On the reliability of results in vocational assessment:
the case of work-based certifications

A project report prepared by City & Guilds as part of the Office of Qualifications and Examinations Regulation's Reliability programme by

Helen Harth
City & Guilds, UK

Dr Bas T Hemker
Cito, The Netherlands

Ofqual/11/4824 January 2011

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

Table of contents

Executive Summary

As part of the Office of Qualifications and Examinations Regulation's (Qfqual) two year Reliability Programme, the present research study investigates the reliability of assessment decisions in work-based vocational qualifications in England and factors that may affect the results. These qualifications are mainly used to confirm occupational competence of employees or for licence to practise. Their assessment regime frequently incorporates observation of naturally occurring evidence in the workplace, as well as oral or written questions, professional discussion, and review of portfolio and product evidence. Their assessment customarily involves a high degree of internally set and marked assessments that require evidence accumulation by the candidate and decision making by a human assessor. The total package of evidence used for judging, evaluating or interpreting someone's competence status can vary from candidate to candidate since the quality or quantity of the evidence may be influenced by local variations in assessment opportunities.

In the context provided for the project, reliability refers to 'the consistency of outcomes that would be observed from an assessment process were it to be repeated. High reliability means that broadly the same outcomes would arise. Unreliability can be attributed to "random", unsystematic causes of error in assessment results'. For vocational assessment this means that if a candidate was assessed again, by a different assessor or carrying out a different work task, would the person still be classified in the same category (ie either competent or not yet competent). Reliability of decisions is then about inter-rater (assessor/IV) agreement or consistency of decisions that can be influenced by diverse sources of error, such as for instance the complexity of the assessment procedures.

The present study aims to advance a conceptualisation of reliability theory in the case of assessment methods used in vocational certification. The methodology for data collection involves the collection and scoring of centre-devised assessment records from 324 candidate portfolios and of internal verifier reports in two vocational areas, Hairdressing and Electrotechnical Engineering, and three qualifications – Level 3 Electrotechnical Services (Electrical Installation – Buildings and Structures), Hairdressing NVQ (National Vocational Qualification) across several pathways at levels 1, 2 and 3 and the new NVQ Certificate/ Diploma in Hairdressing/ Barbering/ Combined Hair Types comprising of several pathways at levels 1, 2 and 3. Based on this live assessment data, the procedure for estimating inter-'item' reliability of binary assessor decisions is not standard, using a coefficient similar to Cronbach's alpha and Guttman's lambda. It further involves estimating inter-rater agreement and inter-rater reliability.

The main findings can be summarised as follows:

- The results suggest that inter-rater (assessor/IV) agreement is high (Gower coefficient ranging from .90 to .99) and inter-rater (assessor/IV) reliability (Cohen's kappa) is 'substantial' (for electrotechnical services) or 'almost perfect' (for hairdressing).

- Inter-'item' reliabilty (using a coefficient similar to Cronbach's alpha and Guttman's lambda) could only be estimated for the electrotechnical services and the results show high values.

- The procedures presented here confirm that it is possible to estimate the reliability of these qualifications, although changes would need to be made in the types of records used by assessors and the feedback given to candidates if this is to be carried out routinely.

- The flexibility required in the structure of these qualifications may prevent such procedures from being applied across all vocational qualifications.

- The verification process appears to work effectively in ensuring consistency of decisions and high inter-rater (assessor/IV) reliability, although further research may be required in this area at a time of radical change in English vocational education.

One of the features of this type of vocational assessment is that candidates are entered for summative assessment only 'when ready'. The decision to 'pass' a candidate is taken in a sense before the assessment takes place and the feedback captured by assessors may not include instances when the candidate's evidence had been evaluated but not yet considered to meet particular standard(s). The availability of natural assessment data for workplace assessments, with observation being the main type of evidence currently used in these qualifications, is then limited to the type of feedback captured by assessors. This study suggests that in order to carry out these procedures routinely, the organisations responsible for the quality assurance, control and regulation of such qualifications should consider the characteristics of the vocational assessment system, the assessment data that may be available for analysis and the appropriateness of measurement procedures for estimating assessment reliability and hence validity.

1 Introduction

Learners pursuing work-based, competence-based qualifications, also known as National Vocational Qualifications (NVQs), produce evidence to demonstrate that they possess the skill, knowledge and understanding or the competence required by the criteria contained in units of assessment that make up such qualifications. A variety of types of evidence (eg naturally occurring workplace tasks, questioning, traditional tests) may be presented by different candidates for the same criteria to decide whether or not they have achieved or not yet achieved particular criteria or whole units (Wolf, 1995). Due to the complexity of evidence available, most often human assessors are required to make these decisions.

Reliability in this context relates to the consistency of classifications – if a candidate was assessed on a different day, by a different assessor or carrying out a different task, would the person still be classified in the same category, eg competent or not yet competent? Can we be sure that two candidates at the same competence level are classified in the same way? In real assessment situations it may be expected that a certain amount of variation may affect decisions about a candidate's competence and lead to a degree of uncertainty or error in their results. Validity, on the other hand, is about the inferences that can be drawn based on the assessment decisions. Would the person classified as competent in an occupational area be able to perform at the level required after certification? The consistency of judgement in deciding when sufficient evidence has been provided may impact on the validity as well as the reliability of that assessment decision, which can affect the trust that should be placed in a particular assessment system (Clauser, Margolis & Case, 2006; Wilmut, Wood & Murphy, 1996; Brookhart, 2003).

The Regulatory Arrangements of the Qualifications and Credit Framework (QCF), the framework for recognising and accrediting vocational qualifications in England, Wales and Northern Ireland, states that assessments in vocational qualifications are required to:

- Be valid in relation to the learning outcomes against the stated assessment criteria.
- Produce sufficient evidence from learners to enable reliable and consistent judgements to be made about achievement of all the learning outcomes against the stated assessment criteria.
- Be manageable and cost effective.
- Be accessible.

(Ofqual, 2008, paragraph 5.3, p26)

In order to achieve these quality standards, awarding organisations put in place complex quality assurance systems that involve, amongst other things, the sampling of assessment decisions taken within centres (training providers, employers, colleges). Limited reliability (and validity) work has been carried out however for these competence-based qualifications in the United Kingdom (UK) (Greatorex, 2000; Johnson, 2006; Crisp & Novakovic, 2008).

This may be due in part to characteristics of the English work-based, criterion-referenced programmes. In this context, where the evidence produced by candidates needs to be predominantly naturally occurring, judged dichotomously (achieved/not yet achieved) over an unlimited number of attempts and only when the candidate is 'ready' for assessment, the feedback thus generated by assessors is not in the form of scores which are readily available for a reliability analysis. This project therefore aims to advance our understanding of the reliability theory and measurement procedures suitable in this context. It does not however explore issues surrounding the validity of decisions, which is outside the scope of this study.

Outline of this report

In section 2 we outline the research aims for this project followed by a description of the context of work-based vocational qualifications and their assessment regime in section 3. The important conceptualisation of reliability of decisions in vocational qualifications is presented in section 4. Section 5 describes the methods applied for data collection and analysis, while section 6 details the estimates under investigation for inter-rater (assessor) agreement and internal consistency for the qualifications we included in the study.

2 Research aims

In order to advance our understanding of the methodologies suitable for work-based qualifications and of the quality criteria that should be expected in vocational assessment, this study aims to:

1. Provide a background of work-based qualifications and their assessment regime.

2. Review the literature on reliability methods that could apply to our context and identify possible threats to the reliability of these decisions.

3. Provide a suitable methodology for collecting assessment data.

4. Formulate suitable procedures for estimating the inter-rater (assessor/IV) reliability and internal consistency of assessor decisions.

5. Provide a detailed discussion on the findings, including recommendations for developing a policy on reliability.

3 The context of work-based qualifications

Since their inception over 20 years ago, NVQs have been taken up by people in the workplace or other settings that replicate a working environment, in vocational areas such as construction, engineering, service industries, health and social care, business administration or management. In England, Wales and Northern Ireland accredited qualifications are regulated through joint credit systems or frameworks, such as the National Qualifications Framework (NQF) or the Qualifications and Credit Framework (QCF)[1], which allow levels of achievement to be compared across qualifications.

3.1 Purposes

The qualifications have been used primarily for employment purposes, specifically for confirmation of occupational competence against the national occupational standards[2] (NOS) and for licence to practise (Ofqual, 2009). The results may also be used to monitor learner completion rates, especially important for qualifications approved for public funding, provide feedback to candidates for future improvement, evaluate the effectiveness of assessor performance or for access into higher education (Kingston, 2007). Given the purposes associated with these competence-based qualifications, the assessment decisions are normally high-stakes, regardless of the assessment design.

3.2 Units of assessment

In order to achieve a qualification at a certain level, candidates are required to prove that they meet a set of criteria contained in a unit of assessment (the smallest component of a qualification), which are based on the NOS. QCF units are at different levels and use the same template, consisting of learning outcomes and associated assessment criteria (see for example City & Guilds, 2009a).

A number of mandatory or general and optional or specialist units may be required for the achievement of a qualification, with credit being awarded for completion of a unit. The rules of combination then state which credits can be combined to complete a specific qualification. Different pathways are available to candidates which support the varying demands required by industry sectors (QCDA, 2010). For example, a hairdressing qualification can be offered at different levels across a number of pathways and units, as shown in Appendix 1, Figures 1 and 2.

3.3 Occupational standards

In the NVQ Code of Practice[3] for competence-based qualifications, competence is about persons who possess 'the ability to carry out activities to the standards required' (NVQ Code of Practice, QCA, 2006, p37). A similar meaning of competence has been conveyed in the QCF unit writing guidelines, where units of assessment are linked to NOS to 'focus on the knowledge, skills and understanding, which, applied together, form the

[1] The NQF, which was introduced for this purpose, is currently being replaced by a new framework, the QCF that in addition indicates the size of qualifications (measured in learning hours or credits), as well as their level (see City & Guilds, 2009a).
[2] National Occupational Standards (NOS) are statements that 'describe what a person needs to do, know and understand in a job to carry out the role in a consistent and competent way' in a particular environment (UK Commission for Employment (UKCES) & Skills and the Alliance of Sector Skills Councils (SASSC), 2010)
[3] The NVQ Code of Practice was developed for qualifications on the NQF, but no longer applies to qualifications in the QCF (see UKCES, 2008).

competence required by employers for certain roles and functions' against which an individual's performance can be judged (QCDA, 2010, p11; see also Wolf, 1995, p30 for a discussion). Standards are the descriptions of the following elements that make up a unit:

- Learning outcomes are equally weighted in terms of achieving a unit. Taken together, they describe the competence that a candidate who has credit for the unit should possess and so may cover diverse sub-domains. Table 1 displays an example of a level 1 unit from hairdressing.

- Assessment criteria that specify the standard of performance a learner must meet to demonstrate mastery or achievement of the learning outcome (Ofqual, 2008, paragraphs 1.4d and 1.5a).

- Range of achievement[4] which describes the circumstances, context, combinations of methods, number or frequency of occasions and levels of responsibility in which competence can be demonstrated (QCDA, 2010) (see Table 1).

A candidate can only be judged occupationally competent when the person has provided sufficient evidence to fulfil all the requirements of the unit.

[4] In a QCF unit, the full range or scope can be expressed either in the additional information about the unit or it may also be included in the assessment criteria (see FAB/JCQ, 2010; example in Table 1, Unit additional assessment requirements).

Table 1 – Example of performance and knowledge criteria from level 1 'Plait and twist hair using basic techniques' unit (NDAQ, 2010)

Unit additional assessment requirements	The assessment of this unit needs to meet the requirements within the Habia Hairdressing and Barbering Assessment Strategies [the standard setting body for the hair, beauty, nails and spa industries]: […] 3. The assessor will observe the learners performance on at least 3 occasions which must include observation of: - a minimum of 5 cornrows - a single French plait - a series of small two strand twists covering a minimum of 25% of the head. 4. The learner must show that they have: - used all the types of products: a) sprays; b) serums; c) gels. - created all the types of plaits and twists: a) multiple cornrows; b) French plait; c) two strand twists. […]
Learning Outcome	*Assessment Criteria*
2. Be able to plait and twist hair	2.1 prepare the client's hair following instructions from the stylist 2.2 control tools to minimise the risk of damage to the hair and scalp, client discomfort and to achieve the desired look 2.3 part the sections cleanly and evenly to achieve the direction of the plait(s) and twists 2.4 secure any hair not being plaited or twisted to keep the section clearly visible 2.5 maintain a suitable and even tension throughout the plaiting and twisting process 2.6 control and secure the client's hair, when necessary 2.7 apply suitable products, when used, to meet manufacturers' and stylist's instructions 2.8 consult with the client during the plaiting and twisting process to ensure the tension is comfortable 2.9 adjust the tension of plaits, when necessary, avoiding damage to the hair and minimising discomfort to the client 2.10 make sure that the direction and balance of the finished plait(s) and twists meets the stylist's instructions 2.11 confirm the client's satisfaction with the finished look.
7. Know products and their use	7.1 identify the types of products available for use with plaits and twists and when to use them 7.2 state the importance of using products economically.

3.4 Decision rules

In order to achieve a unit and/or qualification, a complex combination of decision rules are applied (Ryan & Hess, 1999; Chester, 2003). These include conjunctive and complementary procedures, such as:

- Conjunctive procedures that require that all of the learning outcomes must be met for a 'pass' to be awarded. Better performance in some areas cannot compensate other areas, which may not have been achieved.

- Because in principle the QCF allows for a unit to be substituted by another if the rules of combination allow, these equivalent pathways add a complementary rule to the conjunctive rule.

- The assessment may combine different types of assessment methods, tasks and may accept accreditation of prior learning or experience which add a complementary approach to deciding competence.

Table 2 summarises the decision rules to fulfil the requirements for achieving competence-based qualifications.

Table 2 – The approach used in the QCF to combine multiple measures to reach assessment and qualification classification decisions (based on Chester, 2003)

	Conjunctive AND	*Compensatory +/-*	*Complementary OR*
Measures of different constructs	Minimum performance of competence required on all learning outcomes/assessment criteria and all units must be a 'pass' according to the rules of combination	Tests may be used but there must be evidence that all learning outcomes have been achieved	Choice of optional units (a combination of units may be taken depending on the chosen pathway)
Different measures of the same construct	To cover the range and confirm the inferences made, multiple sources of evidence may be required	-	Criteria covered through tests that are not achieved may be covered using additional instruments
Multiple opportunities	-	Unlimited number of re-takes allowed	Evidence is generated until the standard is achieved
Accommodations and alternate assessments	-	-	Accessibility arrangements or using supplementary evidence

3.5 Vocational assessment

The assessment of performance, knowledge and understanding in vocational qualifications customarily involves a high degree of internally set and marked assessments. While the model may use externally set tests, the emphasis is on competence in relation to naturally occurring activity in the workplace. In the case of externally set and marked assessments it is relatively easy to measure and ensure the reliability of such results. Where the assessment is tailored to local conditions however, though nationally recognised, the onus is on the assessor to judge, evaluate or interpret the value of the evidence presented by the candidate and whether sufficient evidence has been produced against the unit content and range (see Ofqual, 2008, paragraph 5.3d, p26).

The assessor observes the candidate at work carrying out tasks, questioning about what they are doing to ensure knowledge and understanding (see Table 3). When the candidate is judged as not yet competent, it may be because they have not achieved all of the assessment requirements either based on sufficient evidence (eg candidate has tried but not met the criteria) or based on insufficient evidence (eg candidate does not have enough opportunity to perform a task and therefore cannot be judged as fully achieving the outcome or unit).

Table 3 – Sources of evidence which may be used in a work-based qualification

Source of evidence	*Types*
Observation of performance at work	The assessor observes the candidate working in order to assess his/her performance against the unit requirements. Evidence may also come from a witness, log or diary of the candidate's work.
Observation of performance of specifically set tasks	Candidates may be required to perform a particular activity (eg simulated task, project or case study)
Questioning	The knowledge and understanding elements of a unit are assessed through oral or written questions
Historical evidence	Prior experience or learning may be used as evidence provided it is sufficiently current

Some aspects of the range may also be covered by simulated tasks, tests, assignments or other means only when appropriate, where naturally occurring evidence is not available or dangerous. To show that they can work to the standards contained in a unit (which may include performance, knowledge and understanding) secondary, supplementary or complementary types of evidence are also used for judging someone's competence status. This means that 'a total package of evidence' can vary from candidate to candidate since the quality or quantity of the evidence is influenced by local variations in assessment opportunities (Mitchell & Bartram, 1994). For instance, assessors may feel compelled to add further evidence such as additional questioning to ensure that a unit standard was achieved. Where tests are used and the candidate has failed to answer a question, the assessor is required to ensure that that particular area of knowledge has been covered through other means, normally by oral questioning or another sitting of the test (only of the elements not achieved in previous sittings). In principle, the cutoff point for these assessments is 100%.

Since the candidate is assessed only when both the assessor and the candidate are reasonably confident that the person will be successful, the decision to pass a candidate is in a sense taken before the actual summative assessment. Even when traditional tests are used (eg multiple choice tests), high success rates on the test items are normally expected. This implies that a large proportion of observations will classify candidates as 'achieving the required standard'.

In summary, the assessment of work-based qualifications or NVQs has the following characteristics:

1. The candidate is assessed on demand, 'when ready' and his/her performance on each

task is scored dichotomously as competent/not yet competent against individual criteria, normally by a human assessor. The assessment is continuous, made up for example of observations of unique naturally occurring/set work tasks or products, professional discussion, assignments, questioning or witness statements. Any single assessment occasion may cover a number of criteria and units across a range or circumstances. Timescales for completion and marking of individual assessments may vary and in principle the process continues until a positive decision has been made, or the process is abandoned.

2. Decisions for each unit in the rules of combination are used as the basis for awarding the qualification when all criteria and/or outcomes need to be achieved (see Table 2 above).

3. Candidate performance and the assessor decisions are internally and externally verified for quality assurance and control.

Candidates practise the skill or conceptual knowledge continuously through carrying out tasks in the target domain. The assessor or tutor may provide feedback, support, help or reminders until the person is increasingly more independent and could be considered 'ready for assessment'(eg Collins, Brown & Newman, 1989; Brown, Collins & Duguid, 1989). This feedback is often oral, but it may also be captured in learner diaries or assessment records which are logged in a portfolio and referenced back to the unit criteria (or NOS) once achieved (eg as judged by the assessor against the unit content). The assessment is therefore continuous and takes on dual formative and summative functions.

3.6 The verification process

It is the responsibility of the awarding organisation to ensure 'the accuracy and consistency of standards in the assessment of units, across units and over time' (Ofqual, 2008, paragraph 5.6c). To this end, such organisations establish verification processes that provide an important quality assurance function.

Internal verifiers (IVs) are required to implement a verification plan, which includes a schedule of activities such as producing a sampling strategy for individual assessors, directly observing sample assessments carried out by assessors, reviewing candidate evidence, conducting candidate interviews. Centres use various standardisation procedures, for instance assessor training, sampling of assessor decisions by the internal/external verifiers, or access to a community of practice by organising networking opportunities. This process intends to support the assessor in conceptualising the standard required and ensures this across centres/regions. Figure 3 in Appendix 2 shows an example of an internal verification plan used by a centre. The report depicts whether the internal verifier agrees with the assessor's decision and also whether he/she considers the candidate to be competent in the unit content.

The external verifier (EV) is independent of the centre and accountable to the awarding organisation. The person is responsible, amongst other things, for ensuring that assessment decisions are fair, consistent and meet the requirements set out by the national occupational standards. They will also sample decisions taken by assessors and internal verifiers by observing staff or reviewing portfolio evidence. Both the IV and the EV are

required to have industry-recognised occupational qualifications as well as relevant assessor/verifier qualifications.

Because local flexibility is paramount for these qualifications, the unit content, verification of assessor decisions, assessor/verifier training, networking and documentation (eg assessor handbooks produced by awarding organisations) are meant to ensure uniformity across different assessment situations and assessors. While test-based decisions may be required in some cases (eg using multiple choice questions), the most frequently used assessment methods emphasise the role of the assessor and the inextricable relationship between assessment and instruction. The standard is contained in the assessment criteria (evidenced through work samples) which represent performance of the skill or craft rather than a cut-score, since no marks are given. The evidence required needs to ensure the right balance among what needs to be assessed, ie the unit content, the minimum amount of evidence that would be considered sufficient and the minimum number of contexts that must be covered, while ensuring enough flexibility of assessment is allowed. In addition, candidates are put through formal assessment only when the tutor/assessor and the candidate have the confidence that competence in a particular area has already been achieved (see Wolf, 1995).

The issue is then whether the assessment activities in this context lead to valid and reliable decisions that serve their uses and purposes well. Characteristics of the assessment system (decisions rather than scores, work tasks, criteria and range rather than test items or a cutoff point), its purposes, the high stakes status of such certificates and the decision rules imposed by the qualifications framework lead to a particular conceptualisation of reliability.

4 Conceptualising reliability of decisions in vocational qualifications

Reliability is about quantifying the precision, stability or consistency of decisions based on candidates' performance on a task (Traub & Rawley, 1991). Reliability is therefore inversely related to measurement error, the variation that can be expected if the assessment procedure were to be repeated. Reliability can be calculated using a variety of statistical methods. High values for these reliability estimates indicate a small amount of measurement error (AERA, APA, NCME, 1999; Haertel, 2006). Reliability estimates would then indicate the degree of confidence that should be placed in someone's results, although the decision should also be valid.

4.1 Consistency of classification decisions

In contrast to assessments where the emphasis is on standardisation or one or more cut scores to define the decision rule, criterion referenced dichotomous (pass/fail) decisions based on unique workplace evidence pose alternative challenges to measurement theories and approaches to reliability developed for traditional assessments may not be suitable (Mislevy, 1994; Brookhart, 2003; Johnson & Johnson, 2009). In the context of 'on-the-job' performance, where each task is different and may not be repeatable in exactly the same circumstances (eg client/job requirements are varied and ever changing), test-retest or parallel form reliability can be more difficult to interpret, depending on the subject of the qualification or level (Traub & Rawley, 1980; Berk, 1980; Verhelst, 2004; Clauser, Margolis & Case, 2006).

Such a conceptualisation is further challenged by the fact that the candidate is allowed in principle to accumulate evidence until a positive decision can be reached. It is true of course that equivalent evidence in value, that is equivalent in levels of confidence, is customarily accepted by assessors and verifiers, which means that whether these assessments can be considered parallel depends on how strictly we define the properties of a parallel test. A loose interpretation of parallel assessments may even argue that various units can also function as parallel tests and that the measurement of internal consistency is related to the measure of parallel test, since the correlation between the units may be considered as a lower bound for parallel test reliability.

The concern for this type of mastery interpretation is primarily on the consistency of classification decisions, often referred to as reliability of classifications, rather than score reliability as far as measurement error is concerned (Huynh, 1976; Subkoviak, 1976). This is a measure of agreement or consistency of achieved/not yet achieved decisions across repeated applications of the procedures (Swaminathan, Hambleton & Algina, 1974; AERA, APA, NCME, 1999; also Lee, Hanson & Brennan, 2002; Greatorex, 2002, 2005; Greatorex & Shannon, 2003). Replication in this context would mean that if we were to judge a candidate again on a different occasion, the same judgement should be made based on the evidence provided. The decisions should be consistent across equivalent evidence packages for different assessors, providers, regions and time (Murphy et al, 1995; Wilmut, Woods & Murphy, 1996). Further assessor decisions can be considered valid and reliable when they accurately reflect the level of performance, which has been consistently demonstrated by the candidate.

High agreement among independent raters (assessors and verifiers) regarding a candidate's classification can be further evidence of the consistency of decisions.

4.2 Sufficient evidence should lead to uniform levels of confidence

Candidates in the workplace are continuously seeking to provide evidence that confirms their mastery of the skill over a period of time and the final decision to certify a candidate taking a vocational qualification is the result of an assessment process based on a large number of decisions coming from multiple occasions and multiple measures rather than a single administration of a test.

Reliability of these decisions can then be conceptualised in terms of the assessor's level of confidence in his/her judgement that the candidate has produced evidence of sufficient quality and quantity that would indicate that the person can do the job based on the standards contained in the unit of assessment[5]. The aim should be to minimise the chances of judging someone as competent who has not actually met the unit requirements, since holding someone back would be less serious when the assessment is on demand and continuous, although such a situation could increase the cost of assessment. This would also ensure the assessment is cost effective in achieving optimal levels of confidence (Mitchell & Bartram, 1994). A key issue would then be whether candidates' evidence against the unit content that are equivalent in value are of sufficient quality and quantity and also associated with equivalent levels of confidence by different assessors. Reliability in this way would be about applying assessment standards uniformly across a qualification.

In addition, reliable assessment decisions using a number of different packages of evidence is about converging or accumulating the evidence that support the same inference rather than joining of scores (Mislevy, 1994). Although different in detail amongst candidates, the amassed body of portfolio evidence should give a consistent level of confidence over candidates and assessors. Mislevy (1994) uses analogies from scientific research, medical diagnosis or legal reasoning to define reliability as the 'weight of evidence' and the 'relevance' of a particular component (eg assessment within a unit) and how they relate to the inferences made. The nature of the decisions would then argue for an alternative conceptualisation of reliability in terms of sufficiency of information to infer a candidate's competence (Ofqual, 2008; Smith, 2003).

4.3 Validity

The accuracy or consistency of decisions do not include the broader issue of validity of decisions. The question here is whether the certificate (based on these decisions) represents the ability the assessment intends to measure so that the person could confidently be employed on this basis and would the person be able to perform to the standards required by the qualification after certification (Clauser, Margolis & Case, 2006; Wolf, 1995). In this context, an important aspect of the quality of vocational assessment would be the validity of assessor decisions and the criteria that can be used to evaluate those (Wools, Eggen & Sanders, 2010). Such a study is however beyond the scope of this research.

[5] In reaching their decisions for instance, assessors, internal and external verifiers are required to ensure the validity, authenticity, currency and sufficiency (known as the VACS rule) of the evidence provided (eg see City & Guilds, 2009b, p29).

4.4 Factors influencing reliability

Approaches to evaluating reliability consider inconsistency in the classifications required in the vocational assessment system described here to be a type of error (Nichols & Smith, 1998). Assessors and verifiers involved in deciding when sufficient evidence has been provided by candidates may overestimate the evidence, judging too many people as competent when they are not, resulting in high false positive rates (ie people who should have failed but passed), or under-estimating the evidence, judging too many people as not yet achieving the standard, resulting in lower false positive rates but increasing the false negative rates (ie candidates who should have passed but failed).

Other influencing factors may have to do with characteristics of workplace assessment – eg the decision making process, characteristics of the assessment system, the verification process or the decision rules – which are all thought to affect the reliability of these decisions (Driessen et al, 2005; Wolf, 1995; Murphy et al, 1995; Eraut et al, 1996; Greatorex, 2002, 2005; Greatorex & Shannon, 2003; Johnson, 2008a, b; Lane & Stone, 2006; Cronbach & Gleser, 1957; Cronbach et al, 1997; Chester, 2003; Douglas, 2007; Good, 2002). In this case, the expectation is that lower estimates of reliability will be obtained (Murphy et al, 1995). Despite being a threat to the reliability of decisions however, such characteristics of the vocational assessment system are desirable since they can ensure flexibility for candidates, professional relevance and practicality for 'on-the-job' assessment and should not diminish our trust in the system.

Due to limited previous reliability research in this area, it is not yet possible to substantiate these claims, while certain features of these assessments should minimise the risk to the inconsistency and variability of these results. As we mentioned previously, in the context of work-based qualifications in the UK, the reliability of assessment decisions and/or certifications is not normally measured at an operational level and reliability is expected to be optimised by centres and awarding organisations through internal and external verification procedures (Ofqual, 2008, paragraph 5.6.c). This study aims to address this issue by proposing methods for data collection and analysis.

4.5 Estimating decision consistency

For the types of decisions and the nature of these vocational assessments, the methods discussed below may be appropriate.

Reliability indices used to quantify classification consistency are analogous to reliability measures that can show how stable a classification of competent/not yet competent is for each criterion, unit and qualification. This is expressed in terms of classification accuracy and it is a measure of the probability of classifying or misclassifying a candidate as meeting the required occupational standards (Clauser, Margolis & Case, 2006).

The reliability of test scores (equation 1) is defined as the ratio of true score variance to total observed score variance, hence error directly influences the reliability index ($\rho_{XX'}$, equation 1). In the case of single assessment administrations, estimates of reliability such as Kuder-Richardson 20 (KR-20) or coefficient alpha can be used (eg Cronbach, 1951). Cronbach's alpha (α) coefficient is an estimate of reliability known to be a lower bound for the reliability and then it would underestimate the true reliability (Lord & Novick, 1968; Osburn, 2000). A better estimate for reliability value is considered to be Guttman's

lambda-2 (λ_{-2}) with the following relationship with Cronbach's alpha and reliability in Equation 1 below (eg Sijtsma, 2009).

$$\alpha \leq \lambda_{-2} \leq \rho_{XX'} \qquad \textit{(Equation 1)}$$

In the context of classroom assessment, Buckendahl, Yang and Ferdous (2003) evaluate the level of agreement between reliability analyses using the coefficient alpha as a measure of internal consistency and a proposed decision consistency strategy (percent agreement) that uses teacher judgments of student proficiency on a written task. In the case of workplace assessment, where each score may represent different types of evidence, Cronbach's alpha may be viewed as a measure of how well the sum of units capture the expected score in the entire domain. Generalizability (g-) theory is another approach that has been suggested as a suitable alternative for work-based qualifications (Johnson & Johnson, 2009). Cronbach's alpha may however also be considered as an unbiased estimate of the generalizability (eg Brennan, 2001).

4.6 Estimating assessor agreement

Because the assessment system relies on assessor decisions to gauge candidate competence classification which is then verified by an internal and an external verifier using a complex sampling matrix, inter-rater (assessor/IV) reliability that quantifies the closeness of these independent ratings may be a useful measure of the quality of these decisions (Ebel & Frisbie, 1991). Verification and standardisation procedures and high pass rates may lead to high levels of assessor agreement.

A useful measure for inter-rater agreement where we have small variances in judgments is Gower's coefficient (Gower, 1966, 1971). This coefficient is equal to 1 minus the ratio of the sum of the absolute differences of the two judgements of two raters (assessors) over n objects and n times the range of the judgements per object. As a result of the absolute method of norming, Gower's coefficient can be interpreted as a measure of average agreement between judges per object (Zegers, 1991). In the case of dichotomous scores, as it is the case with pass-fail decisions, Gower's coefficient is identical to the proportion agreement. This coefficient is useful where there is limited variation in scores or judgments and in the case of extreme pass rates.

The procedures proposed by Cohen (1960, 1968) may also be suitable to the types of dichotomous ratings made by two raters (assessors) in these qualifications (see Huyhn, 1978 for the relationship between kappa and other parameters). Cohen's coefficient kappa estimates how much classifications will improve decision consistency relative to a random classification (Gwet, 2002). For example, two assessors A and B classify N candidates as positive (+) if they achieve a unit (criteria) and negative (-) if they have not, as described in Table 4. In this example, a is the number of candidates which were classified as achieved by both raters A and B, b is the number of candidates classified as achieved by rater B and not yet achieved by rater A, c the number of cases classified as not yet achieved by rater B and achieved by rater A and finally d is the number of cases for which both raters classify candidates as not yet achieved. In this example, a and d are cases of agreement, while b and c are cases of disagreement between the two raters.

Table 4 – Distribution of N candidates by assessor and response

	Rater A		
Rater B	+	-	*Total*
+	a	b	B_+
-	c	d	B_-
Total	A_+	A_-	N

The overall agreement probability P_a is given by the formula in equation 2:

$$P_a = \frac{a+d}{N} \qquad \textit{(Equation 2)}$$

where a and d are as described above. In our example, the index P_a is the proportion of candidates that both assessors classified in the same categories. The inter-rater reliability as measured by Cohen's kappa statistic defined in equation 3 below. A

$$\kappa A = \frac{P_a - P_e(\kappa)}{1 - P_e(\kappa)} \qquad \textit{(Equation 3)}$$

where $P_e(\kappa)$ is Cohen's measure of the likelihood of agreement by chance, expressed as in equation 4 below.

$$P_e(\kappa) = P_{A_+} P_{B_+} + P_{A_-} P_{B_-} \qquad \textit{(Equation 4)}$$

and P_{A+}=A+/N, P_{A-}=A-/N, P_{B+}=B+/N, P_{B-}=B-/N are rater-specific classification probabilities, where P_{A+}, P_{B+}, P_{A-}, P_{B-} are the probabilities that respectively raters A and B classify a candidate as positive or negative. Kappa may also be given as in Equation 5 below:

$$\kappa = 1 - \frac{\text{observed disagreement}}{\text{expected disagreement}} \qquad \textit{(Equation 5)}$$

with

Expected Disagreement = N (P_{A-} P_{B+}+ P_{A+} P_{B-}) using the same notations as above.

The value thus obtained varies from -1 (perfect disagreement) to 1 (perfect agreement), where a value equal to 0 means that, given the probabilities of passing according to the assessor and the IV as fixed values, the classifications are not better than those obtained by chance.

Values closer to 1 mean that something other than random factors account for the two independent ratings. Cohen's kappa can be used since it is considered to be a more robust measure of agreement than a straightforward percent agreement calculation and it also takes into account the agreement occurring by chance. Kappa would indicate how well the two ratings agree not with the 'true' classification but the 'true' agreement, beyond that expected by chance. In the situation presented by work-place assessment, the value

of kappa would therefore represent the extent to which assessors and internal verifiers agree in their ratings (competent/not yet competent). A number of factors (eg prevalence, bias and independence) can however influence the magnitude of kappa (Gwet, 2006).

It is important to note that kappa can be considered a conservative measure of agreement because it uses the frequencies of the observed categories. Landis and Koch (1977) suggest the commonly used interpretations for Cohen's kappa values described in Table 5 below.

Table 5 – Common interpretations for Cohen's kappa coefficient (from Landis & Koch, 1977)

Kappa			*Interpretation*
<.00	-		'poor'
.00	-	.20	'slight'
.21	-	.40	'fair'
.41	-	.60	'moderate'
.61	-	.80	'substantial'
.81	-	1.00	'almost perfect'

Note that the maximum value of kappa can only be obtained in the case of identical marginal distributions. Because it is difficult to interpret the value of kappa without consideration to the context of each assessment, interpretations such as those of Landis & Koch in Table 5 have been challenged. They are however still able to provide a broad description of the agreement between raters where no historical data is available as it is the case with our study (see Gwet, 2002, 2008).

It may thus be possible to use coefficients of agreement for measuring inter-rater (assessor/IV) reliability, and depending on characteristics of the data, Cohen's kappa seems to be a feasible option. In the case of inter-item consistency where traditional measures may not be suitable, a non-standard method may have to be applied while g-theory may also be suitable (Johnson & Johnson, 2009).

5 Method

Our study seeks to offer suitable procedures for the measurement of inter-rater (assessor/IV) and inter-'item' reliability. Our approach to the data collection and analysis therefore involved

1. Identification of the most appropriate data collection strategy and the qualifications which should be included in the study;

2. Data collection for these qualifications;

3. Compilation of necessary data for analysis;

4. Data analysis which included the generation of suitable procedures for estimating the reliability of assessment results (inter-rater reliability and inter-'item' consistency).

In the first stage of the data collection, we investigated the types of assessment, procedures and assessment records used by centres in the certification process for work-based qualifications. The main aim of this activity was to find the most appropriate method for the data collection. A researcher visited at least one centre offering each of the qualifications targeted in the study to collect information about centre devised and marked assessments and exemplars of assessment records used in the process.

For these purposes, assessment records were collected for a number of qualifications. The criteria for inclusion were that the qualification is awarded by City & Guilds, either in the QCF or the NQF, with the purpose of confirming occupational competence, delivered in the workplace or an NVQ, and that the accreditation end date has not yet passed. Another criterion for inclusion was that it has at least 1000 candidates registered in order to optimise our chances of collecting a sufficiently large sample.

5.1 Participant recruitment

Tutors or assessors from centres (training providers, colleges and employers) delivering City & Guilds qualifications across all English regions were contacted (phone and email) asking them to participate. Willing volunteers were emailed a study information sheet explaining broadly what was expected from them and the purposes of the study; they were compensated for their time and effort. In total 22 centres which delivered one or several of the nine qualifications being investigated participated in the study, although not all qualifications could be selected for inclusion in the data entry and analysis.

5.2 Materials

For the purposes of estimating decision consistency, our aim was to collect data that already existed as part of the candidates' usual assessment process, with minimum disruption to assessors and candidates. For this study we were interested in the assessor/IV decisions on whether a candidate had achieved or not yet achieved a particular criterion. We therefore requested the following materials from anonymous candidate portfolios across all of the units included in the portfolio (described in Table 6):

- All assessment records available from candidate portfolios.
- IV reports.

The records were either photocopied or photographed by centre staff or by a researcher with the centre's permission. We did not require copies of any work products/exemplars which are normally included in a candidate's portfolio. As much as possible, full portfolios were requested, but due to the fact that portfolios belong to the candidate who may not be available upon completion of the qualification, not all the records included were from complete portfolios.

Four types of assessment records were used by the centres sampled in our study to track their learners' progress and provide feedback to candidates as well as an internal verification report filed separately by the centre (see Table 6). They normally covered an assessment event such as an observation of performance in the workplace, occasionally supplemented by a professional discussion, witness testimony, examination of a work product, oral/written questions (open-ended, multiple choice), assignments and so forth. An assessment record would therefore cover one or several units and their related learning outcomes and assessment criteria.

Table 6 – Types of records identified in our sample

No	*Type*	*Assessment record*
1	Work site assessment observation record/feedback sheet	Captures the assessor decision against a unit/learning outcome or assessment criteria as competent/not yet competent. The types of evidence can be performance assessment, questioning, witness testimony.
2	Evidence location and summary of achievement sheets	Links the items of evidence to the criteria that have been achieved, but with no feedback on what was not yet achieved on a particular piece of evidence. Items are added only once achieved and it may allow the IV and EV to indicate which units have been sampled for verification purposes.
3	Performance evidence record	It may be self-reflective account and signed off by the assessor/witness. Links candidate evidence (performance on a task, questioning, witness statement) to achieved criteria and units.
4	Planning, feedback & judgement records (or progress review action plan meeting)	Identifies the criteria that are yet to be demonstrated and the actions required to gather the necessary evidence. It does not necessarily record criteria which have been demonstrated but were not yet achieved.
5	Internal verification report	This is a form completed by the IV for every verifier activity and indicates the units and methods of assessment sampled, whether the evidence is valid, authentic, current, consistent and sufficient, whether the IV agrees with the assessor's decision regarding the candidate's competence status and whether the assessor considers the candidate to be competent/not yet competent.

Work site assessment observation or feedback records (type 1) are used by some centres to capture a candidate's classification into 'competent'/'not yet competent'. The records

differ in the amount of detail included across centres, with some recording candidate competence classification against an assessment criterion and/or learning outcome while others only against whole units (see Figure 4, Appendix 2). These types of recording forms are sometimes used jointly to help the learner identify what has been achieved and what is yet to be achieved by allowing the assessor to give (extensive) qualitative feedback to the candidate (see Figures 5 and 6, Appendix 2).

We were also interested in the internal verification reports since they capture the IV's agreement with the assessment decisions made by the assessor and also the IV's judgement regarding candidate classification of competent/not yet competent. IV or other witness agreements (eg work-place manager, colleague or client) were also available on some of the assessment records included in the sample.

An additional generic study template was created for centres which were willing to participate in the study but did not use observation records as described above (type 1 record). The assessor and the internal verifier was asked to carry out their assessment as usual but independently from each other, following their internal verification plan or assessment schedules and fill in this standardised form to record their judgements on the candidate's performance.

5.3 Qualifications

Upon consideration of each type of assessment record, only qualifications and centres using type 1 records (described in Table 6) were selected for data entry. In the end, three qualifications from those looked at were suitable for our study, the two Hairdressing qualifications across three levels and several pathways and Level 3 Eletrotechnical Services (see Table 7).

Table 7 – Qualifications selected for data collection

Qualification name	*NDAQ Reference*	*City & Guilds qualification code*
Eletrotechnical Services (Level 3) (NQF)	100/2854/7	2356
Cert/Dipl Hairdressing (Levels 1,2,3) (QCF)	500/6662/6, 500/6355/8, 500/6509/9, 500/6573/7, 500/6574/9	3008
NVQ Hairdressing (Levels 1,2,3) (NQF)	500/1193/5, 100/3243/5, 100/3244/7, 100/3245/9	3014

Electrotechnical Services NVQ (Electrical Installation – Buildings & Structures) (City & Guild reference number 2356)

This qualification is aimed at electricians employed in the industry as proof that the learner meets the national occupational standards for an electrician based on evidence produced in the workplace. The learner would be self-employed or employed by an employer. The learner would have the opportunity to produce evidence from carrying out electrical installations on domestic or public premises. Due to the nature of this work, the assessor directly observes the candidate's performance when carrying out work tasks, but may also use oral questioning, reflective accounts of a particular job produced by the

candidate of a task or project he/she worked on, or testimonies from managers and work colleagues to confirm competence in areas which cannot be observed directly.

The Electrical Installation – Building & Structures pathway studied here comprises of four general (mandatory) units and four specialist (trade or optional) units. Due to large candidate registrations on this pathway it was possible to collect data for one pathway only. Note that there are no options of (groups of) units within the pathway.

Hairdressing Qualifications (City & Guild reference numbers 3014 and 3008)
The first hairdressing qualification included in the study is the NVQ in Hairdressing (3014) at levels 1, 2 and 3 (NQF). The qualification at level 1 is designed to accredit the skills of trainees and other salon staff as hairdressing assistants, providing support for other colleagues. Level 2 covers the skills required by a hairdresser, while level 3 reflects the roles of senior salon staff. Learning takes place through demonstration and instruction. Figure 1 in Appendix 1 shows the qualification structure in more detail.

The second Hairdressing qualification is the QCF NVQ Certificate/Diploma in Hairdressing/Barbering/Combined Hair Types (3008). This qualification combines the skills required for working in a hairdressing salon or barber shop, covering the full range of job options, from trainees to senior stylists. Learners have a choice of 6 pathways, having to accumulate credit from a number of mandatory and optional units, depending on credit size. Note that units in the rules of combination can be at a level different from that of the qualification. Because this is the updated QCF version of the 3014 qualification, some units may overlap. This complex structure is depicted in Figure 2, Appendix 1.

5.4 Data entry strategy

For the qualifications selected in the study, in order to be able to estimate the internal consistency of decisions, in the data entry we included only assessment records which contained specific feedback on candidates' classification status such as a tick against the candidate classification (eg 'competent NVQ assessment', 'pass', 'not yet competent', 'more training needed', 'fail') (see Appendix 2 for specific examples of such records). A trained researcher scored the assessor or internal verifier decisions for entry onto a data file and classified them in one of the categories described in Table 8 below.

In the case of inter-rater reliability, we are looking for at least two judgements of candidate's performance: that by the assessor and that by the internal verifier. These are found either in the internal verification reports or on assessment records, where the internal verifier signed off the decision made by the assessor, either based on direct observation at the time of assessment or on inspection of portfolio evidence. The data entry strategy differentiated full agreement from agreement 'with comments', where the internal verifier requested an action point to be fulfilled (eg authenticate evidence, include further evidence such as photographs) in spite of agreement being recorded. It is of course a matter of interpretation whether agreement 'with comments' should be different from full agreement and what a missing value would mean in this context. It is important to note that the results will depend on the interpretation of the data.

Table 8 – Codes used for the assessor decision categories

Witness status	*Decision code*	*Decision description*	*Examples*
Assessor	1	Achieved	Meeting the required standard/NVQ competent
	2	Achieved	With comments/action points
	0	Not yet achieved	Not meeting the required standard/referred assessment
	4, 5, 6, 8, 9	Not yet achieved	- Insufficient evidence to judge - Still to be assessed - Needing further development (work experience) - Ready for summative assessment (formative feedback only) - Could not carry out the task due to external circumstances (eg client allergic to hair colouring product, lack of opportunity, etc.)
	7	Not filled in	
Internal Verifier	1	Agree	
	2	Agree	With comments
	0	Disagree	
	7	N/A	Not needed for the task

Achieved sample

Although portfolio records were selected from a number of qualification, the data set used for this analysis included 20,885 valid entries representing assessor decisions (Table 9 summarises the data available per qualification). Note that a number of decisions may be required for the completion of a learning outcome, unit and then of the qualification.

Table 9 – Number of records per qualification

Qualification name	*Ref no*	*N Units*	*Records (freq)*	*Candidates (N)*	*Assessors (As'r)*	*As'r decisions*	*IVs*	*IV decisions*
Eletrotechnical Services (Level 3) (NQF)	2356	8	316	130	20	8,402	9*	2114
Cert/Dipl Hairdressing (QCF)	3008	43	864	101	42	4,030	10	485
NVQ Hairdressing (NQF)	3014	34	1,977	93	23	8,453	6	3074
3014/3008	-	51	-	194	51	12,483	14	3559
Total	**3**	**59**	**3,157**	**324**	**71**	**20,885**	**23**	**5,673**

* there was data from one unidentified IV

Eletrotechnical Services (Level 3) (NQF)

For this qualification, assessment records were available from 130 candidate portfolios and three centres, two further education colleges and one private training provider. The candidates were assessed by 20 different assessors (for some of the records the identity of the assessor was not available, only a signature). Results across all of the 8 required units

were recorded for 53 of the 130 candidates. The types of assessment used for each unit are described in Table 10 below. The assessments took place in the workplace or college.

Table 10 – The number of decisions recorded for each unit and type of assessment, 2356

Type of assessment	*301**	*302*	*303*	*304*	*305*	*306*	*307*	*308*	*Total*
Observation of candidate performance	635	150	402	761	612	637	509	181	3887
Oral open ended questions	67	42	88	193	116	160	135	20	821
Witness testimony/comment	28	.	.	9	33	26	3	.	99
Records of past activity/product evidence	6	.	2	3	1	13	4	6	35
Professional discussion	56	56	26	80	78	114	51	80	541
Review of portfolio evidence	273	273	272	272	272	272	272	273	2179
Documentary/written evidence	8	3		1	6	4	1	3	26
Candidate self-assessment/reflective account	14	9	7	9	16	31	15	20	121
Written open ended short answer questions	18	33	15	23	16	24	20	25	174
Observed product	12	20	15	3	63	45	35	6	199
Missing	48	24	21	52	59	55	39	22	320
Total	**1,165**	**610**	**848**	**1,406**	**1,272**	**1,381**	**1,084**	**636**	**8,402**

**301-308 are the qualification's unit reference numbers*

Due to features of the work carried out by electricians taking up this qualification, the evidence generated may be directly observed by an assessor and verifier, but it is also based on candidate reflective accounts of the tasks they carry out, supplemented by witness statements, photographic/video evidence, risk assessments, site plans and other documentary evidence they are required to use or produce as part of their job (Table 10 above).

Hairdressing qualifications (3008 and 3014)

Table 9 summarises the number of records entered for each qualification. The number of units, possible combinations of units and pathways in both qualifications was substantial, in contrast to the Electrotechnical qualification 2356 which had only one combination of units, while none of the candidates entered in the study had a complete set of units. As a result, in the case of 3014, there were only 14 candidates with the same combination of units, while for 3008 this number was 10. The main form of assessment recorded for both qualifications was observations of work tasks with 1597 decisions against learning

outcomes/assessment criteria in the case of 3008 and with 7419 decisions for 3014. The maximum number of observations recorded for a pair of units is 47 candidates and the number of combinations with 40 or more observations is limited (only 6 combinations of units). The set of seven units that has at least 30 observations for each combination of units (ranging from 34 to 47, with an average of 39 observations) contains units G5, G7, G10, G12, G13, H6 and H9.Because of some overlap of units in the two qualifications, it was possible to carry out some of the analysis together.

5.5 Quality assurance

The data was entered by four researchers and took an average of approximately 30 minutes per candidate portfolio to enter. The data entries were verified by a second researcher and cleansed for quality assurance purposes before analysis.

Confidentiality

The data we required from centres was purely for the purposes of this study and so all the information collected from centres, including centre details, assessor and candidate identity, was kept confidential and used solely for this analysis. Participants' names or any other personal means of identification have been kept anonymous so that individual performance cannot be known.

Challenges for data collection

Because the records we were interested in for the purposes of this study are not routinely collected by awarding organisations for analysis, the data collection for this study was challenging. In addition, not all centres involved in the qualifications investigated in this study employ assessment records which can be used in a reliability study. The types of assessment records varied across qualifications and across centres who were offering a particular qualification. This meant that a significant number of records collected had to be disregarded in the analysis.

Another challenge was the number of pathways and choice of units which need to be available to vocational learners pursuing a variety of roles in their chosen occupations. This meant that in the case of hairdressing qualifications, lower numbers of candidates were available per unit or group of units or no candidates who performed all units required for the qualification, even though large numbers of portfolios were entered in the study.

Despite these issues, we found that a large number of centres in the qualifications entered for data collection support procedures which allowed us to evaluate assessor/internal verifier agreement. Also, the Electrotechnical Services pathway registered a large number of candidates which made it possible for us to estimate the internal consistency of decisions.

6 Results

The data set resulting in this study allowed us to investigate two methods for measuring two types of reliability for the qualifications included here: inter-rater (assessor/IV) reliability and inter-'item' (unit) reliability.

6.1 Inter-rater agreement and reliability for the Electrotechnical Services qualification

For the inter-rater agreement and reliability analysis we considered all decisions for which we had two ratings. The passing rates are given in Tables 11 and 12 below. Percent agreement calculations of assessor-internal verifier full agreement was recorded in more than 90% of records and agreement 'with comments' in more than 96% of cases (see Table 13). This means that the Gower coefficient for Electrotechnical Services ranges from .90 to .96, which can be considered high. In total there are 1,282 entries for which there was an internal verifier decision based on an IV report, although we did not have matching candidate assessment records for all these reports. In this case, only the agreement was entered in the absence of the assessor decision. Because the number of 'agree with comments' decisions was small we did not add this as a different score category.

Table 11 – Assessor decisions passing rate for qualification 2356

Assessment decision	*N*	*Probability*	*Binary*	*Probability*
Fail	56	0.01		
Fail/missing	754	0.1	0.11	Fail
Pass w/ comments	100	0.01	0.89	Pass
Pass	6,562	0.88		
Missing	930			

Table 12 – Internal verifier decisions passing rate for qualification 2356

IV decision	*N*	*Probability*	*Binary*	*Probability*
Fail	91	0.07	0.07	Fail
Pass w/ comments	120	0.09	0.93	Pass (/)
Pass	1,071	0.84		
Missing (Sys Mis)	7,120			

Table 13 – Witness agreement for qualification 2356

	N	*%*	*% non-missing*
Disagreed	82	1	3.9
Agreed	1,912	22.8	90.4
Agreed w/ comments	120	1.4	5.7
Data on witness agreement	2,114	25.2	100
Missing	6,288	74.8	
Total	8,402	100	

Observed agreement and the probabilities of the decisions by assessor and internal verifier are used to calculate Cohen's kappa following two rules:

1. the ‘mild’ rule where ‘pass’ and ‘pass with comments’ agreements are considered to be observed agreement (comments allowed).

2. the ‘strict’ rule where only ‘pass’ without comments is considered to be observed agreement (no comments allowed).

Table 14 – Cohen's Kappa for qualification 2356 (mild rule)

	Assessor	*IV*	*Both (Expected)*
Pass (with or without comments)	0.89	0.93	0.828
Fail	0.11	0.07	0.008
Agreement	*Expected*	*Observed*	*Kappa*
	0.836	0.961	0.763

Table 15 – Cohen's Kappa for qualification 2356 (strict rule)

	Assessor	*IV*	*Both (Expected)*
Pass (without comments)	0.88	0.84	0.734
Fail	0.12	0.16	0.02
Agreement	*Expected*	*Observed*	*Kappa*
	0.754	0.904	0.612

In the case of the electrotechnical qualification, the verification procedures includes to a large extent review of portfolio evidence after the assessment decisions have been taken, rather than concomitantly. The aims are to check the assessor decision, whether the candidate has achieved the stated criteria and also to provide feedback in the form of action points that would help the assessor better conceptualise the standards they are working to. Because of this time lapse, candidates may add further evidence to the portfolio, often following assessor feedback, which may result in different decisions by the two raters. This is reinforced in the data by the number of cases rated as ‘not yet achieved – not enough evidence’ by the assessor, but getting a ‘pass’ from the internal verifier. In our data set 17 cases were classified as a ‘fail’ by the assessor but passed by the internal verifier. For the same cases, the internal verifier agreement was rated ‘agreed with comments’. Similarly, some candidates may have missing values in the case of assessor ratings, but be rated by the internal verifier. This may also explain why there is a higher percentage of ‘failures’ from assessors than from internal verifiers.

The consequence for the kappa values found here is that the true inter- rater (assessor) reliability that would be obtained if the ratings were made on exactly the same evidence would be higher than the kappa in Table 14. Following the ‘strict’ rule, a somewhat smaller value for kappa is found (see Table 15). According to the classification by Landis and Koch (1977, above) however, in both cases the kappa is considered to be ‘substantial’.

6.2 Inter- rater agreement and reliability for the Hairdressing qualifications

Following similar procedures as for qualification 2356 above, we analysed the inter-rater (assessor/IV) reliability for the Hairdressing qualifications. The pass rates for these qualifications are again very high as depicted in Table 16. Further, Table 17 describes the

cases of assessor-verifier agreement, while the frequencies of assessor decision by witness agreement are captured in Table 18.

Table 16 – Assessor decisions passing rate for qualifications 3008 and 3014

	3008	*3014*	*Total*	*% 3008*	*% 3014*	*Total*
Fail	427	102	529	0.11	0.01	0.04
Fail/missing	1	118	119	0.00	0.01	0.01
Pass w/comments	55	2	57	0.01	0.00	0.00
Pass	3,008	5,703	8,711	0.75	0.67	0.70
Missing (incl system missing)	539	2,528	3,067	0.13	0.30	0.25
Total	4,030	8,453	12,483			

From Table 17 it follows that the Gower coefficients for hairdressing qualifications range from .99 to 1.00, which can be considered extremely high. The number of internal verifier decisions recorded is limited compared to the total number of cases recorded (around 12% for 3008 and 36% for 3014). In only 7 cases the disagreement is about candidates' competence status (the same records for which the assessor recorded a 'pass' were considered a 'fail' by the internal verifier).

Table 17 – Witness agreement for qualifications 3008 and 3014

	N			*3008*	*3014*	*Total*
Witness agreement	*3008*	*3014*	*Total*	*% of non missing*		
Disagreed	0	7	7	0	0.2	0.2
Agreed	479	3,039	3,518	100	99	99.1
Agreed, with comments	0	24	24	0	0.8	0.7
Data with witness agreement	479	3,070	3,549	*% missing of total*		
Not needed for task	6	4	10	88.1	63.7	71.8
Missing	3,545	5,379	8,924			
Total	4,030	8,453	12,483			

Table 18 – Frequencies of assessor decision by witness agreement for 3008 and 3014

Decision	*Disagree*	*Agree*	*Agree w/ comm*	*Missing*	*Total*
Fail (0)	.	41	.	393	434
Pass (1)	.	3,284	3	5,424	8,711
Pass with comment (2)	.	2	.	55	57
Fail (4)	.	.	.	27	27
Fail/missing (5)	.	28	.	39	67
Fail (6)	.	0	.	68	68
Missing (7)	.	97	1	303	401
Fail/missing (8)	.	30	.	22	52
Missing (9)	.	9	.	30	39
System Missing	7	27	20	2573	2627
Total	7	3,518	24	8,934	12,483

Based on this description of the data (Tables 16-18) we were able to calculate values for Cohen's kappa given in Table 19 for each qualification and for the case when the two qualifications are joined, for both the 'strict' and 'mild' cases.

Table 19 – Cohen's Kappa values for qualifications 3008, 3014 and combined

Rule	*3008*	*3014*	*30xx*
Mild: Agreement and Pass (comments allowed)	1	0.971	0.979
Strict: Full Agreement and full Pass (no comments allowed)	1	0.953	0.954

The values for kappa found here are considered to be very high. The value for kappa in the case of the 3008 qualification indicates complete agreement between the two raters, who in the cases included in the study agree that all the evidence judged was sufficient for competent performance. In the case of qualification 3014 and 3008 joined with 3014 these results are also good. The results indicate that inter-item reliability in these cases is also good, which is a necessary condition for high reliability of the certification.

6.3 Reliability estimates on the basis of inter-'item' relations

For the cases where we do not have information on the 'items' and not every candidate is observed on each item, estimating the inter-item reliability may not seem feasible. If we consider units as items however, then inter-'item' or unit relationships may function as a useful measure of internal consistency of the certification.

Internal consistency estimates for Electrotechnical Services

For the Electrotechnical Services pathway (2356-31), there are a reasonable number of candidates that had a score on each unit, which allows us to estimate their reliability on the basis of inter-item (unit) consistency of a coefficient similar to Cronbach's alpha. The procedures involved in estimating the reliability on the basis on inter-item relations for this qualification are non-standard. Firstly, we score the units as if they were items based on some assumptions. Secondly, we need to consider appropriate ways of dealing with missing data. The third and final step is to consider what measure of internal consistency is most appropriate in this situation.

a. Candidate points on a unit

Similar to the inter-rater (assessor/IV) analysis above, two rules can be considered for awarding candidates a score on each unit:

1. The 'strict' rule: a score of 1 is given for a 'pass' only for the assessor decisions without comments;

2. The 'mild' rule: a score of 1 is given for a 'pass' both for with or without comments.

Below we analyse the reliability in both cases. Other types of decisions on a unit (or learning outcome/assessment criteria) such as 'fail – insufficient evidence to judge' or 'fail – not evidenced/still to be assessed' were also considered observations. An assessor decision was not recorded against all or parts of a unit, when the candidate could not perform the task (only one case) and system missing (no unit recorded for a particular candidate) were considered to be missing data. Table 20 shows the number of observations for each unit.

Table 20 – Frequency of assessor decisions per unit for qualification 2356

*Assessment decision**	*301*	*302*	*303*	*304*	*305*	*306*	*307*	*308*	*Total*	*%*
Fail (0)	7	14	2	5	2	6	6	12	54	1
Pass (1)	947	366	644	1,187	1039	1,150	868	407	6,608	79
Pass with comm (2)	15	11	11	15	11	11	15	10	99	1
Fail/missing (5)	1	1	1	1	1	1	1	1	8	0
Missing (7)	57	73	44	57	71	73	56	47	478	6
Fail/missing (8)	86	97	94	89	94	87	88	109	744	9
Missing (9)	38	41	38	36	39	37	36	43	308	4
System missing	14	7	14	16	15	16	14	7	103	1
Total	1,165	610	848	1,406	1,272	1,381	1,084	636	8,402	100

** see key in Table 8. 301-308 represent unit reference numbers that make up the 2356 pathway studied here*

A unit is scored for an individual candidate as in equation 6 below. As described above, a candidate may receive a score of 1 for 'pass' or 0 for 'fail' according to two rules. Because an observation is scored either 1 or 0, the maximum number of points on a unit for a candidate is equal to the number of observations, while the minimum number of points is 0 (no observations). Unit points therefore range from 0 to 100.

$$\text{Unit points} = 100 \times \frac{\text{score on unit for candidate}}{\text{total number of observations for candidate}} \qquad \textit{(Equation 6)}$$

b. Imputations for Missing values

The cases entered in the data set are not all from full or completed portfolios which results in missing data. The number of candidates with observations per unit is given in Table 21 below.

Table 21 – Number of candidates scored per unit

Unit	*301*	*302*	*303*	*304*	*305*	*306*	*307*	*308*	*Total*
N candidates	112	64	81	113	108	111	106	69	130
%	0.86	0.49	0.62	0.87	0.83	0.85	0.82	0.53	

Units that were observed for more than 80% of candidates are considered to be popular units (301, 304-307). In Table 22 the number of candidates is given by number of units observed. The number of candidates for whom none of the units was observed is equal to 15: 8 of them had no entries ('system missing') on all 8 units, whereas the other seven had a record on all 8 units, but the record was 7 'not filled in' (see Table 8). If we remove these 15 candidates from our analyses, we find that the popular units are observed for 94% to 98% of the 115 remaining candidates included in the analysis. The less popular units are observed for 56% to 70% of candidates.

Table 22 – Frequency of candidates with observations for a number of units observed

Number of units observed	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*Total*
N candidates	15	0	0	2	10	17	22	11	53	130

In Table 22 it is shown that there were no persons with records on only 1 or 2 units. The number of persons with scores on more than half of the units is 103. A total of 53 candidates have scores on all 8 units. Note that, in case of varying numbers of candidates, the number of observations depends on the pair of units observed. The pair of units that was least frequently observed was 302-308 with 57 observations, whereas the pair with most observations was of units 301-304 with 111 candidates (see Table 21).

The first imputation and easiest is to consider that a missing value is an indication that the unit is not achieved and the unit score = 0 (Imputation 0). Other imputations are possible based on the ability of the candidate (observed total unit scores for candidate) and difficulty of the item/unit (average score of the unit over candidates). This was considered reasonable for candidates with less than 50% of units missing (observation of at least 5 units), which means that 103 candidates are used in these analysis. Two types of imputations are considered for the missing values (Equations 6 and 7):

$$\text{Imputation i} = \frac{\text{Mean unit points on observed units of candidate}}{\text{Maximum unit points}} \times \text{Mean unit points of the missing units}$$

(Equation 7)

$$\text{Imputation j} = \text{Mean of unit points on observed units of candidate} \times \frac{\text{Mean unit points of missing units}}{\text{Mean unit points of all units}}$$

(Equation 8)

Different analyses were performed based on the different possibilities:

- Imputation - 4 levels: No imputations, Imputation 0, Imputation i, Imputation j
- Set of units - 2 levels: All 8 units, The 5 popular units
- Restriction on candidates - 5 levels: no restriction (include candidates with observations on any number of units), at least 1 unit observed, at least 5 units observed, all 5 popular units observed, all 8 units observed

After eliminating combinations which yield the same results or are not sensible, we investigated 10 different ways to deal with missing values.

c. Reliability estimates on inter-item relations

Two estimates were calculated here, Cronbach's alpha and Guttman's lambda. Based on these assumptions, we obtain 2 (strict/mild rules) x 10 (ways to deal with missing data) x 2 (types of coefficients) = 40 measures of reliability on the basis of inter-item relations. Table 23 provides the measures of reliability on inter-item relations obtained following these procedures.

Although the difference between the mild and strict scoring rules can be considered small, it was found that the strict scoring yields higher internal consistency. This is to be

expected given the difference between the two rules that can result in lower scores on units.

Table 23 – Measures of reliability based on inter-unit relations

					Alpha		*Lambda*	
Imputations	*Units*	*Cand w/Unit*	*N cand*	*N units*	*Strict*	*Mild*	*Strict*	*Mild*
No	All	no restriction	Varying	8	0.954	0.952	0.968	0.967
No	All	All 8	53	8	0.962	0.961	0.97	0.969
No	Popular Units	All 5 Popular	97	5	0.98	0.98	0.98	0.98
is 0	All	no restriction	130	8	0.875	0.874	0.882	0.881
is 0	All	At least 1	115	8	0.662	0.645	0.694	0.682
is 0	All	At least 5	103	8	0.654	0.647	0.697	0.693
is 0	All	All 5 Popular	97	8	0.712	0.704	0.757	0.753
is 0	Popular Units	At least 1	115	5	0.741	0.705	0.757	0.728
type i	All	At least 5	115	8	0.952	0.951	0.957	0.956
type j	All	At least 5	115	8	0.963	0.962	0.968	0.967

Given the definition of the two coefficients, we expected to find smaller values for alpha than for lambda. The main differences were found in the type of imputation used. These were not large however, with the largest differences being found in the case of imputation 0. Results for imputation 0 varied over the types of candidates included, and dropped below .70 when the 15 candidates without assessor decisions recorded were excluded. However, it may be that giving a unit score=0 in case of missing values may not fully explain the nature of the missing data since this does not necessarily relate to candidate's ability as a measure of the observed units. Candidates 'fail' to achieve a unit due to circumstances, eg because they may not have the opportunity to prove competence rather than because they fail to achieve the standard. Therefore, the results for imputation 0 are not the main results for this study of inter-item relations. Reliability estimated with varying numbers of candidates yielded the almost the same result for alpha in the case of imputation i, whereas imputation j yielded the same results for lambda-2.

Internal consistency estimates for hairdressing qualifications (3008 and 3014)
The qualification design for 3008 and 3014 allows candidates to take different groups of units, but with none of the candidates taking all of the units. The options here would be to join groups of units (based on the qualification structures) or only consider types of units (eg mandatory units). These combinations are still too scattered for the estimation of internal consistency. In the variance covariance matrix, almost 80% of the covariance cells are empty. Joining units, removing units (optional units so only mandatory units are evaluated) or both strategies may improve these numbers. The results would however be very limited, both in content as in the number of candidates on which they are based; therefore we did not carry out this analysis.

7 Discussion

The main aim of this study was to formulate suitable procedures for estimating the reliability of work-based qualifications. Our results suggest that inter- assessor (rater) agreement is 'high' (Gower coefficient ranging from .90 to .99) and inter-rater (assessor/IV) reliability (Cohen's kappa) is 'substantial' for the Electrotechnical pathway (kappa > .75) and 'almost perfect' for the hairdressing qualifications (kappa > .95). Slightly lower values for Cohen's kappa were found when only 'pass without comments' is considered agreement (the strict rule). This may be explained by the fact that for this qualification the internal verifier reviews the portfolio evidence at a later time, when its content may have changed, eg with more evidence added following feedback from an assessor.

Furthermore, a number of features of vocational assessment may influence the magnitude of kappa. For instance, the high prevalence of candidates classified as 'competent' or the requirement that ratings are independent may overestimate the value of kappa. Since candidates are assessed by the same person on more than one occasion, each package of evidence is different from candidate to candidate or because the IV checks the assessor's judgements may mean that the assessor/IV judgements are not necessarily independent. Where assessors and IVs may have given their ratings on different products, this can yield to an underestimation of kappa. Such limitations are all related to the fact that we are working with real life data. In addition, the fact that the marginals are extreme and that the kappa corrects for chance does not inflate the value of kappa, but rather makes is slightly more unstable and smaller.

The data available from the Electrotechnical Services pathway allowed us to estimate the internal consistency of decisions by estimating a coefficient similar to Cronbach's alpha by means of considering units as items. In this case, reliability estimates had values larger than .95, considered to be very high, especially in the context presented by vocational assessment. The correlation between the units, assuming that they are all measuring to a large extent the same ability, is not dissimilar from the concept of test-retest reliability. The fact that the correlations between the different evidence types for the different units are high suggests that the correlation of decisions on the same unit (that is test-retest) is probably at least equally high if not higher. In the case of the Hairdressing qualifications, due to their design which allows for the availability of a large number of pathways, estimating the reliability based on inter-item relations could only provide limited results and therefore this study was not executed.

These results suggest that the reliability of these qualifications is likely to be very good when compared to values reported for other qualification types such as general qualifications in the UK. In the case of vocational qualifications, including those based on the NVQ/occupational competence model studied here, where no previous results have yet been reported, there is no comparative data (although see Murphy et al, 1995). Depending on how we deal with missing data, smaller values were found in the case of imputations where missing scores were considered as 'failures'. A missing score however cannot necessarily be attributed to a failed task, since in the context of vocational assessment there may either be multiple attempts to pass a unit, or a candidate may not have had the opportunity to gather the required sufficient evidence for the units to be

signed off, ie the unit/portfolio is not yet completed. Also, units are not necessarily independent and when criteria are met in one unit this may carry over to other units, although the paper work may not reflect this.

7.1 Implications of our findings

Limited access to assessment data and other logistical issues have so far restricted advances in educational measurement theory of assessment decisions for work-based vocational qualifications. This is a first study to provide a workable strategy for investigating the types of data to be found in practice, ways to collect assessment data and finally suggest simple procedures for estimating the reliability of these qualifications in the form of assessor agreement as well as internal consistency estimates. The procedures used here suggest that standard test theory can be extended and reinterpreted to address problems in the assessments of skills and knowledge acquisition of the type used in vocational education. Due to a lack of standardisation of workplace tasks however, since each candidate had their own set of indicators, some assumptions had to be made for the coefficients to be estimated.

While a generalizability study may seem a sensible way to further investigate the reliability of these qualifications (eg Johnson & Johnson, 2009), the data to be found in practice makes it unfeasible to perform such a study since the real life data collection makes it difficult to differentiate between the various sources of measurement error. However the measurement error, or better the lack of measurement error, may be considered as a measure of generalizability. Therefore Cronbach's alpha may be considered as an unbiased estimate of the generalizability (Brennan, 2001). Because we do not need the assumption of parallel tests or units for this to be true, the internal consistency estimates found here may give an indication of the results from a generalizability study. This means that even in our case where we have a heterogeneous set of indicators for the units (unit scores consist of different types of evidence), Cronbach's alpha can be viewed as a measure of how well the sum of units capture the expected score in the entire domain. Consequently, such a study would not yield fundamentally different results from the procedures we presented here and therefore we do not consider this to be useful in this context.

Although a certain amount of variation can be expected in any assessment system, in contrast to the inherent threats to the reliability of decisions in vocational assessment described here, it may be that taken together, the assessor judgements can lead to a correct classification that meets the purposes of a qualification. The more measurements are available, the higher the reliability of the decision based on the measurements (eg Traub & Rowley, 1991). In addition, the fact that candidates are signed off only when deemed competent, with multiple retakes allowed, this may reduce the measurement error, resulting in strong reliability estimates as was found to be the case in this study (Rudner, 2001; Clauser, Margolis and Case, 2006).

The high stakes nature of the assessment decisions involved in these qualifications means that the implications of a wrong decision can have serious consequences for the individuals, employers, the general public or other stakeholders. By implication, those responsible for the quality of the assessments and qualifications should ensure that the right decisions are made. Awarding organisations do this through a quality assurance system which, on the basis of the results presented here, appears to function effectively.

7.2 Recommendations

Specific recommendations cannot be made for each of the challenges encountered, since for example certain characteristics are dictated by the nature of the assessment system that uses observation by an assessor of naturally occurring work tasks as the main type of evidence. In addition, the complex designs of some of these qualifications are necessary in order to meet diverse industry requirements, while the decision rules are a requirement of the qualifications framework. Although the challenges in the data collection were significant, some of the assessment records available can be used to estimate the reliability of certification decisions. The following suggestions can be made however to support further reliability work:

1. Reliability studies in this area should take into account these issues and aim as much as possible to assess reliability across a wide range of qualifications, evidence types and using complete portfolios for a large sample of candidates. This would minimise the effects from some of the system's design features.

2. Further work should be carried out with those involved in the assessment of vocational qualifications, including centre staff, assessors and verifiers to broaden our conceptualisation of quality in assessment that could also include estimates of reliability of the type described here. Measures of agreement could be used for example in standardisation training. In this case, accuracy or agreement may overestimate the reliability of decisions taken during live assessments, but this could be overcome by carefully controlling for the materials used in training.

3. In addition, because candidates are assessed 'when ready', the possibility exists for assessors to be more inclined to consider the person as competent (since the expectation is that the candidate is so) (eg Kazdin, 1977). Frequent training of assessors and verifiers and their independent status are strategies which are likely to mitigate such risks to the quality of the decisions. Effective development and use of computer technology for the assessment of vocational competence and standardisation of decisions may have the potential to contribute to consistent judgements (eg Kratochwill, Doll & Dickson, 1985 in the area of behavioural assessment).

4. Since inter-rater (assessor/IV) agreement was found to be stronger when the assessor and internal verifier were believed to assess the same evidence, this leads us to conclude that if we would like to know the true inter-rater-reliability we should make sure that the product that is reviewed should not be altered. This however may not be possible due to characteristics of the assessment system whereby the assessor as well as the internal verifier provide feedback to the candidate and/or assessor for further improvement.

5. The adoption of electronic (e-)portfolios is likely to enable practitioners to use such a system to store assessment records in a standardised format across qualifications and centres. Used for the purpose of assessing candidates and recording feedback, e-portfolios can enable access to standardised assessment data that could benefit reliability studies that use procedures such as those presented here (eg see Ridgway, McCusker & Pead, 2004; Rees & Sheard, 2004).

7.3 Further research

We accept that this study focused on one type of vocational qualification, namely work-based qualifications. Recent changes in the design of vocational qualifications, which blur the differences between NVQs and vocationally related qualifications (VRQs), suggest that other vocational qualifications with a purpose different from the work-based qualifications presented here should also be investigated. Such qualifications may use different assessment types (including assignments, projects, case studies, traditional examinations) and may involve extended instruction and training.

Furthermore, we do not underestimate the importance of other quality principles that are essential in this context. Validity is such a principle that merits particular attention in the context of vocational assessment and criteria for evaluating the appropriateness of these interpretations should also be developed (Wools, Eggen & Sanders, 2010). For the types of assessments described here, reliability and validity are harder to differentiate than in other cases such as classifications resulting from applying a cutoff point on a continuous score. Where scores are available, a reliability study would investigate correct classifications and misclassifications (eg Lee, Hansel & Brennan, 2002; Haertel, 2006). In the case of vocational assessments where the evidence is unique and varied, it means that it is also more difficult to define the 'true score' without crossing the boundary between definitions of validity and reliability. The main issue in such a study is whether the person is good enough to do the job in real life and so percentages of misclassifications are more difficult to provide in the traditional sense of reliability research. By extension, in our view a reliability study should resemble a validation study.

Following a conceptualisation of reliability that consists of both social and scientific values, Parkes (2007) follows the concepts developed by Kane (2006) of an argument-based validation to suggest a set of reliability arguments. The model includes classical reliability measures, but extends the analysis to the values associated with the scores that include:

- A determination of the social and scientific values of dependability.
- Consistency.
- Accuracy.
- The purpose and the context of assessment.
- What is reproducibility in the context.
- Investigating the evidence.
- Constructing an argument for or against the inferences made.

In this view, replication is not about '[…] pointing to the eight group meetings during the project period as "replications". This is where contextual factors and theoretical considerations become critical' (Parkes, 2007, p5). The interpretative argument proposed by Parkes proposes a strategy for using multiple inferences underlying score

interpretation and use that broadens the conceptual underpinnings of reliability practice that can develop into additional methods and methodologies.

Because this study was based on real life data, we were not able to evaluate the misclassifications of the whole certification procedure. In addition, since there is no limit on the time allowed to accumulate evidence or number of re-takes, the number of candidates who should pass but fail (false negative) may increase, and at the same time the number of candidates who should fail but pass (false positive) becomes smaller. This reinforces our conclusion that a study on validity is so important for vocational assessment. In this case, the number of tries should be taken into account. More broadly, research should further focus on the mechanisms involved in such assessment decisions, the factors that may influence the interpretations made based on the evidence presented to human assessors such as the quality of the evidence or the relationship between assessor and candidate (see Greatorex 2002, 2005). It could be that advances in other fields such as neuropsychology that investigates the neural processes of human decision making may also inform such an enquiry (eg De Martino et al, 2006, 2008).

Finally, the data contained in this study was drawn entirely from City & Guilds qualifications. In the newly established QCF however, where units may be shared among awarding organisations in the absence of shared assessment practice, verification procedures or a Code of Practice (eg QCA, 2006), it is possible that our findings do not replicate across qualifications designed by other awarding organisations. Further studies are therefore recommended in the future.

Whilst more research is needed to apply core psychometric concepts to the context of vocational assessment, there should be a clear understanding of its intrinsic quality. This study is only a first step in addressing this issue for vocational qualifications.

8 References

AERA/APA/NCME (1999). Standards for Educational & Psychological Testing. American Educational Research Association. Washington, DC;

Berk, RA. (1980). A Framework for Methodological Advances in Criterion-Referenced Testing. Applied Psychological Measurement 4: 563-573;

Brennan, R.L. (2001). Generalizability theory. New York: Springer;

Brookhart SM (2003). Developing Measurement Theory for Classroom assessment purposes and uses. Educational Measurement: Issues and Practice, 22(4), 5-12;

Brown JS, Collins A, Duguid P (1989). Situated Cognition and the Culture of Learning, Educational Researcher 18: 32-42;

Buckendahl CW, Yang Y and Ferdous A (2003). An alternative strategy for estimating decision consistency reliability. University of Nebraska, Lincoln, Retrieved from UUUhttp://www.unl.edu/buros/biaco/pdf/pres03buck02.pdf on 19 October 2010;

Chester MD (2003). Multiple measures and high stakes decisions. A framework for combining measures. Educational Measurement: Issues and Practice, 22(2), NCME;

City & Guilds (2009a). Bringing clarity to the QCF: a guide to the new framework. Retrieved from http://www.cityandguilds.com/documents/Centre%20(Generic)/QCF-booklet-clarity-v2.pdf on 19 October 2010;

City & Guilds (2009b). Levels 1-3 NVQ and SVQ Qualifications in Hairdressing, Barbering and Combined Hair Types (3008/3009) – Assessors' handbook. Retrieved from http://www.cityandguilds.com/documents/ind_hairdressing/3008_Assessors_guide_final(2).pdf on 19 October 2010;

Clauser BE, Margolis MJ & Case SM (2006). Testing for Licensure and Certification in the Professions. In . In RL Brennan (Ed) Educational Measurement, pp701-730;

Cohen J (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, 37-46;

Cohen J (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin, 70, 213-220;

Collins A, Brown JS, & Newman SE (1990). Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. In LB Resnick (Ed.), Knowing, learning, and instruction: Essays in honor of Robert Glaser (pp. 453-494). Hillsdale, NJ: Lawrence Erlbaum;

Crisp V & Novakovic N (2008). Towards a methodology for evaluating the equivalency of demands in vocational assessments between colleges/training providers. A paper presented at the International Association for Educational Assessment Annual Conference, September 2008, Cambridge, UK;

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

Cronbach LJ & Gleser GC (1957). Psychological tests and personnel decisions. Urbana: U Illinois Press;

Cronbach LJ, Linn RL, Brennan RL, Haertel EH (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. Educational and Psychological Measurement, 57(3), 373-399;

De Martino B, Harrison NA, Knafo S, Bird G and Dolan RJ (2008). Explaining Enhanced Logical Consistency during Decision Making in Autism. J. Neurosci 28, 10746-10750;

De Martino B, Kumaran D, Seymour B & Dolan RJ (2006). Frames, Biases, and Rational Decision-Making in the Human Brain. Science 313 (5787), 684;

Douglas, K. M. (2007). General Method for Estimating the Classification Reliability of Complex Decisions Based on Configural Combinations of Multiple Assessment Scores. PhD thesis, University of Maryland, USA;

Driessen EW, Tarwijk JV, Overeem K, Vermunt JD & Van der Vleuten CPM (2005). Conditions for successful use of portfolio for reflection. Medical Education, 39, 1230-1235;

Ebel RL & Frisbie DA (1991). Essentials of Educational Measurement. New Jersey: Prentice-Hall Inc;

Eraut M, Steadman S, Trill J & Parkes J (1996). The Assessment of NVQs. Research Report No 4, University of Sussex: Brighton;

FAB/JCQ (2010). Writing QCF Units: How much detail to provide, Guidance Note 3, Version 1, March;

Good R (2002). Using discriminant analysis as a method of combining multiple measures of student performance. Paper presented at the annual meeting of the AERA, New Orleans, April;

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika, 53, 315-328;

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. Biometrics, 27, 857-871;
Greatorex J & Shannon M (2003). How can NVQ assessors' judgements be standardized? Paper presented at the annual conference of the British Educational Research Association, Edinburgh, September;
Greatorex J (2000). What research can an awarding body carry out about NVQs? A paper presented at the British Research Association Conference, University of Cardiff, September;
Greatorex J (2002). Two heads are better than one: standardizing the judgements of National Vocational Qualification Assessors. A paper presented at the British Educational Research Association Conference, Exeter, September;
Greatorex J (2005). Assessing the evidence: different types of NVQ evidence and their impact on reliability and fairness. J of Vocational Education & Trainingm 57(2), 149-164;
Gwet K (2002). Inter-rater reliability: dependency on trait prevalence and marginal homogeneity. Statistical Methods for Inter-Rater Reliability Assessment, 2, 1-9;
Gwet K (2008). Computing inter-rater reliability and its variance in the presence of high agreement. British Journal of Mathematical and Statistical Psychology, 61, 29-48;
Haertel EH (2006). Reliability. In RL Brennan (Ed) Educational Measurement, pp65-107;
Huynh H (1976). On the reliability of decision in domain-referenced testing. JEM, 13, 253-264;
Huynh H (1978). Reliability of multiple classifications. Psychometrika, 43, 317-325;
Johnson M (2006). A review of vocational research in the UK 2002-2006: Measurement and accessibility issues. International J of Training Research, 4(2), 48-71;
Johnson M (2008a). Assessing at the borderline: Judging a vocationally related portfolio holistically. Isues in Education, 18(1);
Johnson M (2008b). Exploring assessor consistency in a Health and Social Care qualification using a sociocultural perspective. J of Vocational Education & Training, 60(2), 173-187;
Johnson S & Johnson R (2009). Conceptualising and interpreting reliability. Ofqual/10/4706;
Kane MT (2006). Validation. In RL Brennan (Ed) Educational Measurement, pp17-64;
Kazdin AE (1977). Artifact, bias and complexity of assessment: The ABCs of reliability. J of Applied Behav Analysis, 10(1), 141-150;
Kingston P (March 2007). No Accounting for Taste, The Guardian, Retrieved from http://www.guardian.co.uk/education/2007/mar/06/furthereducation.uk2 on 19 October 2010;
Kratochwill TR, Doll EJ & Dickson P (1985). Microcomputers in behavioral assessment: Recent advances and remaining issues. Computers in Human Behav, 1(3-4), 277-291;
Lane S and Stone CA (2006). Performance assessment. In RL Brennan (Ed) Educational Measurement, pp387-430;
Lee W, Hanson BA & Brennan RL (2002). Estimating consistency and accuracy indices for multiple classifications. Applied Psychological Measurement, 26, 412–432;
Lord FM & Novick MR (1968). Statistical theories of mental test scores. Reading MA: Addison-Welsley Publishing Company;
Mitchell L & Bartram D (1994). The place of knowledge and understanding in the development of national vocational qualifications and Scottish vocational qualifications. Moorfoot, Sheffield: Employment Dept;
Mislevy RJ (1994). Can there be reliability without 'reliability'? ETS, Princeton, NJ;
Murphy R, Burke P, Content S, Frearson M, Gillispie J, Hadfield M, Rainbow R, Wallis J & Wilmut J (1995) The Reliability of Assessment of NVQs. Report presented to NCVQ, School of Education, University of Nottingham;
National Database for Accredited Qualifications (NDAQ, 2010). Plait and twist hair using basic techniques. Retrieved from http://www.accreditedqualifications.org.uk/unit/Y6001037.seo.aspx?OwnerRef= on 19 October 2010;
Nichols PD & Smith PL (1998). Contextualising the interpretation of reliability data. Educational Measurement: Issues and Practice, 17(3), 24-36;
Ofqual (2008). Regulatory arrangements for the Qualifications and Credit Framework. Ofqual/08/37/26;
Ofqual (2009). Identifying purposes for qualifications in the Qualifications and Credit Framework. Ofqual/09/3985;
QCDA (2010). Guidelines for Writing Credit-Based Units of Assessment, Version 4, QCDA/10/4725;
Osburn HG (2000). Coefficient alpha and related internal consistency reliability coefficients, Psychological Methods, 2000, 5, 343-355;
Qualifications and Curriculum Authority (2006). NVQ Code of Practice. Retrieved from http://www.ofqual.gov.uk/files/qca-06-2888_nvq_code_of_practice_r06.pdf on 19 October 2010;
Parkes J (2007). Reliability as argument. Educational Measurement: Issues and Practice, 26(4), 2-10;

Rees C & Sheard C (2004). The reliability of assessment criteria for undergraduate medical students' communication skills portfolios: the Nottingham experience. Medical Education, 38(2), 138-44;
Ridgway J, McCusker S and Pead D (2004). Literature review of e-assessment. Futurelab, Bristol;
Rudner L (2001). Computing the expected proportions of misclassified examinees. Practical Assessment, Research & Evaluation, 7(14);
Ryan JM & Hess RK (1999). Issues, strategies and procedures for combining data from multiple measures. Paper presented at an annual meeting of AERA, Montreal;
Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. Psychometrika, 74(1), 107-120;
Smith JK (2003). Reconsidering Reliability in Classroom Assessment and Grading. Educational Measurement: Issues and Practice, 22(4), 26–33;
Subkoviak MJ (1976).Estimating reliability from a single administration of a mastery test. JEM, 13, 265-276;
Swaminathan, H., Hambleton, R.K., & Algina, J. (1974). Reliability of Criterion-Referenced Tests: A Decision-Theoretic Formulation. Journal of Educational Measurement, 11(4), 263-267;
Traub RE & Rawley GL (1980). Reliability of Test Scores and Decisions. Applied Psych Meas 4 (4), 517-545;
Traub, R.E. & Rowley, G.L. (1991). Understanding reliability. Educational Measurement: Issues and Practice, 10(1), 37-45;
UK Commission for Employment (UKCES) & Skills and the Alliance of Sector Skills Councils (SASSC) (2010). National Occupational Standards Quality criteria with Explanatory Notes. Second draft, Retrieved from http://www.ukces.org.uk//upload/pdf/NOS_Quality_Criteria_Second_Draft_040210_1.pdf on 19 October 2010;
Verhelst ND (2004). Classical Test Theory. In Manual for relating Language Examinations to the Common European Framework of Reference for Languages (CEFR), Reference Supplement, Section C, Language Policy Division, Strasbourg;
Wilmut J, Woods R & Murphy R (1996). A Review of Research into the Reliability of Examinations. A discussion paper prepared for the School Curriculum and Assessment Authority, retrieved from http://www.nottingham.ac.uk/shared/shared_cdell/pdf-reports/relexam.pdf on 19 October 2010;
Wolf A (1995). Competence-based Assessment. Open University Press: Buckingham;
Wools S, Eggen T & Sanders P (2010). Evaluation of Validity and Validation by Means of the Argument-based Approach. CADMO, 1, 63-82;
Zegers, F.E. (1991). Coefficients for Interrater Agreement. Applied Psychological Measurement, 15, 321-333.

Acknowledgements

The first author is grateful to Andrew Boyle for discussions and many helpful comments and suggestions on an earlier version of this paper. The authors would also like to thank their reviewers Dr Saskia Wools, Dr Huub Verstralen, members of the Reliability Programme Technical Advisory Group, Inga Fitzgerald, Andrew Stone and Vicky Foley. We would also like to thank the many assessors and candidates who took part in this research study and without whom this would not have been possible. Finally, we are indebted to Ofqual for their financial support and to Dr Qingping He and Joanna Taylor for their support during this project.

Appendix 1: Examples of qualification structures

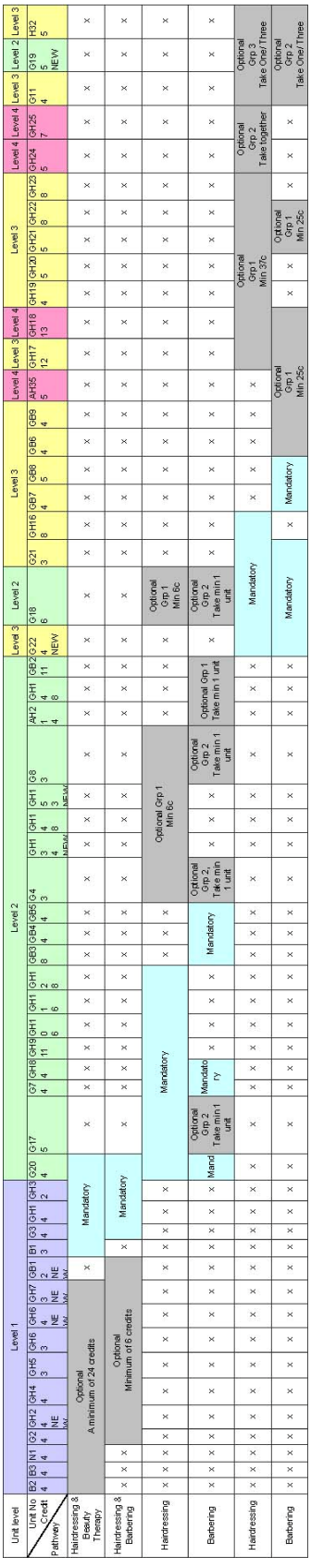


Figure 1 – Example of qualification structure of the Certificate/Diploma Hairdressing (QCF)

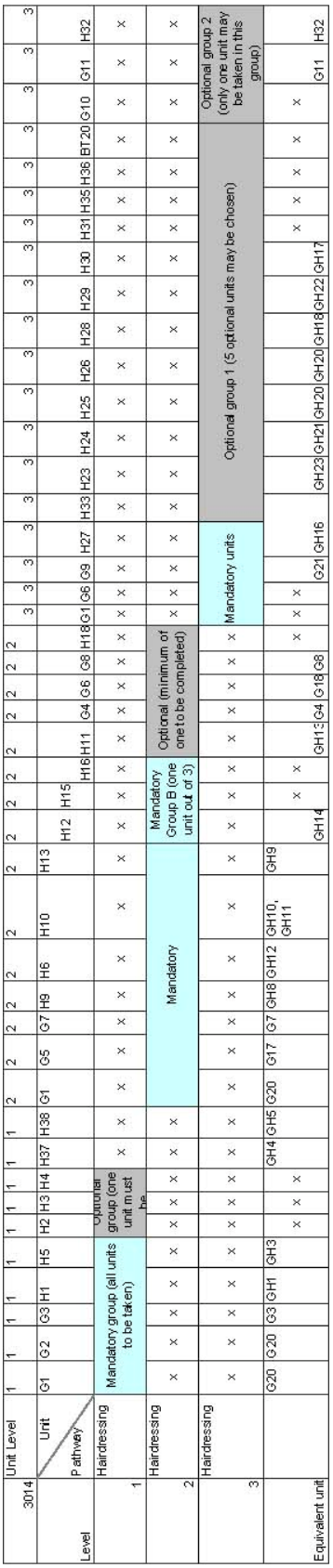
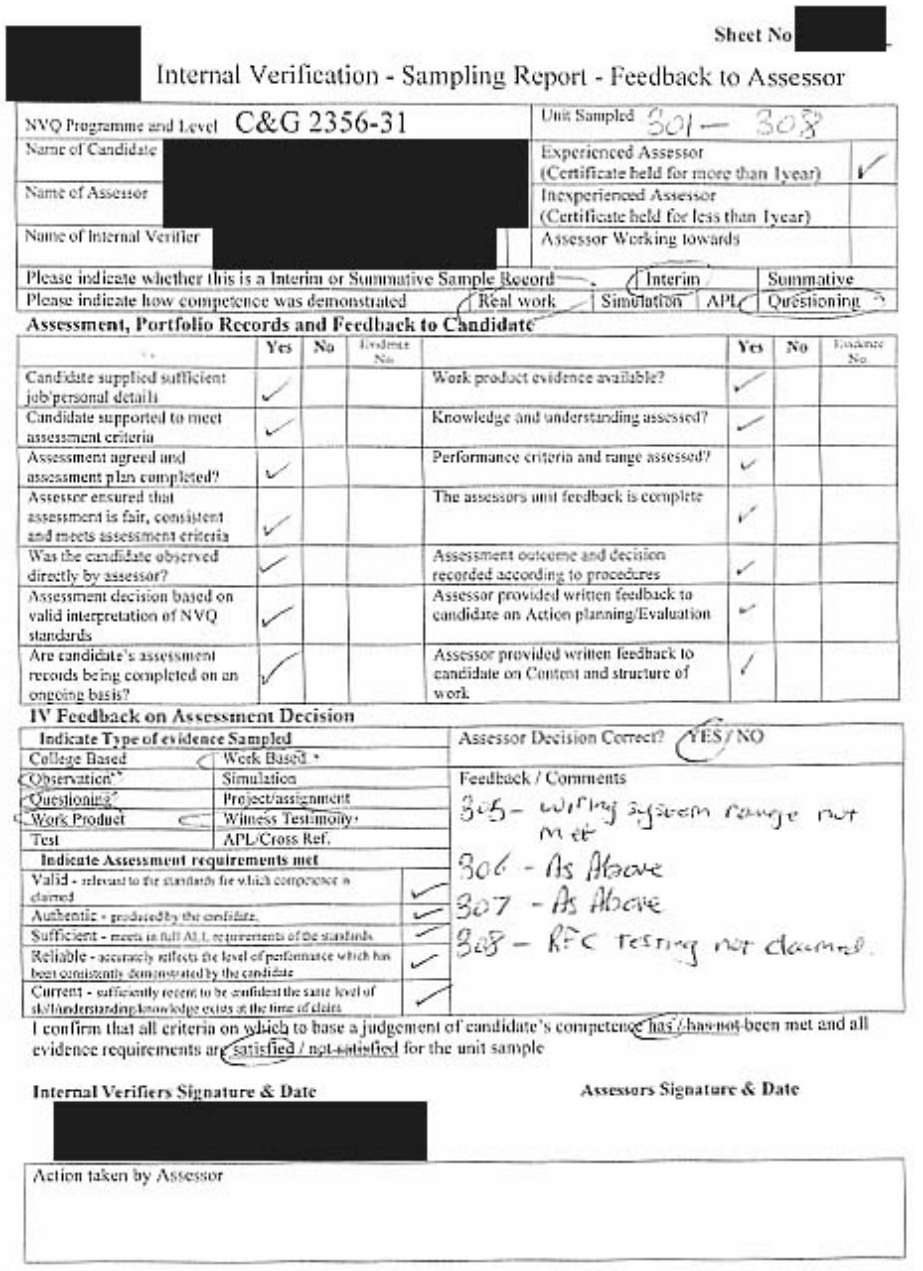


Figure 2 – Example of qualification structure of the NVQ Hairdressing (NQF)

Appendix 2: Examples of assessment records



Sheet No [redacted]

[redacted]

Internal Verification - Sampling Report - Feedback to Assessor

NVQ Programme and Level C&G 2356-31		Unit Sampled 301 – 308	
Name of Candidate [redacted]		Experienced Assessor (Certificate held for more than 1year)	✓
Name of Assessor		Inexperienced Assessor (Certificate held for less than 1year)	
Name of Internal Verifier		Assessor Working towards	
Please indicate whether this is a Interim or Summative Sample Record		Interim	Summative
Please indicate how competence was demonstrated	Real work	Simulation / APL	Questioning

Assessment, Portfolio Records and Feedback to Candidate

	Yes	No	Evidence No		Yes	No	Evidence No
Candidate supplied sufficient job/personal details	✓			Work product evidence available?	✓		
Candidate supported to meet assessment criteria	✓			Knowledge and understanding assessed?	✓		
Assessment agreed and assessment plan completed?	✓			Performance criteria and range assessed?	✓		
Assessor ensured that assessment is fair, consistent and meets assessment criteria	✓			The assessors unit feedback is complete	✓		
Was the candidate observed directly by assessor?	✓			Assessment outcome and decision recorded according to procedures	✓		
Assessment decision based on valid interpretation of NVQ standards	✓			Assessor provided written feedback to candidate on Action planning/Evaluation	✓		
Are candidate's assessment records being completed on an ongoing basis?	✓			Assessor provided written feedback to candidate on Content and structure of work	✓		

IV Feedback on Assessment Decision

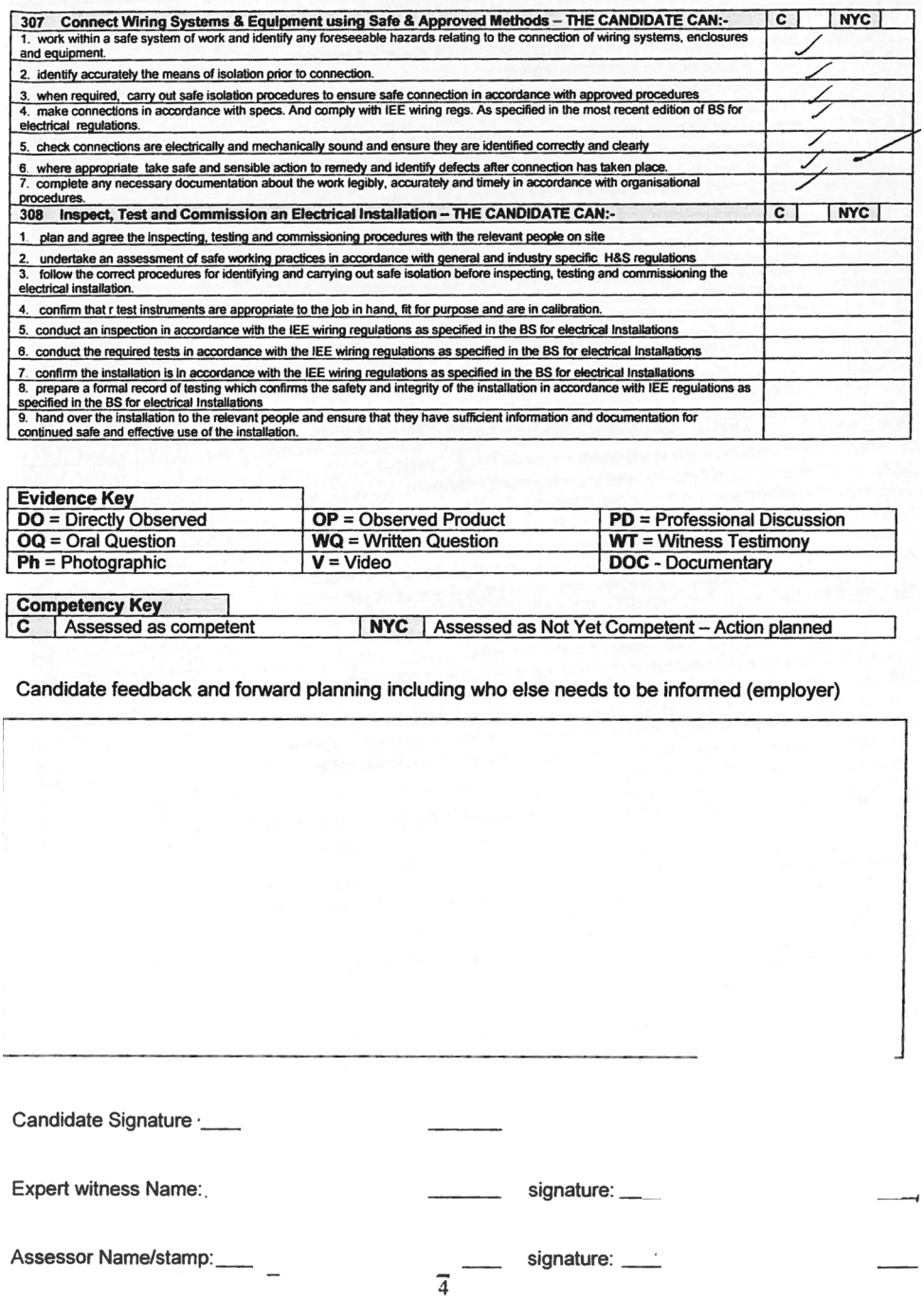
Indicate Type of evidence Sampled		Assessor Decision Correct? YES / NO
College Based	Work Based *	
Observation*	Simulation	Feedback / Comments
Questioning*	Project/assignment	305 – wiring system range not met
Work Product	Witness Testimony*	306 - As Above
Test	APL/Cross Ref.	307 - As Above
Indicate Assessment requirements met		308 – RFC testing not claimed.
Valid - relevant to the standards for which competence is claimed	✓	
Authentic - produced by the candidate.	✓	
Sufficient - meets in full ALL requirements of the standards	✓	
Reliable - accurately reflects the level of performance which has been consistently demonstrated by the candidate	✓	
Current - sufficiently recent to be confident the same level of skill/understanding/knowledge exists at the time of claim	✓	

I confirm that all criteria on which to base a judgement of candidate's competence has / ~~has not~~ been met and all evidence requirements are satisfied / ~~not satisfied~~ for the unit sample

Internal Verifiers Signature & Date [redacted] **Assessors Signature & Date**

Action taken by Assessor

Figure 3 – Example of an Internal verification report (Type 5, Table 6)



307 Connect Wiring Systems & Equipment using Safe & Approved Methods – THE CANDIDATE CAN:-	C	NYC
1. work within a safe system of work and identify any foreseeable hazards relating to the connection of wiring systems, enclosures and equipment.	✓	
2. identify accurately the means of isolation prior to connection.	✓	
3. when required, carry out safe isolation procedures to ensure safe connection in accordance with approved procedures	✓	
4. make connections in accordance with specs. And comply with IEE wiring regs. As specified in the most recent edition of BS for electrical regulations.	✓	
5. check connections are electrically and mechanically sound and ensure they are identified correctly and clearly	✓	
6. where appropriate take safe and sensible action to remedy and identify defects after connection has taken place.	✓	
7. complete any necessary documentation about the work legibly, accurately and timely in accordance with organisational procedures.	✓	
308 Inspect, Test and Commission an Electrical Installation – THE CANDIDATE CAN:-	**C**	**NYC**
1. plan and agree the inspecting, testing and commissioning procedures with the relevant people on site		
2. undertake an assessment of safe working practices in accordance with general and industry specific H&S regulations		
3. follow the correct procedures for identifying and carrying out safe isolation before inspecting, testing and commissioning the electrical installation.		
4. confirm that r test instruments are appropriate to the job in hand, fit for purpose and are in calibration.		
5. conduct an inspection in accordance with the IEE wiring regulations as specified in the BS for electrical Installations		
6. conduct the required tests in accordance with the IEE wiring regulations as specified in the BS for electrical Installations		
7. confirm the installation is in accordance with the IEE wiring regulations as specified in the BS for electrical Installations		
8. prepare a formal record of testing which confirms the safety and integrity of the installation in accordance with IEE regulations as specified in the BS for electrical Installations		
9. hand over the installation to the relevant people and ensure that they have sufficient information and documentation for continued safe and effective use of the installation.		

Evidence Key		
DO = Directly Observed	**OP** = Observed Product	**PD** = Professional Discussion
OQ = Oral Question	**WQ** = Written Question	**WT** = Witness Testimony
Ph = Photographic	**V** = Video	**DOC** - Documentary

Competency Key			
C	Assessed as competent	**NYC**	Assessed as Not Yet Competent – Action planned

Candidate feedback and forward planning including who else needs to be informed (employer)

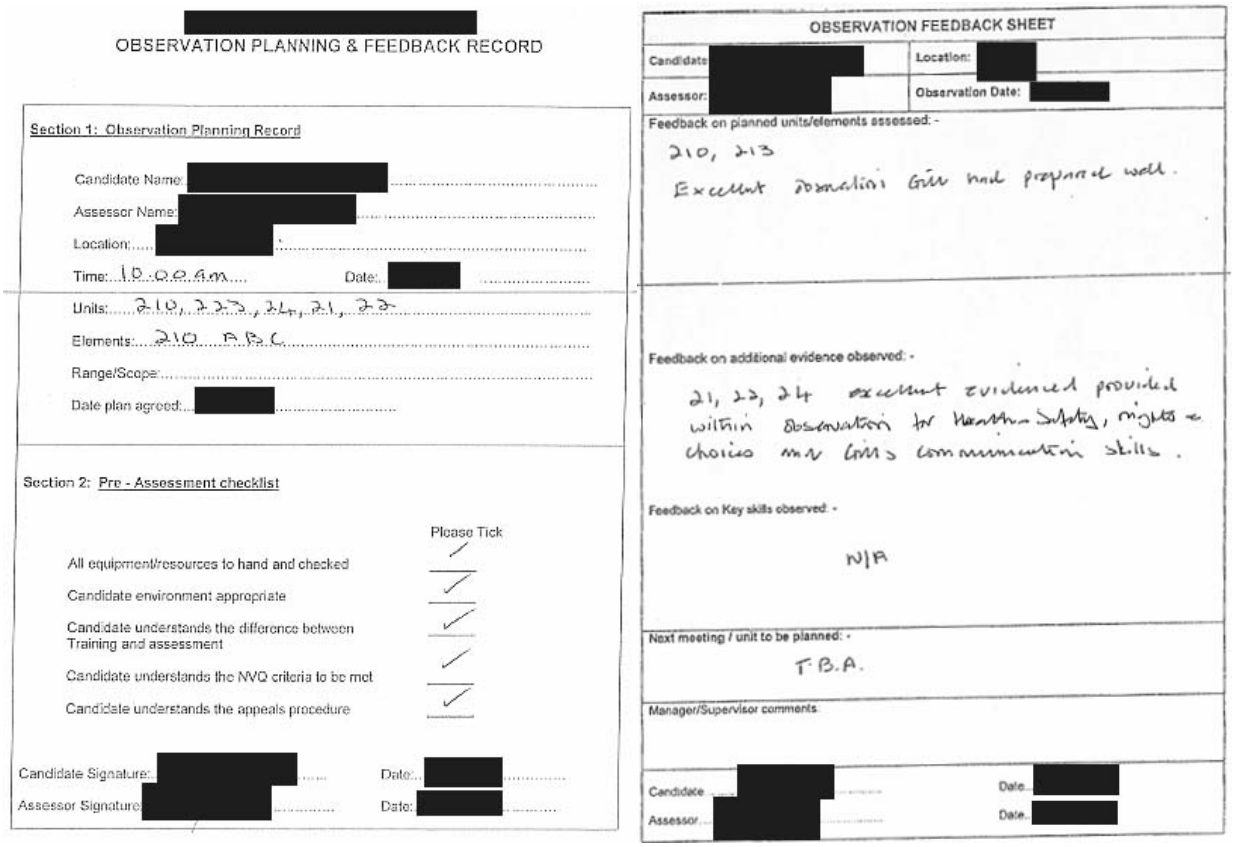
Candidate Signature

Expert witness Name: signature:

Assessor Name/stamp: signature:

4

Figure 4 – Example of observation record (Type 1, Table 6) where candidates are judged as competent (C) or not yet competent (NYC) against performance criteria



[redacted]

OBSERVATION PLANNING & FEEDBACK RECORD

Section 1: Observation Planning Record

Candidate Name: [redacted]

Assessor Name: [redacted]

Location: [redacted]

Time: 10.00am Date: [redacted]

Units: 210, 213, 21, 21, 23

Elements: 210 P1 P2 C

Range/Scope:

Date plan agreed: [redacted]

Section 2: Pre - Assessment checklist

	Please Tick ✓
All equipment/resources to hand and checked	✓
Candidate environment appropriate	✓
Candidate understands the difference between Training and assessment	✓
Candidate understands the NVQ criteria to be met	✓
Candidate understands the appeals procedure	✓

Candidate Signature: [redacted] Date: [redacted]

Assessor Signature: [redacted] Date: [redacted]

OBSERVATION FEEDBACK SHEET

Candidate: [redacted] Location: [redacted]

Assessor: [redacted] Observation Date: [redacted]

Feedback on planned units/elements assessed: -

210, 213

Excellent discussion. Girl well prepared well.

Feedback on additional evidence observed: -

21, 22, 24 excellent evidenced provided within observation for Health & Safety, rights & choice and Core Communication skills.

Feedback on Key skills observed: -

N/A

Next meeting / unit to be planned: -

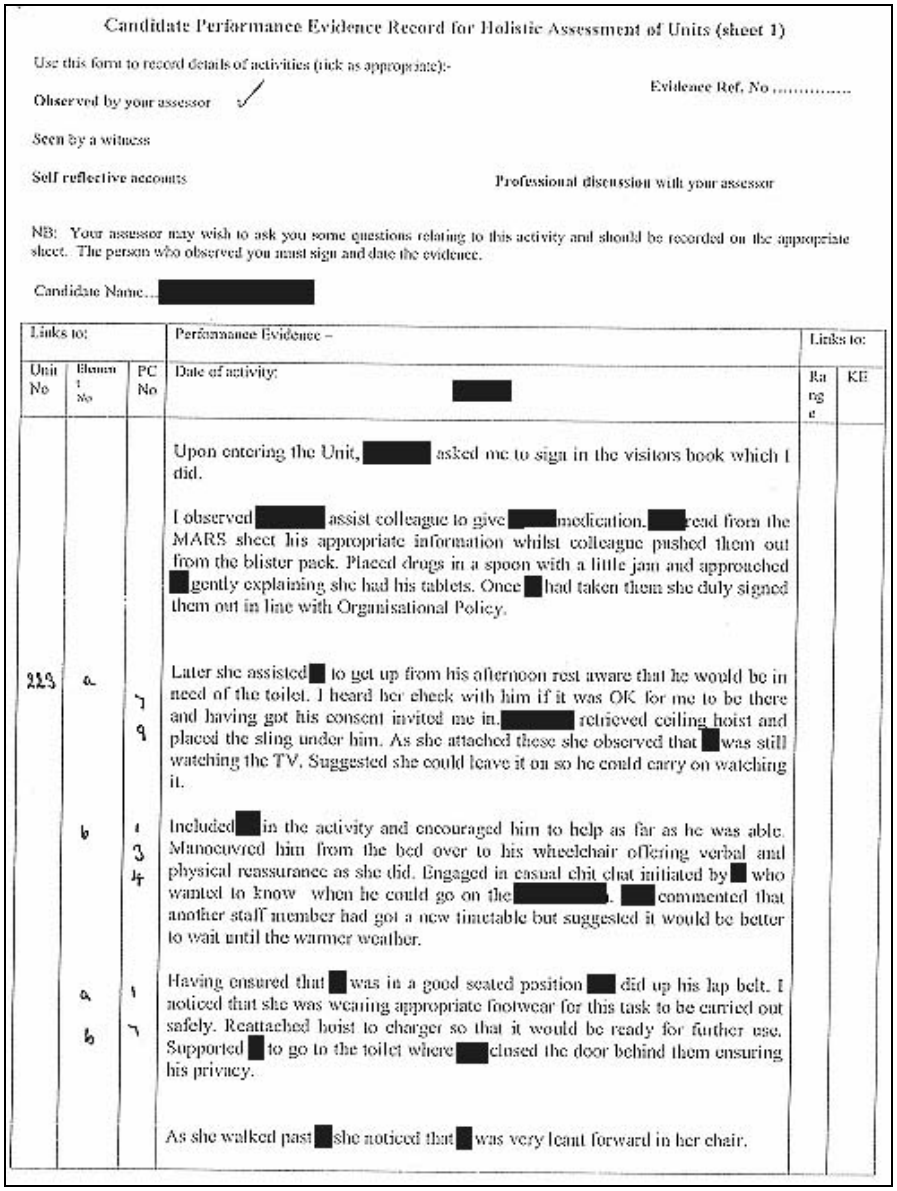
T.B.A.

Manager/Supervisor comments

Candidate [redacted] Date: [redacted]

Assessor [redacted] Date: [redacted]

Figure 5 a and b – Example of a observation record that provides qualitative feedback to the candidate (Type 1 record, see Table 6)



Candidate Performance Evidence Record for Holistic Assessment of Units (sheet 1)

Use this form to record details of activities (tick as appropriate):-

Observed by your assessor ✓ Evidence Ref. No .....

Seen by a witness

Self reflective accounts Professional discussion with your assessor

NB: Your assessor may wish to ask you some questions relating to this activity and should be recorded on the appropriate sheet. The person who observed you must sign and date the evidence.

Candidate Name...[redacted]

Links to:			Performance Evidence – Date of activity: [redacted]	Links to:	
Unit No	Element No	PC No		Range	KE
			Upon entering the Unit, [redacted] asked me to sign in the visitors book which I did.		
			I observed [redacted] assist colleague to give [redacted] medication. [redacted] read from the MARS sheet his appropriate information whilst colleague pushed them out from the blister pack. Placed drugs in a spoon with a little jam and approached [redacted] gently explaining she had his tablets. Once [redacted] had taken them she duly signed them out in line with Organisational Policy.		
223	a	1 9	Later she assisted [redacted] to get up from his afternoon rest aware that he would be in need of the toilet. I heard her check with him if it was OK for me to be there and having got his consent invited me in. [redacted] retrieved ceiling hoist and placed the sling under him. As she attached these she observed that [redacted] was still watching the TV. Suggested she could leave it on so he could carry on watching it.		
	b	1 3 4	Included [redacted] in the activity and encouraged him to help as far as he was able. Manoeuvred him from the bed over to his wheelchair offering verbal and physical reassurance as she did. Engaged in casual chit chat initiated by [redacted] who wanted to know when he could go on the [redacted]. [redacted] commented that another staff member had got a new timetable but suggested it would be better to wait until the warmer weather.		
	a b	1 7	Having ensured that [redacted] was in a good seated position [redacted] did up his lap belt. I noticed that she was wearing appropriate footwear for this task to be carried out safely. Reattached hoist to charger so that it would be ready for further use. Supported [redacted] to go to the toilet where [redacted] closed the door behind them ensuring his privacy.		
			As she walked past [redacted] she noticed that [redacted] was very leant forward in her chair.		

Figure 6 – Example of a observation record (Type 3, Table 6)

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 2011.

Office of Qualifications and Examinations Regulation
Spring Place
Herald Avenue
Coventry Business Park
Coventry
CV5 6UB