* task. We believe that the current orthodoxy in English competence-based qualifications, that competence can only (barring few exceptions) be assessed in the workplace, on real-world tasks, is misguided. It risks assessing candidates on genuine tasks, but which may not always be authentic and may not always be suitable for revealing to the assessors what they really want to know about relevant aspects of candidate's competence. Furthermore, this might (in combination with funding and time pressures) even promote bad practice (both on the part of assessors and of candidates). We believe, and some of our respondents have even suggested this themselves, that a certain amount of assessment in simulated, standardised conditions, designed to make certain important aspects of relevant constructs observable, is preferable and would improve overall validity of qualification results and individual assessment methods. After all, studies in medical education and assessment have found that 'direct [workplace] observation provides necessary and unique information regarding learner skills while maintaining a correlation to overall performance' as well as other measures e.g., multiple choice tests, simulations, etc. (Fromme, Karani, & Downing, 2009). There is no reason why this could not be the case in VQ assessment.

As well as the important issues outlined above, it has been difficult to estimate how much WBOA is carried out in VQs – perhaps unsurprisingly, given its equivocal status in VQ assessment. WBOA is spread across several 'assessment types/methods' in regulatory and awarding organisation databases – and in fact some of those categories are inadequate. For example, a portfolio is a repository of information, rather than a method of assessment (Stone, 2012). It is, of course, difficult to produce a set of descriptive categories that will work without fail across the broad swathe of occupations that are covered by VQs. However, there probably needs to be concerted attempts by interested parties – regulators, AOs, SSCs, etc. – to rationalise

and improve descriptive databases. At the moment they lack coherence and inhibit the extent to which VQs can be described meaningfully. Perhaps, again, a leaf can be taken out of medical assessment's book. In that context, assessment methods, including various methods of workplace observation, seem to be well defined (Baillie & Rhind, 2008). Once they are well defined, it is possible (if necessary) to standardise the methods and easier to gain understanding of their properties. This is important in a context when government reviewers are calling for new approaches to vocational assessment (Richard, 2012).

Different evidence types that might be useful in validation

In this project we conducted short data collection exercises as examples of the type of work that could be carried out in 'live' validation exercises. In the section that follows, we consider the lessons we have learned from doing analysis of different types.

Opinion gathering by questionnaire and interview

Given mainly time and logistic constraints, we did not evaluate formally the content validity of our instruments and the reliability of data sets that we gathered from them, but we consider that we did manage to develop meaningful instruments (questionnaire and interview guide) and gather robust data by their use. We gathered nearly 500 questionnaire responses relatively easily, and they gave – by and large – a consistent and coherent set of findings. In addition, we interviewed 18 assessors, IQAs and QCs about their practice. These respondents were thoughtful and insightful, and helped us to understand the practice of WBOA.

The instruments were designed by following the proposed validation framework and posing questions related to various assumptions and threats to validity that were previously identified as likely. It is our impression, based on (by and large) meaningful data that we managed to gather, that this approach was helpful, and enabled us to address a number of relevant validity issues in a systematic way – which, in turn, confirmed to us the usefulness of the proposed validation framework.

It would also be necessary in a 'live' validation to collect the opinions of other stakeholders with respect to a qualification, assessment method, etc. Those involved in the provision of qualifications will tend, generally, to have a rosier view of them than other stakeholders (He, Boyle, & Opposs, 2011). Moreover, also, when consulting non-expert stakeholders, validators may come up against the challenge of needing to explain technical assessment concepts clearly and comprehensively, yet without unduly guiding respondents (He, Boyle, & Opposs, 2011). We have already referred to the issue of how to interpret minorities of negative responses. If we were to use our instruments in live validation exercises, we would need to pilot them in advance, and remove any non-functioning items. In addition, ideally the sample of respondents would be based on an appropriate sampling frame, which was not used in the current case.

As discussed in more detail in the relevant sections, the questionnaire and interview responses generally confirmed our views as to whether including the various assumptions and threats to validity into the framework was justified, and therefore, most of those that had been there prior to our fieldwork, were subsequently retained in the framework. We would suggest that, for each individual qualification in live validation exercises, similar instruments might be used in the first instance to provide an indication regarding which assumptions outlined in the framework might be in need of more thorough investigation in specific cases, and which might be seen as sufficiently plausible a priori.

Developing indices to provide quantitative evidence of validity/quality of aspects of WBOA

One of our objectives in carrying out this work was to investigate the feasibility of deriving statistics to show the quality of assessments, or elements thereof (analogues to 'questions' or 'items'). We set out the reasons why this might be a good idea, and those why it might not be, but have come to the view that – on balance – it is potentially useful to use statistics to show the measurement properties of WBOA.

We used two data sources for our analysis; first, the data file gathered during the (Harth & Hemker, 2012) reliability research, and secondly, an example file extracted from City and Guilds' own e-portfolio system, Learning Assistant. We were able to derive useful measures from both these sources – even though neither was designed for the purpose to which we put it. We believe that it was sensible to re-use the (Harth & Hemker, 2012) file, given the labour that went into its collection and collation.

Data collection was self-evidently much easier from the e-portfolio. For this bespoke activity, there was still considerable data formatting to be done before analysis could be carried out. But this is a problem that could be resolved if and when novel analyses were integrated into 'business as usual' within awarding organisations.

Perhaps contrary to what is sometimes thought (cf. Stone, 2012), the use of e-portfolio does not – in and of itself – introduce standardisation into assessment practice. Rather, current e-portfolio systems are often set up as 'workflow systems' and assessment validity is not the guiding principle behind their design. What e-portfolios do afford (apart from the relatively pain-free gathering of data for bespoke exercises) is the potential for awarding organisations to standardise assessment practice almost 'covertly' (Boyle, 2009, p. 33). Whilst Harth & Hemker (2012) found a myriad of portfolio forms in use when they visited centres for their research, the e-portfolio can be updated by its owner and centres can be 'led' to provide data to the e-portfolio owner's specification.

We believe that we have provided useful approaches to understanding something akin to item facility, investigated the usefulness of IRT information functions, and tried to show how 'absent' learning outcomes might say something about their relative difficulty. Also, we have used chi-squared as an example of how to

say something about the severity or leniency of raters and have replicated Harth & Hemker's (2012) work in deriving a value of Cronbach's alpha.

Several conclusions flow from our analysis. Firstly, we believe that the data derived from WBOA (and competence-based assessment in general) have very different properties to data that are derived from standardised tests. Variance will tend to be much lower, and distributions are often skewed, rather than symmetrical. Further, the concept of 'fail' or 'zero' is debatable in this context. For clarity, this is not because of some flaw in data collection, these are legitimate features of data from WBOA and/or competence-based assessment. They flow from the purposes and assumptions of that assessment method/tradition. It does not follow that data sets with the features we describe in this paragraph are somehow inferior to those with more well-established characteristics, which are derived from the administration of standardised examinations. What we are observing is difference, rather than one being better or worse than the other.

However, some implications flow from WBOA data being different to standardised exam data. We recommend, having carried out and reflected on this work, that 'plain' statistical analyses are carried out, laden with as few assumptions as possible. We found non-parametric methods to be particularly appropriate. Before starting on this endeavour, we supposed that it might be best to use very specialised, complex inferential methods to analyse our 'different' data (multi-dimensional IRT or multivariate g-theory, for example). However, having conducted this work, our view has swung strongly away from such an approach. Complex inferential methods are dependent on many assumptions and often these simply do not apply to the case of WBOA. It is much better to use techniques that would be found in a general statistics textbook, rather than those that are only found in recent editions of *Psychometrika* or the most recent version of *Educational Measurement*. What we need is pretty transparent analysis that can be understood by as many as possible, and replicated and expanded upon. It may well be that – at some point in the future – we need to bring complex inferential methods back into the picture, but for now it is better to stay with vanilla approaches.

Conclusions

We believe that we have taken a useful first step towards describing and defining WBOA in terms that will be understandable to an audience wider than just VQ practitioners. We also hope that the approach we have taken of emphasising that WBOA is an assessment method that needs to comply with certain general assessment and validity principles has helped to derive a validation framework that is comprehensive and general enough, yet does justice to some of the specific properties of WBOA.

Although validation could be seen as a continuing (never-ending!) process, we believe that we have demonstrated through our necessarily brief fieldwork, that it is possible to follow the proposed framework selectively yet still obtain useful insight into aspects of validity of WBOA. Therefore, the fact that the framework is fairly detailed and likely to be difficult to apply operationally in full should not prevent the use of some of its aspects in validation exercises as well as in everyday awarding organisation practice.

A number of issues that require further research have emerged from this study, the most important of which are outlined in our recommendations overleaf. We hope that the current increased focus on VQs, both from the regulator and more generally, will help to drive these research strands forward to the benefit of the VET sector.

Recommendations

1. Although we believe that the proposed framework is sound and fairly comprehensive, we strongly recommend that it is fully piloted on a representative sample of competence-based VQs.
2. We also recommend that any pilots of the framework are carried out with a view to prioritising the threats which exist to validity, which will help in operational validation.
3. One possible approach to implementing some aspects of the framework into VQ awarding organisation practice might be to use it in structuring and focusing the operational verification processes (which, in any case, could be seen as continuous internal validation exercises) to reflect the framework as far as this is practically feasible.

Our other recommendations, outlined below, reflect our current view of the main validation questions (all based on the proposed framework) that future pilots should seek to answer based on validation exercises in individual qualifications. Some of our high-level recommendations (which probably go beyond WBOA-related issues) would require cross-awarding organisation cooperation and agreement from relevant Sector Skills Councils (SSCs) and funding bodies to investigate and potentially implement, e.g., combined workplace/simulated assessment in competence-based VQs. The regulator might wish to promote and facilitate debate and further research into such issues.

4. Review of constructs:
    - Design exercises to investigate whether current Learning Outcomes and Assessment Criteria assessed by WBOA appropriately reflect what practitioners actually assess and believe are important aspects of competence.
    - Investigate the possibility of expressing the construct as underlying knowledge/skills/abilities rather than lists of tasks where this is possible.
5. Review of available work-based assessment methods using insights from other fields where they are widely used (e.g., medical education):
    - Based on this review, develop criteria to judge whether the current form of WBOA used in VQs is the most appropriate for the constructs (that should be) assessed.
    - Investigate the potential of different 'versions' of WBOA (or different assessment methods altogether) to enable better targeting of assessment (e.g., those focusing on procedural skills vs. those focusing on integration of skills and less easily definable aspects of competence such as reasoning, professionalism, communication, and integration in the workplace).
    - Translate insights from these investigations into coherent database categories in order to ensure that WBOA (or its versions) is recognised and recorded as an assessment method there.
6. Review of 'orthodoxies':

- Notwithstanding the overall support for using different forms of workplace assessment – is workplace assessment always the most appropriate? Could some carefully designed and targeted simulations not contribute useful additional evidence of competence (and mitigate negative washback)? A genuine, real-world task may not always be authentic, i.e. might sometimes not allow the assessors to see the abilities/skills that they are interested in.
- Is observational assessment really unproblematic in competence-based VQs – or, should we be taking a hard look at levels of standardisation and agreement between assessors/IQAs/QCs and designing methods to promote and monitor this more effectively?
- Does the final binary judgement of whether someone is competent or not yet competent really preclude the use of meaningful global scales (or comparative judgement techniques) to help standardise or help investigate the consistency of assessor judgements?
- It is possible (though not straightforward) to produce appropriate and useful statistical indicators of WBOA quality – relevant research in this area should be continued. This might become easier if standardised e-portfolio forms were used across centres, and if the forms were designed in a way that might support relevant data collection and analysis.
- Review outcomes-based funding arrangements insofar as these present threats to the integrity of WBOA practices and decisions (cf. recommendations by the Richard Review).

References

AERA, APA, & NCME. (1999). *Standards for educational an psychological testing.* Washington, DC: AERA.

AlphaPlus Consultancy Ltd. (forthcoming). *Validation of Regulated Assessments.* Coventry: The Office of Qualifications and Examinations Regulation.

Ashworth, P. D., & Saxton, J. (1990). On 'competence'. *Journal of Further and Higher Education, 14*(2), 3-25.

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Baillie, S., & Rhind, S. (2008). *A Guide to Assessment Methods in Veterinary Medicine.* Edinburgh: Royal College of Veterinary Surgeons Trust. Retrieved December 20, 2012, from http://www.live.ac.uk/documents/assessment_guide.pdf

Baird, J.-A., Beguin, A., Black, P., Pollitt, A., & Stanley, G. (2012). The Reliability Programme: Final Report of the Technical Advisory Group. In D. Opposs, & Q. He (Eds.), *Ofqual's reliability compendium* (pp. 771 - 838). Coventry, United Kingdom: The Office of Qualifications and Examinations Regulation.

Bartram, D., & Mitchell, L. (1992). *The design, development and delivery of National Vocational Qualifications and Scottish Vocational Qualifications: the place of knowledge and understanding.* Unpublished guidance note.

Benett, Y. (1993). The Validity and Reliability of Assessments and Self-assessments of Work-based Learning. *Assessment & Evaluation in Higher Education, 18*(2), 83-94.

Bieri, C., & Schuler, P. (2011). Cross-curricular competencies of student teachers: a selection model based on assessment centre admission tests and study success after the first year of teacher training. *Assessment & Evaluation in Higher Education, 36*(4), 399-415.

Borsboom, D. (2005). *Measuring the mind.* Cambridge: Cambridge University Press.

Borsboom, D., Cramer, A. O., Kievit, R. A., Zand Scholten, A., & Franic, S. (2009). The end of construct validity. Revisions, new directions, and applications. In R. W. Lissitz (Ed.), *The concept of validity* (pp. 135-170). Charlotte, NC: Information Age Publishing, Inc.

Boyle, A. (2009). The Contribution of Centralisation and De-Centralisation to Public Confidence in Examinations and Qualifications. *CADMO, 17*(2), 19-38.

Boyle, A., & Rahman, Z. (2012). *The internal reliability of some City & Guilds tests.* Coventry: The Office of Qualifications and Examinations Regulation.

Bramley, T., & Dhawan, V. (2012). Estimates of reliability of qualifications. In Q. He, & D. Opposs (Eds.), *Ofqual's reliability compendium* (pp. 217 - 321). Coventry: The Office of Qualifications and Examinations Regulation.

Braun, E. M., Hammad, S., & Hannover, B. (2011). Self-rated competences and future vocational success: a longitudinal study. *Assessment & Evaluation in Higher Education, 36*(4), 417-427.

Brennan, R. L., & Johnson, E. G. (1995). Generalisability of performance assessments. *Educational Measurement: Issues and Practice, 14*(4), 25-27.

Brockmann, M., Clarke, L., & Winch, C. (2011). *Knowledge, skills and competence in the European labour market.* London: Routledge.

Carroll, G., & Boutall, T. (2010). *Guide to Developing National Occupational Standards.* Retrieved April 11, 2012, from http://www.ukces.org.uk/assets/ukces/docs/publications/evidence-report-44-developing-occupational-skills-profiles-for-the-uk-a-feasibility-study.pdf

Cedefop. (2008). *Terminology of European education and training policy: A selection of 100 key terms.* Thessaloniki: Cedefop.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2008). *Building a validity argument for the Test of English as a Foreign Language.* London: Routledge.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice, 29*(1), 3-13.

City & Guilds. (1994). *Producing a portfolio of evidence for an NVQ.* London: City & Guilds of London Institute.

City & Guilds. (1997). *NVQ/SVQ Handbook.* London: City & Guilds of London Institute.

City & Guilds. (2003). *Principles for the development and delivery of assessments.* London: City & Guilds of London Institute.

City & Guilds. (2006a). *N/SVQ guide for centres and candidates: centre guide.* London: City & Guilds of London Institute. Retrieved May 1, 2012, from http://www.cityandguilds.com/documents/ind_sport/EN-12-0001.pdf

City & Guilds. (2006b). *Guidance on internal verification of N/SVQs.* London: City & Guilds of London Institute.

City & Guilds. (2011a). Splitting skills and know how - position statement. London: City & Guilds of London Institute.

City & Guilds. (2011b). *Levels 1-3 NVQ Qualifications in Hairdressing, Barbering and Combined Hair Types (3008): Assessors guide.* London: City & Guilds of London Institute.

City & Guilds. (2011c). *Guidance for centres: supporting customer excellence centre manual.* London: City & Guilds of London Institute.

Colley, H., & Jarvis, J. (2007). Formality and informality in the summative assessment of motor vehicle apprentices: a case study. *Assessment in Education: Principles, Policy & Practice, 14*(3), 295-314.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement, 2nd ed.* (pp. 443–507). Washington, DC: American Council on Education.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer, & H. Braun (Eds.), *Test validity* (pp. 3-17). Hillsdale, NJ: Erlbaum.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. E. Linn, *Intelligence: Measurement, theory, and public policy* (pp. 147-171). Urbana: University of Illinois Press.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302.

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), 265-286.

Crossley, J., & Jolly, B. (2012). Making sense of work-based assessment: ask the right questions, in the right way, about the right things, of the right people. *Medical Education, 46*, 28–37.

Eisner, E. (1991). *The enlightened eye: Qualitative inquiry and the enhancement of educational practice.* New York: Macmillan.

Epstein, R. M., & Hundert, E. M. (2002). Defining and Assessing Professional Competence. *Journal of the American Medical Association, 287*(2), 226-235.

Eraut, M. (1994). *Developing Professional Knowledge and Competence.* London: Falmer Press.

Eraut, M., Steadman, S., Trill, J., & Parkes, J. (1996). *The assessment of NVQs.* Brighton: University of Sussex Institute of Education.

European Commission. (2007). *Key Competencies for Lifelong Learning: European Reference Framework.* Luxembourg: Office for Official Publications of the European Communities, 2008. Retrieved March 27, 2012, from http://ec.europa.eu/dgs/education_culture/publ/pdf/ll-learning/keycomp_en.pdf

European Commission. (2008). *The European Qualifications Framework for Lifelong Learning (EQF).*

Fischer, K. W., Bullock, D. H., Rotenberg, E. J., & Raya, P. (1993). The dynamics of competence: how context contributes directly to skill. In R. H. Wozniak, K. W. Fischer, & (Eds), *Development in Context: Acting and Thinking in Specific Environments* (pp. 93-117). Hillsdale, NJ: Erlbaum.

Fitzpatrick, R., & Morrison, E. J. (1971). Performance and Product Evaluation. In R. L. Thorndike (Ed.), *Educational Measurement* (pp. 237-270). Washington D.C.: American Council of Education.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 8*(9), 27-32.

Fromme, H. B., Karani, R., & Downing, S. M. (2009). Direct Observation in Medical Education: A Review of the Literature and Evidence for Validity. *Mount Sinai Journal of Medicine*, 365-371.

Further Education Unit. (1984). *Towards a Competence-Based System: An FEU View.* London: Further Education Unit.

Gijbels, D. (2011). Assessment of vocational competence in higher education: reflections and prospects. *Assessment & Evaluation in Higher Education, 36*(4), 381-383.

Gordon, J., Halasz, G., Krawczyk, M., Leney, T., Michel, A., Pepper, D., . . . Wiśniewski, J. (2009). *Key Competences in Europe: Opening Doors For Lifelong Learners Across the School Curriculum and Teacher Education.* Retrieved March 12, 2012, from http://ec.europa.eu/education/more-information/doc/keyreport_en.pdf

Greatorex, J. (2000). What research could an Awarding Body carry out about NVQs? *Paper presented at the British Educational Research Association Conference.* University of Cardiff, UK. Retrieved March 29, 2012, from http://www.cambridgeassessment.org.uk/ca/digitalAssets/113935_What_Research_Could_an_Awarding_Body_Carry_Out_about_NVQs.pdf

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

Haertel, E. H. (1990). From expert opinions to reliable scores: Psychometric for judgment-based teacher assessment. *Paper presented at the Annual Meeting of the American Educational Research Association.* Boston, USA.

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

Hagar, P., Gonczi, A., & Athanasou, J. (1994). General Issues about Assessment of Competence. *Assessment & Evaluation in Higher Education, 19*(1), 3-16.

Halász, G., & Michel, A. (2011). Key Competences in Europe: interpretation, policy formulation and implementation. *European Journal of Education, 46*(3), 289-306.

Harth, H., & Hemker, B. (2012). On the reliability of results in vocational assessment: the case of work-based certifications. In D. Opposs, & Q. He (Eds.), *Ofqual's reliability compendium* (pp. 321-364). Coventry, United Kingdom: The Office of Qualifications and Examinations Regulation.

Hauer, K. E., Holmboe, E. S., & Kogan, J. R. (2011). Twelve tips for implementing tools for direct observation of medical trainees' clinical skills during patient encounters. *Medical Teacher*, 27-33.

He, Q. (2012). Estimating the Reliability of Composite Scores. In D. Opposs, & Q. He (Eds.), *Ofqual's reliability compendium* (pp. 523 - 555). Coventry: The Office of Qualifications and Examinations Regulation.

He, Q., Boyle, A., & Opposs, D. (2011). Public perceptions of reliability in examination results in England. *Evaluation & Research in Education, 24*(4), 255-283.

Herling, R. W. (2000). Operational Definitions of Expertise and Competence. *Advances in Developing Human Resources, 2*(1), 8-21.

Hodkinson, P. (1992). Alternative Models of Competence in Vocational Education and Training. *Journal of Further and Higher Education, 16*(2), 30-39.

Hoskins, B., & Fredriksson, U. (2008). *Learning to Learn: What is it and can it be measured?* European Commission Joint Research Centre. Retrieved December 20, 2012, from http://publications.jrc.ec.europa.eu/repository/bitstream/111111111/979/1/learning%20to%20learn%20what%20is%20it%20and%20can%20it%20be%20measured%20final.pdf

Hyland, T. (1993). Competence, Knowledge and Education. *Journal of Philosophy of Education, 27*(1).

IFF Research and the Institute of Employment Research (IER). (2012). *Evaluation of Apprenticeships: employers (BIS research paper number 77).* London: Department for Business, Innovation and Skills.

Jessup, G. (1991). *Outcomes: NVQs and the Emerging Model of Education and Training.* London: Falmer Press.

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

Johnson, S., Johnson, R., Miller, L., & Boyle, A. (2012). *Reliability of Vocational Assessment: an evaluation of level 3 electro-technical qualifications.* Coventry: Office of Qualifications and Examinations Regulation.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527-535.

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

Kane, M. T. (2004). Certification testing as an illustration of argument-based validation. *Measurement, 2*(3), 135–170.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (pp. 17-64). Westport, CT: ACE and Praeger Publishers.

Kane, M. T., Crooks, T., & Cohen, A. (1999). Validating Measures of Performance. *Educational Measurement: Issues and Practice, 18*(2), 5-17.

Konrad, J. (1999). Assessment and Verification of National Vocational Qualifications. *Paper presented to International Lifelong Learning Conference .* University College Worcester, UK. Retrieved April 27, 2012, from http://www.leeds.ac.uk/educol/documents/000001074.htm

Kopelow, M., Schnabi, G., Hassard, T., Klass, D., Beazley, G., Hechter, F., & Grott, M. (1992). Assessing practising physicians in two settings using standardised patients. *Academic Medicine, 67*, 19-21.

Lane, S., & Stone, C. A. (2006). Performance assessment. In R. L. Brennan (Ed.), *Educational measurement* (pp. 387-431). Westport, CT: ACE and Praeger Publishers.

Le Deist, F. D., & Winterton, J. (2005). What is Competence? *Human Resource Development International, 8*(1), 27-46.

Linacre, J. M. (2009). *A User's Guide to WINSTEPS® & MINISTEP. Rasch-Model computer programs. Program Manual 3.68.0.* Chicago: Winsteps. Retrieved July 9, 2012, from http://www.winsteps.com/a/winsteps3682.pdf

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 5-21.

Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher, 36*, 437-448.

Lucas, B., Spencer, E., & Claxton, G. (2012). *How to teach vocational education: A theory of vocational pedagogy.* London: The City and Guilds of London Instite.

Mansell, W. (2011, June). *Is 'rigorous' school accountability supported by OECD data?* Retrieved December 20, 2012, from Warwick Mansell NAHT blog: http://www.naht.org.uk/welcome/news-and-media/blogs/warwick-mansell/?blogpost=464

McNamara, T. (1996). *Measuring second language performance.* London: Longman.

McNamara, T. (2006). Validity in Language Testing: The Challenge of Sam Messick's Legacy. *Language Assessment Quarterly, 3*(1), 31-51.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement (3rd ed.)* (pp. 13-103). New York: Macmillan.

Messick, S. (1994). The Interplay of Evidence and Consequence in the Validation of Performance Assessment. *Educational Researcher, 23*(2), 13-23.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62*, 229-258.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher, 23*(2), 5-12.

Murphy, R., Burke, P., Content, S., Frearson, M., Gillespie, J., Hadfield, M., . . . Wilmut, J. (1995). *The reliability of assessment of NVQs.* University of Nottingham, School of Education. Nottingham: University of Nottingham. Retrieved May 14, 2012, from http://www.nottingham.ac.uk/shared/shared_cdell/pdf-reports/nvqrelrep.pdf

National Council for Vocational Qualifications. (1995). *NVQ criteria and guidance.* London: National Council for Vocational Qualifications.

Noel, G. L., Herbers, J. E., Caplow, M. P., Cooper, G. S., Pangaro, L. N., & Harvey, J. (1992). How well do Internal Medicine faculty members evaluate the clinical skills of residents? . *Annals of Internal Medicine*, 757–765.

Norcini, J. J. (2005). Current perspectives in assessment: the assessment of performance at work. *Medical Education, 39*(9), 880–889.

Norcini, J. J., & Burch, V. (2007). Workplace-based assessment as an educational tool. *Medical Teacher*, 855-871.

Norris, N. (1991). The Trouble with Competence. *Cambridge Journal of Education, 21*(3), 331-341.

Oates, T. (2013). *Towards a new VET. Effective vocational education and training.* Cambridge: Cambridge Assessment.

OECD. (2005). *The Definition and Selection of Key Competencies: Executive Summary.* Retrieved from http://www.oecd.org/dataoecd/47/61/35070367.pdf

Office of Qualifications and Examinations Regulation (Ofqual). (2011). *Invitation to tender for the investigation of the validity of regulated assessment OF154.*

Office of Rail Regulation. (2007). *Developing and Maintaining Staff Competence.* (Second ed.). London: Office of Rail Regulation. Retrieved December 20, 2012, from http://www.rail-reg.gov.uk/upload/pdf/sf-dev-staff.pdf

Opposs, D., & He, Q. (2011). *The Ofqual assessment validity project.* Coventry: The Office of Qualifications and Examinations Regulation.

Pollitt, A. (2004). Let's stop marking exams. *Paper presented at the 30th Annual Conference of the International Association for Educational Assessment.* Philadelphia, USA. Retrieved July 9, 2012, from http://www.cambridgeassessment.org.uk/ca/digitalAssets/113942_Let_s_Stop_Marking_Exams.pdf

Pollitt, A. (2009). Abolishing marksism and rescuing validity. *Paper presented at the 35th Annual Conference of the International Association for Educational Assessment.* Brisbane, Australia.

Pollitt, A. (2012). The method of Adaptive Comparative Judgement. *Assessment in Education: Principles, Policy & Practice, 19*(3), 281-300. doi:10.1080/0969594X.2012.665354

Pollitt, A., & Ahmed, A. (1999). A New Model of the Question Answering Process. *A paper presented at the 25th Annual Conference of the International Association for Educational Assessment.* Bled, Slovenia.

Pollitt, A., & Ahmed, A. (2007). Improving the quality of contextualized questions: an experimental investigation of focus. *Assessment in Education: Principles, Policy & Practice, 14*(2), 201-232.

Qualifications and Curriculum Development Agency (QCDA). (2010). *Gidelines for writing credit-based units of assessment, Version 4, QCDA/10/4725.*

Redfern, S., Norman, I., Calman, L., Watson, R., & Murrells, T. (2002). Assessing competence to practise in nursing: a review of the literature. *Research Papers in Education, 17*(1), 51-77.

Rethans, J., Sturmans, F., Drop, R., van der Vleuten, C., & Hobus, P. (1991). Does competence of general practitioners predict their performance? Comparison between examination setting and actual practice. *British Medical Journal, 303*, 1377-1380.

Richard, D. (2012). *The Richard Review of Apprenticeships.* London: Department for Business, Innovation & Skills. Retrieved December 20, 2012, from http://www.schoolforstartups.co.uk/richard-review/richard-review-full.pdf

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*, 119-144.

Schuwirth, L. W., Southgate, L., Page, G. G., Paget, N. S., Lescop, J. M., Lew, S. R., . . . Baron-Maldonado, M. (2002). When enough is enough: a conceptual basis for fair and defensible practice performance assessment. *Medical Education, 36*, 925–930.

Semta. (2009). *Common Requirements for National Vocational Qualifications (NVQ) in the QCF.* Retrieved April 4, 2012, from http://semta.org.uk/store/files/Common_requirements_NVQs_in_the_QCF1.1_12_06_09.pdf

Shaw, S., Crisp, V., & Johnson, N. (2012). A framework for evidencing assessment validity in large-scale, high-stakes international examinations. *Assessment in Education: Principles, Policy & Practice, 19*(2), 159-176.

Stasz, C. (2011). *The purposes and validity of vocational qualifications.* Cardiff: SKOPE.

Stone, A. (2012). What does 'e-portfolio' mean in the vocational sector? *International Journal of e-assessment, 2*(2). Retrieved December 3, 2012, from http://journals.sfu.ca/ijea/index.php/journal/article/viewFile/40/42

Swanson, D. B., Norman, G. R., & Linn, R. L. (1995). Performance-Based Assessment: Lessons From the Health Profession. *Educational Researcher, 24*, 5-35.

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

Tchibozo, G. (2011). Emergence and outlook of competence-based education in European education systems: an overview. *4*(3), 193-205.

Torrance, H., Colley, H., Garratt, D., Jarvis, J., Piper, H., Ecclestone, K., & 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.

Toulmin, S. (1958). *The uses of argument.* Cambridge: Cambridge University Press.

UK Commission for Employment and Skills (UKCES). (2011). *NOS Strategy 2010-2020 (revised 2011).* London: UK Commission for Employment and Skills (UKCES).

Unwin, L., Fuller, A., Turbin, J., & Young, M. (2004). *What determines the impact of vocational qualifications? A literature review.* London: Department for Education & Skills. Retrieved December 3, 2012, from https://www.education.gov.uk/publications/eOrderingDownload/RR522.pdf

Walker, M. (2010). Key note speech to the CIEA National Conference. *Chartered Institute of Educational Assessors (CIEA) National Conference.* Retrieved May 1, 2012, from http://www.ciea.org.uk/Home/news_and_events/events_listing/CIEA_national_conference_2010_review/Keynote1_1_MickWalker.aspx

Weigel, T., Mulder, M., & Collins, K. (2007). The concept of competence in the development of vocational education and training in selected EU member states. *Journal of Vocational Education & Training, 59*(1), 53-66.

Weinert, F. E. (2001). Concept of competence: A conceptual clarification. In D. Rychen, & L. Salganik (Eds.), *Defining and selecting key competences* (pp. 17-31). Goettingen: Hogrefe & Huber.

Weiss, C. H. (1972). *Evaluation research: Methods for assessing program effectiveness.* Englewood Cliffs, NJ: Prentice-Hall.

Wenger, E., McDermott, R., & Snyder, W. M. (2002). *Cultivating Communities of Practice.* Boston, MA: Harvard Business Press.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappa, 70*, 703-713.

Wiggins, G. (1993). Assessment: Authenticity, Context, and Validity. *Phi Delta Kappa*, 200-208, 210-214.

Wolf, A. (1995). *Competence-Based Assessment.* Buckingham, Philadelphia: Open University Press.

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

Wolf, A. (2001). Competence-Based Assessment. In J. Raven, & J. Stephenson (Eds.), *Competence in the learning society.* New York: Peter Lang.

Wolf, A. (2011). *Review of vocational education – the Wolf report.* London: Department for Education. Retrieved April 17, 2012, from https://www.education.gov.uk/publications/eOrderingDownload/The%20Wolf%20Report.pdf

Wools, S., Eggen, T., & Sanders, P. (2010). Evaluation of validity and validation by means of the argument-based approach. *CADMO, 18*(1), 63-82.

Appendix 1: Various assessment methods and other sources of information that might contribute to a portfolio of evidence of a candidate's competence

- **Questioning** tends to be used to assess knowledge and understanding underpinning the practical task in order to review or supplement the performance, filling gaps between what was observed and the evidence required (City & Guilds, 2006a). This can include oral questions, essay questions, structured-answer questions, short-answer questions or multiple choice questions. Four broad aspects of knowledge tend to be used for questioning:
  - knowing that - knowledge of pertinent facts (what, when, where, who, etc)
  - knowing how - knowledge of procedures (also 'what if')
  - knowing when - knowledge facts and procedures to be applied to particular problems or situations
  - knowing why - reasons for procedures or results (City & Guilds, 2003).

- **Assignments** are structured activities that are usually externally set, internally marked and externally verified but in some instances can be developed by centres as long as they are externally verified. They can be used to assess knowledge, practical, oral or interpersonal skills over a period of time and are not likely to be continuously observed (City & Guilds, 2003).

- **Projects** could involve designing, planning or doing by individuals or groups. Compared to assignments, projects tend to be: spread over a longer period of time, less structured, less likely to have one correct answer, allow greater discretion to candidates, and can be undertaken in the workplace (City and Guilds, 2003). Projects may cover aspects of work outside the candidate's responsibility, for example, assessing health and safety in the workplace (City & Guilds, 2011c).

- **Professional discussion** may be used to encourage a candidate to explain how they carry out their work, the evidence produced and the standards met and explain certain behaviours relating to their work to the assessor (City & Guilds, 2006a).

- **Historical evidence** usually relates to performance and tends to be considered at an early stage of the assessment process. This may be an important source, which needs to be authenticated and considered by assessors, possibly leading to further assessment i.e. to confirm competence is current (City & Guilds, 1997).

- **Witness** testimony may be used as evidence where it is not possible to observe candidate performance e.g., presence intrusive to work or emergency at work outside planned assessment (City & Guilds, 2006a). Colleagues, customers and managers may witness relevant activities before or during the evidence gathering process and provide testimonies or endorsements. This may give an indication of candidates' consistency of performance (City & Guilds, 1994). There are two types of witness:

- **An expert witness** is someone who regularly and systematically observes candidates performing tasks e.g., line manager. They will be clear about the purpose and use of testimony and briefed by the internal verifier or assessor to ensure they understand the standards to which the evidence relates (City & Guilds, 2006a). In some cases, the expert witness may not be familiar with the standards so may be less able to make an accurate judgement about the candidate's competence (City & Guilds, 2011c).
- **A non-expert witness** is someone not occupationally competent e.g., customers or peers. Their evidence may be less reliable than that of the expert witness as they are unlikely to be familiar with the standards being assessed (City & Guilds, 2006a).

Appendix 2: Table of common terms (characteristics) within a range of definitions of competence identified from the literature

Skills	Knowledge	Context	Attitudes	Performance/ activity/actions	Ability/ abilities	Values	Attributes	Responsibility	Other characteristic	Source
X	X		X	X[42]					Experience	Further Education Unit (1984, p. 3)
X	X	X		X					Understanding	Jessup (1991, p. 26)
X				X[43]	X[44]				Personal effectiveness Understanding	Investor in People (1995. p. 41), cited in Le Deist & Winterton (2005, p. 34)
X	X	X			X[45]	X	X[46]			Cedefop (2008, p. 47)
X	X	X	X							European Commission (2007, p. 3)
X	X	X			X[47]			X	Autonomy	European Commission (2008, p. 11)
X	X		X			X				Hoskins & Fredriksson (2008, p. 11)
X	X	X	X		X[48]					OECD (2005, p. 4)
X	X		X						Behaviour	(France) Le Deist & Winterton (2005, p. 37), Gordon, et al. (2009, p. 36)
		X						X	Autonomy	(Germany) Gordon, et al. (2009, p. 36), Weigel, Mulder, & Collins (2007, p. 57)
X	X	X				X			Communication Use of Emotions Moral development Reasoning Reflection	Epstein & Hundert (2002)

[42] Successful performance in life roles
[43] To standards expected in employment
[44] To perform activities
[45] To apply learning outcomes
[46] Interpersonal attributes
[47] Personal, social, and methodological abilities
[48] To meet complex demands

Skills	Knowledge	Context	Attitudes	Performance/ activity/actions	Ability/ abilities	Values	Attributes	Responsibility	Other characteristic	Source
		X		X						Redfern, Norman, Calman, Watson, & Murrells (2002, pp. 53-54)
X	X[49]	X		X				X	Willingness to undertake activities to agreed standards	Office of Rail Regulation (2007, p. 2)
X	X	X	X	X	X		X			Ashworth & Saxton (1990, p. 22)
X	X		X	X						Gonzi et al. (1993. pp 5-6), cited Eraut (1994, p. 179)
X	X		X	X	X		X		Capacities Competencies Qualities	Halász & Michel (2011, p. 292)
		X		X[50]						Herling (2000, p. 20)
	X[51]	X[52]			X	X				Tchibozo (2011, p. 194)

[49] and experience
[50] Minimally efficient actions
[51] Cognitive
[52] Set of problem situations

Appendix 3: Further examples of LOs and ACs

Table 25: Assessment criteria for LO 2 'Creatively restyle women's hair' in Unit GH16 'Creatively cut hair using a combination of techniques' (L3 NVQ Diploma in Hairdressing, 3008)

a) Explore the variety of looks with your client using relevant visual aids
b) Recommend a look that is suitable for your client
c) Base your recommendations on an accurate evaluation of your client's hair and its potential to achieve the look
d) Suitably prepare your client's **hair** prior to cutting
e) Confirm with your client the **look** agreed at consultation before commencing the cut
f) Establish and follow suitable cutting guideline(s) to achieve the required look
g) Personalise your **cutting techniques and effects** to take account of **factors** that will influence the desired **look**
h) Combine and adapt your **cutting techniques and effects** in an innovative way to achieve the desired **look**
i) Change your own position and that of your client to help you ensure the accuracy of the cut
j) Establish accurate distribution of weight, balance and shape by cross-checking the cut
k) Create outline shapes that are accurate, defined and achieve the look required by your client
l) Remove any unwanted hair outside the desired outline shape
m) Consult with your client during the cutting service to confirm the desired look
n) Take suitable remedial action to resolve any problems arising during the cutting service
o) Make a final visual check to ensure the finished cut is accurate
p) Use **creative finishing techniques** that complement the cut
q) Ensure the finished restyled look complements your client's features and enhances their personal image and that of the salon
r) Confirm your client's satisfaction with the finished **look**

Table 26: Assessment criteria for LO 2 'Be able to inspect electrotechnical systems and equipment' in Unit 317 'Inspecting, testing, commissioning and certifying electrotechnical systems and equipment in buildings, structures and the environment (ELTP06)' (Level 3 NVQ Diplomas Electrotechnical Technology - 2357)

The learner can:

a) assess whether the safe system of work is appropriate to the work activity
b) carry out a visual inspection in accordance with the requirements of the installation specification, the IE Wiring Regulations and IEE Guidance Note 3, that includes:
   - the installation methods of wiring systems and equipment
   - the selection of conductors, cables and cords
   - the selection of protective and isolation devices
   - routing and identification/labelling of conductors, cables and flexible cords
   - presence of means of earthing
   - presence of protective conductors and bonding
   - isolation
   - type and rating of over current protective devices
c) complete a schedule of inspections in accordance with the IEE Wiring Regulations and IEE Guidance Note 3.

Table 27: Assessment criteria for LO 5 'Be able to support individuals and others following an incident of challenging behaviour' in Unit 115 'Promote Positive Behaviour' (Level 3 Diploma for the Children & Young People's Workforce 4227-03/04/05)

The learner can:

a) Demonstrate methods to support an individual to return to a calm state following an incident of challenging behaviour
b) Describe how an individual can be supported to reflect on an incident including:
   - how they were feeling at the time prior to and directly before the incident
   - their behaviour
   - the consequence of their behaviour
   - how they were feeling after the incident
c) Describe the complex feelings that may be experienced by others involved or witnessing an incident of challenging behaviour
d) Demonstrate how to debrief others involved in an incident of challenging behaviour
e) Describe the steps that should be taken to check for injuries following an incident of challenging behaviour.

Table 28: Assessment criteria for LO 4 'Be able to respond appropriately to incidents of challenging behaviour' in Unit 115 'Promote Positive Behaviour' (Level 3 Diploma for the Children & Young People's Workforce 4227-03/04/05)

The learner can:

a) Identify types of challenging behaviours
b) Demonstrate how to respond to incidents of challenging behaviour following behaviour support plans, agreed ways of working or organisational guidelines
c) Explain the steps that are taken to maintain the dignity of and respect for an individual when responding to an incident of challenging behaviour
d) Demonstrate how to complete records accurately and objectively in line with work setting requirements following an incident of challenging behaviour.

Table 29: Assessment criteria for LO 3 'Be able to co-ordinate liaison with other relevant persons during work activities' in Unit 313 'Overseeing and organising the work environment (electrical installation) (ELTP03)' (Level 3 NVQ Diplomas Electrotechnical Technology - 2357)

The learner can:

a) comply with approved procedures to ensure effective co-ordination with other workers/contractors, including steps to resolve issues which are outside the scope of their job role
b) apply communication techniques that are clear, accurate and appropriate to the situation

Appendix 4: Questionnaire questions, framework mapping and main findings[53]

Table 30: Questionnaire questions, framework mapping and main findings

Q. No	Question	Framework inference	Summary of findings
6	Do any of the qualifications that you work with require observation of tasks that are RARELY carried out in the workplace?	DOMAIN DESCRIPTION	Gives reasonable support to validity of WBOA. (331 out of 496 say they have NOT seen rare tasks in assessment.) No sig dif between role types.
7	Do the qualifications that you work with assess ALL the tasks required for the job?	DOMAIN DESCRIPTION/ EXTRAPOLATION	Gives reasonable support to validity of WBOA. (374 out of 496 say that quals assess ALL tasks needed for job.) No sig dif between role types.
8	In relation to inappropriate or missing task(s), what action(s) have you taken?	DOMAIN DESCRIPTION/ EXTRAPOLATION	Large number of blank responses (290 out of 496) Biggest response groups have informed IQA or QC. Few have contacted AO or SSC directly.
9	How much do you agree or disagree with the following statements about the feedback process when you feel tasks are inappropriate or missing?	DOMAIN DESCRIPTION/ EXTRAPOLATION	
	a. I am clear on what to do		Large number of blank responses (291 out of 496) Respondents largely knew what to do (140 out of 205) tend to/strongly agree. No sig dif between role types.
	b. I know that the relevant people have heard me		Large number of blank responses (291 out of 496) Agreement rates fairly high, but slightly lower than sub-item 9a (117 out of 205) Chi squared significant at 5 % level.
	c. I know that the relevant people take necessary action		Large number of blank responses (291 out of 496) Agreement rates fairly high, but lowest of three sub-items (106 out of 205) No sig dif between role types.
11	Which of the following do you most commonly do when forming your judgement about candidates' performance in work-based observations?	JUDGEMENT	Assessors reported referring to series of combinations when assessing.

[53] To make a manageable table, we have included quantitative items only.

Q. No	Question	Framework inference	Summary of findings
			Very few 'confessed' to forming an overall judgement without referring to the assessment criteria.
12	How often would you say that your decision based on work-based observation differs from that of the assessor?	JUDGEMENT	Majority of IQAs and QCs (255 out of 421) reported differing from assessors 'seldom' or 'never'.
14	In general, do you think that the number of observations required or suggested by the relevant assessment strategy is enough to make a confident judgement?	GENERALISATION/ JUDGEMENT	Large majority thought that number of obs was about right (339 out of 496). No sig dif between role types.
16	How important are each of these standards to you when observing a task? Please rank from 1 to 4, with 1 being the most important.	JUDGEMENT	Ranking of stds types broadly supported validity argument – quals stds 'lowest' ranking most important, own stds 'least important' But – still important minorities, e.g., 101 rated 'own stds' as 1 or 2.
17	In work-based observation, how often do you compare a candidate's performance with the following to inform your judgement?	JUDGEMENT	Scale responses to numerical values, and means calculated as follows …
	a. Candidate's previous formally assessed performance		Respondents took into consideration 'often'
	b. Candidate's previous performance (not formally assessed)		'Halfway between' 'often' and 'sometimes'.
	c. Performance of other candidates		'slightly closer to' 'seldom' than 'sometimes'.
18	How lenient or strict do you think you are when making judgements based on observation in the workplace?	JUDGEMENT	Assessors perceived themselves to be fairly strict – mean 8.02 on scale of 1 to 10.
19	In your experience of work-based observational assessment, how often on a task...	JUDGEMENT	Chi squared significant at 1% level. Substantial support for WBOA method. But important minorities who had seen mis-classifications sometimes.
	a. … do candidates PASS but SHOULD NOT		
	b. … candidates DO NOT PASS but SHOULD have		
21	How often do you feel assessors lead/coach candidates during observation in the workplace?	ADMINISTRATION	Strong support for WBOA – 288 responses say 'seldom' or 'never'. Sizeable minority say 'sometimes' (140). No sig dif between IQAs and QCs.
23	In your experience, how consistent or inconsistent are workplace observation standards of assessors for the same qualifications across the centres that you quality assure?	JUDGEMENT	Mean value = 1.93. (2 = somewhat consistent.)
24	How likely or unlikely is it that you will spot problematic assessors (e.g., inconsistent, too severe, too lenient) using the current sampling approach?	JUDGEMENT	Strong support for ability to spot problematic assessors (291 out of 421). Chi squared significant at 1% level.

Q. No	Question	Framework inference	Summary of findings
25	How often do you take part in standardisation activities relevant to observation in the workplace?	JUDGEMENT	Fairly frequent participation. 323 out of 410 respondents attended standardisation at least very few months. Chi squared significant at 1% level; differences tended to be between QCs and the rest – perhaps reflecting different roles.
27	How effective or ineffective do you find your standardisation activities in aligning your own approach and standards with those of other people in your role?	JUDGEMENT	No sig dif between role types. Large majorities (370 out of 386 definite responses) endorsed effectiveness of standardisation activities.

Appendix 5: QC interview schedule

Observation

1. What are the main benefits/advantages/good things about using observation in the workplace to assess and make judgements?
2. What are the main difficulties in using observation in the workplace to assess and make judgements?
    - Would you still support this (instead of, e.g., RWE) as the main assessment method for competence-based qualifications?
3. What are the main issues you find with assessors using observational assessment in the workplace?
    - Do you find that they sometimes over or under assess?
    - Are you happy with how this is monitored and tracked through QA procedures?
    - Can the monitoring (QA) system be improved?
4. How likely it is that you will spot problematic assessors (e.g., inconsistent, too severe, too lenient assessors, discriminatory, personality clash problems) using your current sampling approach?
    - If unlikely – why?
    - Do you rely on your own observation of assessors or do you discuss with IQAs or interviews with candidates?
    - In which situations can the QCs/IQAs intervene/get involved in observation?
5. How likely it is that you will spot inappropriate tasks (e.g., not authentic, too easy/too demanding, inappropriate simulation) using your current sampling approach?
    - If there are such tasks (e.g., rarely occurring tasks that need to be assessed by observation), how do people deal with it (assessors, candidates, salons)?

Competence standards

1. What are L2/L3 candidates ready for when they qualify?
    - Are they job ready?
    - Commercial timings?
    - Level of supervision?

- How difficult might it be moving between workplaces if you have only trained in one?

2. Have you ever found or heard that qualified learners are unable to do certain tasks on the job?

Context, task choice

1. How much are you able to feedback on task inappropriateness or missing tasks (if any) through the QC procedures?
    - Do you feel this has an effect?
    - How much do you feel the SSC is engaged and actively looking for feedback or taking feedback on board? Or is it mainly the Awarding Organisation?
2. Difficult customers/situations/different environments – how to deal with in assessment; dealing with extremes (representative of the rest or not?)
    - Have you experienced an examiner stopping an assessment in such a situation? Did you agree?
    - Do assessors get any advice on this in their training or standardisation? Is this something that is discussed and agreed upon among assessors/IQAs?
3. Do you feel that you always have enough evidence to make a secure judgement? How confident do you generally feel about making the right judgement?

Relationship with candidates

1. In your experience, do you find assessors or IQAs giving appropriate levels of support, guidance and feedback to learners?
    - What about inappropriate help/leading during assessment?
    - How do you deal with this?
2. What about personality clashes, discrimination, positive or negative bias?
    - How do you pick this up?

Assessor training

1. Do you think that the content of assessor qualifications is appropriate to ensure that they do a good job with work-based observational assessment?

Consistency, standardisation

1. How consistent do you feel you are with other QCs in your industry?
    - Do you feel standardisation (if any) is effective?
    - Please give a few examples of standardisation activities you have in your company.
    - Do you need more or less standardisation?
2. How consistent are assessors within centre/nationally? What about IQAs?
    - Please give a few examples of standardisation activities assessors normally undertake?
    - Difference between work-based assessors and peripatetic assessors?
    - Do you feel standardisation (if any) is effective?
    - Do they need more or less standardisation?
    - What activities might work best?

Current procedures

What specifically do you do in order to quality assure work-based observational assessment?

Does it require more targeted procedures? More time? More frequent visits?

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