RELIABILITY ISSUES IN COMPETENCE-BASED ASSESSMENT: CONCEPTS AND ESTIMATES

A literature review prepared by City & Guilds as part of the Office of the Qualifications and Examinations Regulator's Reliability programme by

HELEN HARTH
City & Guilds, UK

Dr PETER VAN RIJN
Cito, The Netherlands

April 2010

CONTACT ADDRESS
Helen Harth, City & Guilds, 1 Giltspur St, London, EC1A 9DD, United Kingdom
Email helen.harth@cityandguilds.com
Phone 44 (0)20 7294 8153; Fax 44 (0)20 7294 2416

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

'THE ESTIMATES OF RELIABILITY OF VOCATIONAL ASSESSMENT' RESEARCH STUDY

Programme name and number	Ofqual Reliability Programme Contract number: 2905
Lead Institution	City & Guilds 1 Giltspur St EC1A 9DD London UK
Project Leader	Mrs Helen Harth helen.harth@cityandguilds.com Tel: 020 7294 8153
Collaborating Institution	Cito Nieuwe Oeverstraat 50 6811JB Arnhem The Netherlands
Project Leader	Dr Peter van Rijn peter.vanrijn@cito.nl
Length of project	10 months
Total cost to Ofqual over its life	£41,960
Proposed start and end dates	4 January to 25 October 2010

Document	
Document title	Reliability issues in competence-based assessment: A literature review
Date created	8 March 2010
Author and project role	Helen Harth, Project manager

Document history		
Version	Date	Comments
1	12 March 2010	First draft for submission to Cito's project team
2	22 March 2010	Second draft for submission to peer reviewers
3	31 March 2010	Third draft for submission to Ofqual's project team

SUMMARY

'The Estimates of Reliability of Vocational Assessment' project is part of the Office of the Qualifications and Examinations Regulator's (QfQual) Reliability Programme. The present research study will be looking at the consistency of assessment decisions and factors which may affect the reliability of results of competence based vocational qualifications in England. OfQual has commissioned this study to City & Guilds and it will be carried out in collaboration with Cito's psychometric research centre. The project runs from January to October 2010.

This comprehensive literature review investigates relevant literature on reliability issues in vocational assessment with respect to the

- individual assessment components of criterion referenced qualifications. These may include a number of assessment types such as observation of performance in a workplace or in simulated conditions, logbooks, professional discussions, projects, examinations, with a variety of item types (eg tasks, multiple choice or open answer) and delivery modes (paper or on screen). These assessments may include internally set and internally marked, externally set and internally marked or externally set and externally marked assessments.
- overall qualification result. This includes the combination of these different assessment types and the decision rule to determine the candidate's final result.

With regard to the background of competence based qualifications used in England, this paper will thus review the research literature into

- Reliability of assessor decisions from criterion referenced competence based assessments such as performance or portfolio assessment
- The methods appropriate to quantify reliability by means of classification accuracy and consistency for pass-fail decisions in criterion referenced assessments

The implications for competence based vocational qualifications in the Qualifications and Credit Frameworks (QCF) are discussed. Some conclusions and recommendations which are supported by the results of this review are made.

The content in this literature review will be used in the final report due for submission to OfQual on the 25 October 2010.

1 INTRODUCTION

The Regulatory Arrangements of the Qualifications and Credit Framework (QCF) is the framework for recognising and accrediting vocational qualifications in England, Wales and Northern Ireland. It states that assessments in vocational qualifications are required to (Ofqual, 2008, paragraph 5.3, p26)

- 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

Although in this approach validity and reliability of the learners' results are important in ensuring confidence in the qualification system, reliability (and validity) aspects have been less investigated (Eraut, Steadman, Trill & Parkes, 1996) than for other types of qualifications (eg general qualifications). However, a vast body of research on the consistency (ie reliability) and the accuracy (ie validity) of criterion-referenced assessment decisions (either assessor judgements or based on a cut score) has been published over the past 50 years, which can expand our understanding of measurement theories that are suited for work-based assessments.

The present review focuses on the methods available for estimating the consistency and accuracy of criterion-referenced assessment decisions that are used for deciding a person's mastery status in competence-based vocational qualifications. The purpose of this review is to provide an overview on the issues and research related to the reliability measures suitable for the assessment types used in these qualifications and provide a framework for expressing the reliability of these types of scores.

The paper is divided into four sections. The first section describes the context, purposes, uses and types of vocational competence-based assessment in the UK. On this background, the reliability of assessor decisions is discussed in the second section, including the link between reliability and validity, whether reliability is important in this context and the factors that contribute to inconsistency. The third section reviews the methods available for investigating the reliability of decisions and their applications to real or simulated data. Strategies available for vocational assessments are summarised. The final section concludes the review with a summary of the challenges and opportunities available for discussing reliability of vocational qualifications.

1.1 Description of competence based assessment

This review starts with a description of competence based assessment of the types used in the United Kingdom (UK) which encompasses the main features of these qualifications in order to support the discussion on the measurement theories suitable in these circumstances. A number of sources on competence-based qualifications in the UK were reviewed (Wolf, 1995, 1998; Eraut et al, 1996; Wilmut, Woods & Murphy, 1996; Ofqual, 2008; QCDA, 2008).

1.1.1 Competence based qualifications

Since their inception over 20 years ago, competence-based National Vocational Qualifications (NVQs) have been used primarily for employment purposes. That is, for confirmation of occupational competence, licence to practice, monitoring learner progression (especially important for funding purposes), providing feedback to candidates for future improvement, and evaluating the effectiveness of assessor performance. People in the workplace or other settings that replicate a working environment normally take up these qualifications. More recently, a small number of these qualifications have been accepted for progression into higher education (Kingston, 2007). The decisions are dependent on the purposes associated with competence-based qualifications and are normally high-stakes, regardless of the assessment design.

These qualifications are outcomes focused, outlining what needs to be achieved but with no prescribed learning programmes. They offer training in vocational areas such as construction, engineering, service industries, health and social care, business administration and management. They are regulated by the Office for Qualifications and Examinations (Ofqual) in England through a set of regulatory criteria, which are designed to provide a legislative base that protects the rights of learners, and enforce obligations of stakeholders (Ofqual, 2008). One such obligation is the provision of quality in assessment. The competence-based assessment development process consists of specification of standards, specification of opportunities to collect sufficient evidence, assessor judgements, learner feedback and quality assurance.

The competence-based approach is based on national occupational standards (NOS) which 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 (QCDA, 2008, p54; UK Commission for Employment (UKCES) & Skills and the Alliance of Sector Skills Councils (SASSC), 2010). In some sectors, demonstration of competence against NOS is required in order to carry out a job (eg run a business) or practice a craft or profession (UKCES & SASSC, 2010; see SEMPTA, 2010 for a full list of NOS purposes). However, the link between NOS and units is an indirect one, because criteria need to be 'demonstrable, observable and measurable' so that their achievement can be assessed (Ofqual, 2008, section 3.2; also see QCA, 2009). In the NVQ Code of Practice[1] for competence-based qualifications, competence is about persons who possess 'the ability to carry out activities to the standards required' (NVQ Code of Practice, 2006, p37). A similar meaning of competence has been conveyed in the QFC unit writing guidelines[2], where units of assessment are linked to NOS to 'focus on the knowledge, skills and understanding, which, applied together, form the competence required by employers for certain roles and functions' (QCDA, 2008, p11; see also Wolf, 1995, p30ff for a discussion).

[1] The NVQ Code of Practice was developed for qualifications on the National Qualifications Framework but no longer applies to qualifications in the QCF (see UKCES, 2008);

[2] Note however that the term competence is not explicitly mentioned in the current version of the regulatory criteria for the QCF (QCDA, 2008; see also Mitchell, 1989; Mulder, Weigel & Collins, 2007)

QCF units are made up of learning outcomes and associated assessment criteria. Outcomes set a clear standard (Ofqual, 2008, paragraph 1.4d) and taken together they describe the occupational skills, knowledge and understanding (or competence) that a candidate who has credit for the unit should posses. Outcomes are equally weighted in terms of achieving a unit (Ofqual, 2008). The associated assessment criteria specify the standard of performance a learner is expected to meet to demonstrate mastery or achievement of the learning outcome (Ofqual, 2008, paragraph 1.5a). The standard is meant to be expressed through examples of range of achievement[3] reflected in the assessment criteria, which define the breadth and depth or scope of a learning outcome and its assessment criteria, describing the circumstances or context, using a combination of methods or at increased levels of responsibility in which competence can be demonstrated (QCDA, 2008). Figure 1 shows the main components of a QCF qualification.

Figure 1: The structure of qualifications on the QCF (QCDA, 2009)

Criteria developed in the previous National Qualifications Framework (NQF), in which range statements were specified separately, have been criticised for being too general in the past, and so leading to local interpretations regarding the required standard of performance (Wolf, 1995, 1998). In its current format, there may be the danger that there will be an even wider variance in the way different learning providers and awarding organisations interpret the learning and assessment requirements (Johnson, 2008a; FAB/JCQ, 2010).

[3] Although the full range or scope can be expressed in the additional information about the unit (QCDA, 2008), it is not clear in the current version of the Regulatory Requirements (Ofqual, 2008) whether these are specified or required to support assessment, or whether the full range should be included in the assessment criteria (see FAB/JCQ, 2010). It is then possible for the range (scope/ evidence requirements) within units, which are generally written by sector skills councils or bodies, to vary in the way it is specified, either within the assessment criteria or within the additional assessment requirements (see Table 1 for an example from hairdressing).

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: […] 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.

Table 1 displays an example of assessment criteria used in a level 1 unit from hairdressing. The context and range is specified in the assessment criteria which should ensure that users of the unit know what is expected of the learner to achieve the learning outcome at the level of the unit, eg at level 1 the completion of routine tasks and procedures (Ofqual, 2008). In addition, learning outcomes and assessment criteria may cover diverse sub-domains, such as skill, knowledge, understanding or behavioural requirements. While the candidate is expected to be able to 'part the sections cleanly' in 2.3, they also need to 'confirm the client's satisfaction with the finished look' in 2.11.

The performance criteria taken together distinguish between satisfactory and unsatisfactory performance in the function covered by the NOS (UKCES & SASSC, 2010). Based on the unit content (learning outcomes, performance criteria which subsume range) assessors and internal verifiers (IVs) (see 1.3) use the evidence provided over repeated occasions to decide whether for a particular assessment criteria they have the confidence that the candidate is

- Competent: the evidence generated meets the assessment requirements
- Not yet competent: the candidate has not yet achieved all of the assessment requirements, either based on sufficient evidence or due to insufficiency of evidence where for example the candidate does not have enough opportunities to perform the tasks

Taken together, the criteria represent the final result for the unit as well as for the qualification. In the UK, due to their uses and purposes, competence based qualifications are not generally graded[4]. Qualifications are at different levels and comprise a number of units with varying credit values which follow a number of allowed pathways (routes), depending on the rules of combination. Table 1 shows an example of a competence-based unit.

1.2 Types and sources of evidence

In the outgoing NQF, assessment is based on a candidate's demonstrated performance of the learning outcomes (Wolf, 1995) as specified in the NVQ Code of Practice (2006). These criterion-referenced assessments are assumed to consist of a number of naturally occurring tasks or procedures as part of a person's usual work activities or simulations of naturally-occurring activities. Achievement of individual criteria is then associated with entire learning outcomes which had to be satisfied in order for the assessor to be able to decide whether sufficient evidence has been accumulated to the agreed standard. In the NQF the emphasis on performance assessment and portfolios of evidence is justified by their authenticity, realism and instructional relevance (see UKCES, 2008, Reckase, 1995). By contrast, while the assessment of QCF competence based units is against outcomes and assessment criteria, the assessment criteria do not include any explicit references to methods or instruments of assessment to be used. Units should be capable of assessment independent of other units. The assessment methodology is therefore not prescribed (Ofqual, 2008, paragraph 1.31), although it is expected to be suitable to the type of achievement and purpose of the qualification (QCDA, 2008, p5-7). It is currently not clear how the assessment of QCF units that allow increased flexibility to users to innovate will differ from the former NVQ Code of Practice.

The evidence matched against specified standards can originate from a number of sources that suit increasingly diverse circumstances, which can include a combination of assessment techniques, for example, observation of actual products of performance or behaviours in a workplace or in simulated conditions, diaries, professional discussions, projects, assignments, on screen simulations and also examinations. Multiple choice questions, essays, and oral examinations could be used to test factual recall and applied knowledge. Table 2 below describes the main assessment methods used in this context.

[4] In a survey of the National Database of Accredited qualifications database (NDAQ) carried out in March 2010, of the 380 competence-based qualifications sampled none were graded beyond a pass.

Table 2: Assessment methods used in competence based units

Assessment activity	Description	Assessor's role
Performance assessment	• Practical observation of work-based performance, which may also be simulated in certain circumstances • The assessors make use of assessment opportunities as they occur naturally (see for example City & Guilds, 2009) and may ask additional questions to amplify the evidence provided. • There is also the potential that one observation may yield evidence for different performance criteria from one or several units.	• Direct judgement of the candidate's performance in terms of typical or minimally expected performance at a particular level of competence over a period of time (see Lane & Stone, 2006 for a review). • Assessor feedback is a key outcome of the assessment process.
Inspection of products	• Final product, such as an object created/ repaired, work diary, photographs of completed work, documents, computer files, sketches are produced during normal work activities or prior to starting training.	• The assessor can accept these sources in addition to observing the candidate in the work place.
Witness testimonies	• Expert witnesses may also provide evidence of working processes where an assessor is not able to be present to observe a candidate's performance for practical reasons (eg remote sites, work products no longer available for assessment).	• Indirect judgement • Depending on the witness' status or level of occupational competence, assessment expertise and familiarity with the national standards, other supplementary evidence may still be required to infer competence.
Professional questioning/ oral questioning	• Learners may also be required to show they have mastered the knowledge and understanding relating to the skill either through the performance of a particular task or it may be supplemented by oral questioning • Used to supplements skills assessment	• Direct judgement • Devised by the assessor to elicit further evidence for the range
Examinations	• multiple choice or open-ended items • delivered on paper or computer • When candidates do not achieve all of the knowledge outcomes, supplementary questioning to cover those gaps may be required	• Internal or external marking by the awarding organisation
Professional discussion	• Used as an opportunity for the candidate to explain certain behaviours and values relating to their work or how they carry out their work	• Assessor lead
Projects/ assignments	• Used when the candidate is required to produce evidence outside their responsibility (eg reviewing a department's operating procedures and making recommendations to management or may also include a multiple choice test that covers knowledge and understanding components of the standards).	• Assessor marked • May be internally or externally set and are generally carried out within a particular time spam, rather than directly observed by an assessor.

Assessment activity	Description	Assessor's role
Logbooks	• The evidence produced by the candidate is logged in a portfolio and referenced to the national occupational standards. This may include records of assessor observations, site visit reports, records of oral questions and candidate answers, photographs, videos, diaries, drawings, plans and so forth • Assessment forms are also included (assessor/ IV judgements) • All materials are cross-referenced in the portfolio according to a specified table of contents • Materials may refer to one or multiple criteria/ units • Used for verification purposes since the evidence can be collected and presented in this format • It is up to the centre to design the logbook, with a variety of approaches for recording assessor judgements and feedback being used (including holistic and analytic approaches) • The portfolio design is up to the training provider (centre) but the awarding body also provides guidance/ portfolio exemplars	• Assessor reviews evidence where this cannot be directly observed • The logbook may also be reviewed by the assessor, internal verifier or external verifier as part of the verification process • Composed of materials selected jointly by the candidate and his/ her assessor to reflect the candidate's work
Alternative assessment instruments	• Diaries may be used to record assessor feedback of occupational performance • Learner journey • Digital portfolios of work-based learning and assessment (see for example Project e-scape, TERU, 2010)	• Learner & assessor lead
Accreditation of prior and experiential learning (APEL)	• formal recognition of skills and knowledge learners already possess and may have been gained previously in structured or unstructured experiences and work • evidence may include work experience in the occupational setting, learning outside the work-place, incidental learning, intentionally planned learning, in-house company training, external courses (see also Walklin, 1991)	• Assessors compare evidence to assessment criteria before an award can be made

In order to achieve a unit and/or qualification, a complex combination of decision rules are applied. These include conjunctive and complementary procedures (Ryan & Hess, 1999; Chester, 2003). Conjunctive procedures require that all of the learning outcomes must be met for a 'pass' to be awarded. In addition, because in principle the framework allows for a unit to be substituted by another (according to the rules of combination), these equivalent pathways add a complementary rule to the conjunctive rule. Complex decision rules are also used to combine different types of assessment methods and tasks which can represent measures of the same or different constructs. Candidates have multiple opportunities to achieve the required standard, while they may also be required to demonstrate competence across a range. Better performance in some areas however cannot compensate other areas, which may not have been achieved. Accreditation for prior learning or experience may be accepted, normally as supplementary evidence. Table 3 summarises the decision rules to fulfil the requirements for achieving a competence-based qualifications.

Table 3: The approach used in 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 providing all learning outcomes are achieved, (although there may be some flexibility in terms of the assessment criteria)	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 is not achieved may be covered using additional instruments)
Multiple opportunities		Unlimited number of re-takes allowed	Evidence is collected until the standard is achieved
Accommodations and alternate assessments			Accessibility arrangements or using supplementary evidence

The requirement to achieve all of the learning outcomes, their assessment criteria for a range requires a large number of judgements to be taken over repeated occasions, which can lead to problems in ensuring the dependability of assessor decisions (Eraut et al, 1996, Wolf & Silver, 1986). In the new QCF, this may be increased by the requirement that units should be capable of assessment independent of other units, which although does not preclude the use of assessment designs that combine several criteria, it nevertheless raises issues for their assessment.

Due to the complex nature of these assessments, it is possible that one piece of evidence (eg performance, work exemplar) is used to contribute towards the achievement of multiple criteria, while competent performance needs to be demonstrated across a range of circumstances and occasions. The emphasis is on the opportunity for the candidate to provide sufficient evidence over a period of time, covering the required range on different occasions for the assessor to have the confidence that the candidate is competent.

Performance assessments and portfolios pose alternative challenges to measurement theories, in contrast to the traditional emphasis on standardization, one or more cut-scores to define the decision rule (Cizek, 2001), multiple-choice formats and automatic marking. While test-based decisions may be required, the description of these assessment methods highlights the role of the assessor (see Table 2, third column) and the inextricable relationship between assessment and instruction. The standard is contained in the work samples which represent minimally acceptable performance of the skill or craft rather than a cut-score (see Wolf, 1995). Variability in the interpretation of the standards and the complexity of the tasks/ activities are perceived as key threats to the validity (The Cambridge Approach, 2009) and reliability of this model (Murphy et al, 1995). It has been suggested that complex decision rules are important to validity and reliability (Chester, 2003). Since the recent introduction of the QCF however, to date there has been limited research published regarding the assessment regimes appropriate for these qualifications or on their measurement characteristics.

1.3 Quality assurance in assessment

It is the responsibility of the awarding organisation (AO) 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, AOs establish procedures which 'require the production of sufficient evidence from learners to enable reliable and consistent judgements to be made about the achievement of all the learning outcomes against the stated assessment criteria' (Ofqual, 2008, paragraph 5.3d, p26). The verification system aims to ensure the consistency of decisions, involving a complex set of relationships between assessors, internal verifiers (IV) and external verifiers (EV) and awarding organisations (while other stakeholders have a vested interest).

Each candidate is assigned one or several work-based or peripatetic assessors who are responsible for formally judging the candidate's evidence against the required assessment standards. They are required to select the most appropriate assessment methods, which meet prescribed quality criteria, help candidates identify opportunities to demonstrate their competence or produce evidence, especially when it is not possible to generate it as part of the normal work practice and when supplementary sources of evidence need to be generated (Fletcher, 1991). Assessors are also required to achieve relevant assessor qualifications for their role in order to be able to operate independently, participate in standardisation events and demonstrate that they are continuously updating their occupational competence and assessment skills. The role of the assessors emphasises the formative function of the vocational assessment which helps the learner to compare their performance to the standard required in their job roles.

The IVs are required to sample assessors' decisions/ judgements throughout the assessment process by directly observing the assessments carried out by the assessor or reviewing candidate evidence. The EV in turn is responsible, amongst other things, for ensuring that assessments are fair, consistent and meet the requirements set out by the national occupational standards. They will also sample decisions taken by assessors and IVs by observing staff or reviewing portfolio evidence. Internally centres employ various standardisation procedures, such as

assessor training, sampling of assessor decisions by the IV/ EV, access to a community of practice (Konrad, 1998). This may support the assessor in conceptualising the standard required and ensure this across centres/ regions. The verification system has been criticised for becoming a 'tick-box exercise' due in part to time and cost constraints, rather than following its intended purposes of checking the reliability of assessment results (Eraut et al, 1996; Wolf, 1998). However, in the current climate of radical change in the qualifications landscape, the verification systems used by AOs are also changing, so further studies are required (see also Konrad, 2000).

1.4 Vocational learners, learning and assessment

The QCF Regulatory Arrangements (2008) aims to answer the flexibility, diversity of circumstances and the wide range of learners who may take up these competence based qualifications. Learners are usually employed in the occupational field they are pursuing and the qualification is a route to confirming occupational competence. The assessment that takes place in employment (or simulated in a college) is continuous and takes on dual formative and summative functions, with a strong relationship between the assessor (who sometimes may also be the tutor) and their trainees, normally far fewer than is expected in a large scale assessment programme (see Brookhart, 2003 for a similar discussion on classroom assessment).

The many complex and important skills required in these occupations are learned to a large degree informally through apprenticeship-like methods that would involve observation of the target process, coaching, learner reflection and successive approximation, instead of didactic or formal teaching (Collins, Brown & Newman, 1989; Brown, Collins & Duguid, 1989). The learner practices the skill or conceptual knowledge continuously through carrying out tasks in a target domain. Coaching will involve guidance from one or several masters through the provision of scaffolding such as feedback, support, help or reminders until the learner is increasingly more independent. The learner will essentially practice the craft until mastery is achieved.

For most of the vocational qualifications, team collaboration is an important part of their work, when they may also have access to a senior member of staff (a master) or peers who may interfere with their performance. Such activities are part of the candidate's training or instruction. The assessment is not necessarily an isolated activity; it is the work performance itself that is judged. Learners also integrate unique prior work or life experiences and new information in order to construct ever more advanced understandings of the skill. Different learners may achieve the same criterion through undertaking different numbers or types of tasks, depending on each context.

For this reason, vocational learners are actively taking part in their learning and assessment, using assessor's continuous feedback. They take the initiative to devise their own work tasks and in collaboration with the assessor, select the evidence required to demonstrate competent status. This continuous feedback loop leads to a positive assessment culture (Wolf, 1995; Johnson, 2008a), where the aim is not to 'fail' the candidate, but to identify the gap between desired and actual performance and allow continuous on-demand ('when ready') assessment. Depending on their level, they are required to produce from routine to more complex activities or tasks, which are based on reflection, independent thought and increasingly more advanced application of skills, knowledge and understanding. Although funding may follow progress through the programme, the primary beneficiary of the assessment result or outcome is the learner. The issue is then whether the assessment activities in this context lead to consistent and accurate decisions that serve their uses and purposes well.

2 RELIABILITY OF DECISIONS

The previous section presented the context in which vocational assessors are required to judge whether an observed piece of evidence fits one or several defined criteria within the specified range, usually after the candidate has performed in the work-place. This section will focus on the reliability of these decisions and overall qualifications, factors that may affect estimates of reliability and its intrinsic link to validity.

2.1 Reliability of decisions in criterion-referenced assessments

In general, reliability can describe the credibility or limitations of the inferences we make about a set of assessment results (Mislevy, 1994). In the Standards of Educational Measurement, reliability is defined as "the consistency of […] measurements when the testing procedure is repeated on a population of individuals or groups" (AERA, APA & NCME, 1999, p25). Reliability of criterion-referenced assessments is then defined as the 'measure of agreement between the decisions made in repeated test administrations' (Swaminathan, Hableton & Algina, 1974, p264) which is the proportion of candidates who are classified the same way on two administrations of the same test (Traub & Rawley, 1980; Berk, 1980; Clauser, Margolis & Case, 2006). The literature surveyed proposes three forms of reliability evidence: decision consistency based on the cut score, the standard error of measurement near a cut score and inter/intra-rater reliability of the decisions for tasks where human judgements are required to decide a person's status with regards to a fixed standard (eg assessment criteria). A reliability study would thus evaluate the inferences made about candidate performance based on the decisions.

In the context of criterion-referenced competence based assessments reliability has been conceptualised as the consistency of assessor judgements (Hambleton & Novik, 1973; Greatorex, 2002; Greatorex & Shannon, 2003; Greatorex, 2005). The assessor decisions regarding a learner's performance on one or several outcomes and their related assessment criteria should be consistent achieved-not yet achieved classification decisions across learners (on the same qualification), across different assessors, providers and time (Murphy et al, 1995; Wilmut, Woods & Murphy, 1996). It is important to specify that reliability is specific to the decisions made by assessors across candidates, centres or regions for the intended purpose of the assessment. Candidates in the work place are continuously seeking to provide evidence that confirms their mastery of the skill over a period of time, so reliability of the decision could be about the assessor's level of confidence in his/ her judgement. Replication in this context would be that if we were to judge a candidate again, the same judgement should be taken based on the evidence provided.

Using statistical procedures, reliability indices can quantify the measurement precision of the assessment results if the procedure was to be repeated (Haertel, 2006). Since correct classifications can be distinguished from misclassifications, reliability estimates would indicate the degree of confidence that should be placed in the decisions (Traub & Rawley, 1991; Haertel, 2006). Other interpretations have looked at measuring the decisions across a number of conditions, such as assessors or tasks (Livingston & Lewis, 1995). An important point however is that the reliability of decisions can mean a number of things, and this meaning depends on their intended uses and interpretations. This in turn determines the way in which reliability is defined, quantified and reported (Traub & Rawley, 1980).

2.2 Validity and reliability

Reliability estimates the consistency of the measurement, the degree to which an instrument measures the same way each time it is used under the same conditions

with the same candidates. Validity, on the other hand, involves the degree to which the assessment measures what it is intended, and not something else. The accuracy and the consistency of scores do not encompass the broader issue of whether the scores or decisions represent the (true) ability the assessment intends to measure (Clauser, Margolis & Case, 2006). For example, using only multiple choice knowledge tests to assess a hairdressing unit may have highly reliable scores, but no validity.

Following the traditions of content and construct based models (see Kane, 2001, 2006), the validity of the interpretations of assessor judgements in competence based qualifications is conceptualised as the extent to which these assessments are 'close to the reality of vocational practice' (Wolf, 1995, p42). A candidate would be required to do the same thing in real practice as in the assessment situation (Messick, 1989; Wolf, 1995; Eraut & Steadman, 1998; Wilmut et al, 2003). If the assessment does not reflect with precision what is expected of the learners in view of the desired outcomes, the results cannot provide accurate or valid information about their performance (Webb, Herman & Webb, 2007).

The consistency of judgement in deciding when sufficient evidence has been provided to the same judge on several occasions or to several judges on one particular instance may impact on the reliability of that assessment decision, while validity may also be affected (Clauser, Margolis & Case, 2006; Wilmut, Wood & Murphy, 1996; Brookhart, 2003). The emphasis is on the assessment, on the type of evidence that would be accepted based on the performance descriptors and on the knowledge and understanding criteria so that assessors can make reasonably objective judgments about whether or not each person has achieved them (Wolf, 1995). Assessors in different contexts (candidates, centres) use different sources of evidence yet the same criteria to classify candidates. It is thus necessary for the decisions to be consistent in view of varying evidence when compared to a fixed outcome.

When human judgements are required to make interpretations of products or performances, validity has been considered to supersede the need for stringent reliability measures of the results (Moss, 1994). It has been further argued that when assessments conform to the assessment criteria specified in the outcomes, the assumption is that the decisions based on the unit specification will automatically be valid, hence comparable and thus reliable (Jessup, 1991, p192). Such an approach to reliability may be justified by the types of measurement methods which can be applied in the NVQ context (Eraut et al, 1996). An alternative strategy would be to propose accuracy and consistency estimates that can be used in the case of competence-based qualifications.

Brennan (2001a) applies the Generalizability theory (GT) to a situation comparable to that found in vocational assessment where tasks and raters vary across replications to show that GT '"blurs" arbitrary distinctions between reliability and validity […] and forces an investigator to concentrate on the intended inferences, whatever terms are used to characterize them' (Brennan, 2001a, p9). In other words, whether factors that can affect the inferences and stability of scores are in the realm of validity or reliability can be considered irrelevant as far as the researcher is able to define the universe of generalisation.

2.3 Is reliability of decisions important?

Concerns about the reliability of vocational assessments and qualifications have been raised since the early 1990s (Eraut et al, 1996; Wolf, 1998, p436). Where judgements are made about someone's occupational competence using complex

criteria (such as those presented in the unit of assessment) the reliability of the assessment decisions may be threatened (Murphy et al, 1995).

The standards on which these units and qualifications are based are intended to be precise enough for appropriate judgements to be possible (Wolf, 1995). Of concern is the ability of assessors to make judgements based on the standards and assessors are thought to compensate for the context in which the candidate operates: 'the key judgements have far more to do with whether someone has actually performed up to the assessor's standard than with the individual performance criteria at all' (Wolf, 1995, p69). An instance where considerations outside the written standards would be for example assessment criteria 2.7 in Table 1 above 'apply suitable products, when used, to meet manufacturers' and stylist's instructions'. However, it could be argued that this flexibility in judgement is a requirement for the system to be able to function across a wide range of contexts. Because of the flexibility required in assessing work-based performance, the consistency and reliability of these judgements have been questioned (Jessup, 1991, p 193).

In the context of classroom assessment, Shepard (2001) considers that reliability is not as important as for large-scale dated examinations because errors in judgements may be corrected soon after. In other views, reliability was less important compared to a strong validity argument. The general consensus in the measurement community is that reliability is important, despite specific concerns regarding the tools available to conceptualise and interpret reliability (Brennan, 1998; Brookhart, 2003; Johnson, 2006). Because it is within the remit of the AO to ensure accuracy and consistency of the assessment results, stringent quality assurance procedures are required of approved centres to enable audit trails (Ofqual, 2008). For this reason the system has been criticised for being dominated by heavy paperwork and for an emphasis on following a process rather than ensuring the consistency of the judgement reached (Wolf, 1995, 1998; Eraut et al, 1996). Reliability studies looking at the consistency or precision of these judgements are however seldom carried out for English competence based qualifications (Jessup, 1989, but see Murphy et al, 1995).

2.4 Factors that contribute to accuracy and consistency

Assessment against competence-based standards described in the unit involves collecting evidence and assessors making judgements on whether a particular criterion has been achieved or not. The general public needs to have confidence that, for example the electricians seeking certification are judged as competent only when able to safely practice a skill (or sub-skill) at a defined level. Sources of measurement error can affect the assessment results, and ultimately the dependability of the decisions based on assessor's judgement. Murphy and colleagues view reliability of assessor judgements as the extent to which '[the results] are free from influences which cannot be related to the valid interpretation of the requirements [of these qualifications]' (Murphy et al, 1995, p4). A number of factors can however influence the quality of the judgements made (Greatorex & Shannon, 2003).

Because repeated measures do not exactly equal one another, measurement always contains an amount of chance error, no matter how constrained the conditions under which the instrument is used (Carmines & Zeller, 1979). Measurement theory however does not view 'error' in the common sense of mistake, something wrongly done or untrue. By contrast, measurement error emerges when repeated measurements record varying results or are inconsistent. The concept of error in the context of educational measurement is to discourage too literal interpretations of observed candidate scores (Jarjoura, 1985).

Overall, competence in a specific occupation at a particular level should have the same features, regardless of context – in a criterion-referenced system, estimates of

reliability should show how stable a classification of competent/not yet competent is for each criteria, 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). Equally, reliability might be interpreted as the proportion of times that same decision would be reached using two parallel assessment instruments (Hambleton & Novick, 1973).

Approaches to evaluate reliability consider inconsistency in the types of assessment described here to be a type of error (Nichols & Smith, 1998). Reduced classification consistency may be the result of increased random, temporary or unsystematic errors which could affect a large proportion of candidates due for instance to candidates, assessors, tasks or situation (Eraut et al, 1996). Classification accuracy is affected by systematic error such as fatigue or centre resources. The interactions and distinctions among factors that may influence the quality of the decisions are central in a competence-based approach since they can have a direct impact on the judgements made based on the candidate's performance (Murphy et al, 1995).

2.4.1 Threats to the accuracy and consistency of assessor judgement decisions

For these assessment types, measurement error would be the result of disagreement amongst assessors (Murphy et al, 1995), task quality (Brennan and Johnson, 1995), the types of evidence (Greatorex, 2005), an assessor's skill and experience, task-sampling variability or authenticity (Shavelson, Gao & Baxter, 1993), occasion-sampling variability (Cronbach et al, 1997), the setting the assessment takes place - school or work based (Murphy et al, 1995; Greatorex, 2002), the unit content made up of learning outcomes and the associated assessment criteria (Driessen et al, 2005; Wolf, 1995), low consistency across tasks (Shavelson, Baxter & Gao, 1993; Parkes, 2000), and administration occasion, time or place.

Similarly the choice of assessment method, mode or scoring procedures, assessor influences, beliefs and occupational competence are threats to the accuracy of the results (Kane, 1982, 2006). Features of performance assessments or direct observations of work-based tasks, such as the context within which the learning and assessment takes place, unlimited opportunities to gather the evidence, extended time periods, collaborative work, choice of task or assessment type and centre resources, although with the potential to better represent the criterion being assessed, pose challenges to the standardisation of the administration and scoring of performance assessment (Lane & Stone, 2006). Furthermore, witness testimonies are one type of evidence which was suggested to lead to decreased levels of agreement between assessors (Greatorex, 2005). It should be noted that witness testimonies are normally used as supplementary sources of evidence, rather than the sole source of evidence.

Thus, if a certain candidate was tired and made many mistakes, the assessor's decision to consider the candidate as not yet competent may be inaccurate, because it does not represent the candidate's true ability. Candidate or centre related systematic errors are threats to the accuracy (ie validity) of the assessor's classifications, which are predictable and are not accounted for in a reliability analysis. By contrast, random errors in the consistency of assessor judgements influence the consistency (ie reliability) estimates of the decisions (Murphy et al, 1995; Wilmut, Wood & Murphy, 1996; Johnson & Johnson, 2009).

In this instance, a reliability study would establish whether, for example, the same person would qualify or not on two different occasions with the same or different assessors and across training providers using the same set of standards. The

problem is that, as Nichols and Smith (1998) observe, the consequence of not having a robust construct theory that defines all of these facets or sources of error can distort our expectations regarding the inconsistency in measurement.

Measurement studies have suggested that the amount of error acceptable however in an assessment programme is dependent on what is expected given the domain (eg science, mathematics, writing or work-based tasks) and on the interpretations made by users of the results (Kane, 2006). The extent to which a reliability coefficient is sufficient 'depends on the seriousness of the decisions being made with the test, and what can realistically be expected of a test in a given situation' (Subkoviak, 1988, p51). For competence based qualifications few studies have reported reliability measures, so historical comparisons will not be possible. Although essentially arbitrary, high reliability measures should be required from these high stakes decisions (Haertel, 2006). Because of the small number of studies on the consistency of assessor decisions, a consensus on the required levels of decision consistency and accuracy have not yet emerged for these assessment types (Greatorex & Shannon, 2003).

2.4.2 Threats to decisions based on test results

In the context of these qualifications, it is also important to differentiate among the reliability of scores obtained using different assessment types which are used to make achieved/ not yet achieved decisions. Studies looking at the reliability of clinical oral examinations found that reliability of scores is affected by content sampling (eg number of items for each criteria) or the number of examiners (Berk, 1980). In a generalizability study, Wass, Wakeford, Neighbour and Van der Vleuten (2003) show that extending testing time to four 20-minute oral examinations, each with two examiners, or five orals, each with one examiner would improve intercase and pass/fail reliabilities. Another study investigating the reliability of clinical oral examinations, Daelmans and colleagues (2001) showed that the reliability when using a number of orals is better than the reliability of the common single oral examination. In the case of written examinations (eg short answer, essays, reports, logbooks), research in the field of medical examinations has found that a large number of examinations but with shorter questions would be required for satisfactory reliability estimates (Day et al, 1990). In the case of multiple choice or true/ false questions, advanced measurement methods exist to estimate the reliability, to equate scores across forms or time, but it has been suggested that for pass/fail decisions of the type described here, one-best answer questions are most appropriate (Case, Swanson & Ripkey, 1994).

2.4.3 Ways to improve the accuracy and consistency of assessor judgement

Despite certain reservations (see above), quality assurance processes (verification, standardisation), frequent training of assessors, occupational expertise, exemplars of work that meet the standard and networks of assessors have been proposed as key requirements to increase the consistency of assessor judgements for a particular qualification, eg of individual assessors across a number of candidates (inter-rater reliability), across assessors (intra-rater reliability) within a centre or across centres/ regions. (Wolf, 1995; Konrad, 1998; Johnson, 2008a, b).

Since reliability measures are not typically produced, several studies have looked at the effectiveness of these procedures using qualitative methods (eg Eraut et al, 1996; Greatorex, 2002; Greatorex & Shannon, 2003). Greatorex (2002) used a questionnaire design to establish best practice in the standardisation methodology used by a representative number of centres, which are known to also positively affect the reliability or consistency of these decisions. Although these studies were based on qualitative methods to establish the factors which may affect assessor decisions and did not seek to measure reliability estimates of these decisions, the authors

found that standardisation activities did not ensure agreement between assessors. They go on to identify a number of issues which have the potential to affect the consistency of these judgements, such as sampling activities of live performance observations, the relationship between assessors and candidate, resources available (including the location of assessors in relation to their candidates), the training available which may seek agreement between assessors rather than levels of uncertainty allowed in the process or sufficiency of evidence accepted by assessor.

Peregrine and colleagues have also raised issues about the different sampling methods of assessor judgements (Peregrine et al, 1994). Internal verification systems may apply a risk-rating matrix when deciding how much monitoring is required, such that less experienced or not yet qualified assessors are classified as being at higher risk of producing inconsistent judgements and so they are monitored more closely by senior assessors/ the IV.

Konrad (1998) proposes that the training of IVs using communities of assessment practice may result in more consistent and comparable decisions. Social interaction between assessors is usually facilitated through networks of assessors (Wolf, 1995), the EV/ awarding organisation and membership to other professional organisations. The requirement for assessors to undertake continuous professional development (CPD) in some areas is intended to maintain the currency of their occupational skills, which may create a shared level of competence, but may not necessarily enable them to learn from each other or share a repertoire of resources (see Wenger, 1998). The effectiveness of such networks in ensuring consistency of assessment decisions however has not been measured so far.

Vocational learners are actively taking part in the assessment process to prove that they are occupationally competent, which would imply that they could also become members of the same community of practice: ‘summative assessment requires that teachers (or other assessors) become members of a community of practice, while formative assessment requires that the learners become members of the same community of practice’ (Wiliam, 1998). The reliability goal is stable information about the gap between candidate performance and the standard expressed in the unit as well as in that particular source of evidence. Reliable classification of candidates into discrete categories as well as assessor feedback on specific areas of improvement are important features of a good quality assessment system. Reliability in this context is about the consistency of decision making given sufficiency of evidence for the purposes expressed in the standards (Wilmut, Wood & Murphy, 1996).

2.4.4 Combining multiple measures into a portfolio or qualification result

It is possible that the type of criterion-referenced assessments presented here, while measuring a common construct or skill, may be made up of multidimensional outcomes, with unidimensional individual criteria. Assessor decisions are based on a candidate’s performance on each of the criteria, so the reliability and validity of decisions for each criterion are of interest (Hambleton & Novick, 1973). Furthermore, reliable assessment decisions from a number of different assessment types is about converging or accumulating the evidence that support the same inference rather than joining of scores (Mislevy, 1994). The final decision to certify a candidate taking a vocational qualification is the result of an assessment process based on multiple occasions and multiple measures, including candidate artefacts, tutor observation and tests rather than a single administration of a test. The use of multiple measures in portfolio assessment have the potential to influence the estimates of the reliability (and validity) of these results (Mislevy, 1994). However, some of these measurements may be more reliable than others, since they are not equally created. When each piece of a portfolio for instance is scored and then aggregated into a composite score, reliability levels may increase (Reckase, 1995).

Further, qualifications consist of several units of assessment. In this case, a qualification result would be less reliable when 'it is internally inconsistent or equivocal, or […] we realise that securing additional information would cause us to revise our beliefs substantially' (Mislevy, 1994). The results integrated from multiple sources would not be exchangeable or repeatable in the sense described by traditional test theories. This is further complicated by the fact that candidates are signed off only when deemed competent, implying that they are able to attempt a particular task an unlimited number of times since multiple attempts at the same task in these types of assessment may reduce the measurement error (Rudner, 2001; Clauser, Margolis and Case, 2006).

Both the reliability of individual measures and how they are combined are important for the total classification reliability since the misclassification errors will be a function of the error scores associated with each measure (Cronbach & Gleser, 1957). Due to the complex set of decision rules combining conjunctive and complementary procedures in these qualifications, established methods for estimating the reliability of single scored tests are not able to estimate the reliability of non-scored multiple assessment measures based on these rules. Combining multiple measures conjunctively may also affect reliability (Cronbach et al, 1997). Different methods able to evaluate the classification reliability of these qualifications are then required (Douglas, 2007; Good, 2002).

In this sense, 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. Using his analogy, reliability of an instrument concerns its credibility, how appropriate it is for the inferences made based on the scores, indirect measures of the construct. In conjunction with other measures, it can lead to a correct classification that meets the purposes of a qualification and so the more measurements are available, the higher the reliability of the decision based on the measurements. If reliability were interpreted in this way, there would be no validity without reliability and vice versa.

3 MEASUREMENT MODELS AND THEIR APPLICABILITY TO COMPETENCE BASED ASSESSMENT

The types of assessments used for employment and licence to practice discussed here are ultimately used to classify persons or decide whether they are competent and not yet competent for each of the criteria (Clauser, Margolis & Chase, 2006). Inconsistency or error in judgement will affect the reliability of the decisions and would imply that there is lower classification accuracy and consistency.

Approaches to studying the precision of the dichotomous decision situation, where the correct classifications can be distinguished from misclassifications, have been concerned with the consistency of the decisions or with an examination of losses (see Traub & Rowley, 1980; Berk, 1980). The literature looking at similar assessment methods (eg performance, portfolio, classroom assessments) argue however that less developed measurement theories for this type of assessment may have drawn conclusions based on methods better suited to other assessment programmes (eg Traub & Rawley, 1980; Nichols & Smith, 1998; Brookhart, 2003; Haertel, 2006; Clauser, Margolis & Case, 2006).

In addition, the decision rules applied in the competence-based qualifications system with multiple conjunctive hurdles (each component is required to be a 'pass' for an overall pass to be awarded) determine the type of reliability study that is suitable in the circumstances. Different context, purposes (both formative and summative), methods (norm vs criterion referenced) and group size may mean that models applied to large scale testing are less useful (Mislevy, 1994; Brookhart, 2003; Johnson & Johnson, 2009). In this context, section 1 shows that measurements of a number of binary criteria are used to make binary decisions regarding each outcome. Once all of the outcomes are achieved, the decisions are used to decide a person's mastery status, as either achieved or not yet achieved a unit. Candidates are assumed to complete a variable number of tasks for the purposes of achieving the same number of outcomes and criteria. The verification system applied for competence-based units requires the IV to sample a proportion of the assessors' live or portfolio decisions. A single assessor however takes the majority of candidate performance decisions.

This section aims to review alternative reliability measures that may be more appropriate for the assessment types used in competence-based vocational qualifications developed in the context of classical test theory or generalizability theory.

3.1 Reliability studies of NVQs

Although no examples were found in the literature of methods suitable to estimate the classification reliability for assessments with varying number or type of tasks (a candidate may perform different tasks to achieve a fixed outcome), a number of studies of NVQs have used classical test theory methods, analysis of variance or qualitative methods of investigation such as interviews, questionnaire and field studies (see Wilmut, Wood & Murphy, 1996; Greatorex, 2000; Johnson, 2006 for reviews). This work has been important in understanding the challenges imposed by work-based assessment in the context presented by these qualifications, but has provided limited evidence for estimating the consistency of these decisions. Limited access to assessment data, logistical issues and time constraints have restricted advances in educational measurement theory of assessor decisions for these qualification types.

Murphy and colleagues (1995) have published a first reliability study involving 31 centres across five qualifications (although no details of the analysis or tables of findings could be reported at the time, see Wolf, 1998). In their brief summary of findings however, they suggest that there was good assessor agreement when reviewing portfolio evidence, despite the fact that most disagreements were due to lack of sufficient evidence, poor presentation or authenticity. There were no concluding findings in the case of assessor decisions based on live observations of performance. Several other outcomes of the study are to do with both the validity and the reliability of the classification decisions, but in the absence of more specific detail it is difficult to evaluate the contributions of this study to a conceptualisation of reliability in competence based qualifications such as the NVQs studied here.

In a later study, Eraut and colleagues (1996) look at the consistency of assessor judgements using qualitative techniques, including interviews, open-ended survey and fieldwork. The respondents to their survey expressed a commonly accepted view of reliability of such decisions which may be affected by the subjectivity of human assessors. They further cast doubt whether the standards expressed in the units of assessment lead to consistent interpretations (p66). Although this is not a measurement study as such, it offers useful background to issues which may affect the validity and reliability of the classification decisions, including the way in which policy may affect the practice of assessment.

Using borderline portfolio evidence relating to one unit of assessment, Johnson (2008a) investigates the cognitive strategies that underpin assessors' holistic judgements of a vocational qualification (delivered in a college) graded portfolio using a think aloud task, a modified Kelly's Repertory Grid interview technique to elicit assessors' perceptions about characteristics of each assessment criteria. An observation of a moderation meeting is used to identify issues which may affect the consistency of judgements. The author concludes that classification consistency was influenced by assessor's shared experiences and differing perspective. Adopting a sociocultural perspective in the traditions of Wenger (1998) and Engestrom (2001), Johnson (2008b) uses similar qualitative and ethnographic methods to identify features that characterise assessors within a community and factors which may affect the consistency of their decisions. Despite limitations of these pilot studies, the author succeeds in describing the context of these qualifications and in identifying the threats to the consistency of decisions which may impact on their accuracy as well.

In a study investigating whether the consistency of assessor judgements is affected by the types of evidence, Greatorex (2005) collected assessors' analytic decisions on two fictitious borderline portfolios as well as the assessor's comments, devised by an assessor to be 'just competent' or 'not yet competent'. An analysis of variance (ANOVA) was carried out to identify any interactions between the different types of evidence, assessor observation, witness testimony, personal statement and written underpinning knowledge questions and answers and found most disagreement in judgements of witness testimony. The study has however a number of limitations, including a reduced sample size (two fictitious portfolios containing 177 decisions were reviewed by 15 and 12 assessors respectively) which is generally related to statistical significance level. Despite this, the study offers a useful method for researching assessor decisions based on portfolio assessments.

In an earlier study using a questionnaire, Greatorex (2002) has highlighted a mixed picture with regards to the standardisation methods used (eg training involving identifying good and bad practice, discussions, feedback). Based on these results, Greatorex and Shannon (2003) devise a standardisation exercise using historical candidate portfolio samples to establish whether these procedures affect the consistency of assessor decisions (inter and intra-rater reliabilities). The methodology

used included a survey, interviews with assessors and candidates and standardisation exercises using anonymous candidate portfolios matched for ability. To delve into the effects of a standardisation procedures, assessors were asked to record their assessment decisions on four portfolios prior to the standardisation event as well as afterwards, including recording reasons for their decisions. The study found varied levels of agreement, depending upon the classification of each portfolios used (borderline not yet competent, borderline just competent). This small-scale study identified a number of issues with the feasibility of data collection in these qualifications, as well with an understanding of how much disagreement should be acceptable amongst assessors.

The study assumes that assessors are make holistic judgements regarding a number of criteria which would make any analysis at the level of the criteria untenable (Wolf, 1995). More recent theories of decision making have proposed a 'two-systems' model of human judgement which arises through a combination of intuitive (parallel system) and analytic processes (Evans, 2003; Kahneman, 2003). These complex decision problems may be reduced to a set of relatively simple component judgements which are then combined (Kleinmuntz, 1990). Recent neuropsychological studies have shown that emotional contextual cues and framing effects have an important role to play in human decision making since they communicate knowledge elements which allow optimal decisions to be made under uncertainty (De Martino et al, 2006, 2008). Relevant to a component judgement approach, although individual demands of judgement of each component may be reduced when compared with a holistic model, may be affected by random errors associated with the assessment procedure, method, task, context, number of assessors, assessor background and the interaction between assessor and candidate. Across tasks and assessors, measurement errors may have a cumulative effect. A suitable measurement framework would have a built-in 'error control mechanism' that is able to help achieve greater consistency (Kleinmuntz, 1990).

3.2 Decision-consistency measurement studies

Indices such agreement amongst independent raters (assessors), decision-consistency coefficients and generalizability coefficients characterise the weight that can be placed on the scores and are not addressed by usual reliability indices. When criterion-referenced classifications concerned, the probabilities of misclassification, ie of placing candidates in the right or wrong category, are considered to be more informative than SEM or reliability coefficients (see Haertel, 2006). The current characteristics of competence-based assessment need to take into account the following factors:

- The measurement method used is criterion-referenced (Traub & Rowley, 1980; Hambleton & Novick, 1973)
- The different item formats (beyond multiple choice);
- Its intended uses, purposes and stakes of the results (Brookhart, 2003; Nichols & Smith, 1998);
- Group size may be large, but often small;
- The type of scores used are continuous but most frequently binary (pass/fail);
- Tasks or items are dichotomously scored;
- Tasks may vary across candidates (type, number);
- Assessor(s) make most of the judgements, but there may also be onscreen multiple choice machine marked tests;
- The number of parallel test forms used, task variability, usually small number of items on knowledge tests but with large numbers of criteria assessed through observation;
- Measurement error is correlated across items/ task scores;
- Assessment is on different occasions, continuous;

- The decision rules (conjunctive/ disjunctive/ compensatory);
- There is no limit in the number of re-takes, assessment is continuous and may be naturally occurring, “when-ready” assessment;
- Multiple sources of measurement error, mainly to do with assessor judgments in performance and portfolio assessment tasks;
- The assessment has formative, summative or both functions;
- The context of instruction and assessment;
- Domain is that of occupational assessment (hairdressing, plumbing, etc)

All these and other issues may influence the confidence that should be placed in a person’s score (Traub & Rowley, 1980; Haertel, 2006).

Studies investigating the reliability of classification have proposed methods based on either classical test theory or generalizability theory methods, suitable for either norm or criterion referenced tests, with continuous score distribution where a number of cut-scores are used to assign candidates into categories or for dichotomous variables which are used to make a binary decision (see reviews in Berk, 1980; Traub & Rawley, 1980). Few studies have however proposed methods which could be used regardless of the type of scores, number and type of task used or for complex decision rules (but see Douglas, 2007; Livingston & Lewis, 1995; Nichols & Smith, 1998; Brookhart, 1998). The focus of this review is on dichotomous assessor judgements leading to a binary decision, as well as the reliability of decisions based on continuous test scores which will be used to develop an approach suitable for binary decisions of competence based assessments.

3.2.1 Classical theory methods

To estimate consistency, the true scores are used to estimate the distribution of classifications on two independent, parallel forms (see Berk, 1980). For single test administrations, estimates of internal consistency such as KR20 or coefficient alpha can be used. Coefficient alpha is used as an estimate of reliability of test scores equal to the ratio of true score to total observed score, hence error of test scores directly influences the reliability index. This type of reliability estimate is thus dependent on the variation of total scores appropriate to norm-referenced tests. As such, the literature on reliability differentiates between measurement models for criterion referenced and norm referenced assessment methods for recognising achievement, which rule out internal consistency measures for criterion-referenced tests (Popham & Husek, 1969). Later studies however take advantage of the possibilities provided by advances in testing technology to suggest that internal consistency estimates may still be suitable (Kane, 1986).

In the context of classroom assessment, Buckendahl, Yang and Ferdous (2003) evaluate the level of agreement between reliability analyses using coefficient alpha as a measure of internal consistency and a proposed decision consistency strategy (percent agreement) that uses teacher judgments of student proficiency on four levels and a written assessment that empirically classifies performance as the two assessments. The study design included four classification categories of proficiency, two representing the standard and two not yet meeting the standard to examine whether the decision consistency was equivalent across these categories. Their analysis shows however that the decision consistency values were low, and conclude that internal consistency values provide better evidence for these scores. However, the discrepancy between these values may be due to the study design which asked teachers to classify candidates into categories defined by performance descriptors they were not familiar with or were subject to interpretation. This study shows in a sense that when the concept of what is measured, eg competence of the standards, it is not well defined, this may result in inconsistent decisions.

The verification system applied for competence-based units requires the IV to sample a proportion of the assessors' live or portfolio decisions. A single assessor however takes the majority of candidate performance decisions. For these qualifications, a study of the reliability of a single assessor's (rater) judgements of candidate performances is therefore required where some sources of evidence may be naturally occurring and therefore different number and type of work-based tasks for each candidate. In classical test theory, reliability coefficients suggested as suitable for single rater's judgements are interclass correlations, test-retest (intraclass correlation) and parallel-forms reliabilities (see Haertel, 2006).

Alternate forms (equivalent tests or parallel forms) and internal consistency designs assume item or task independence. Yet gathering of evidence in the work-place would make the construction of parallel items or entire tests untenable (Nichols & Smith, 1998). In such cases, the probabilities of misclassification (of wrongly passing or failing a candidate) have been proposed as more appropriate, than SEM or reliability coefficients more suitable to continuous scores instead (Haertel, 2006).

Employing a simulation approach, Bradlow & Wainer (1998) use classical test theory to examine the consistency of pas/ fail decisions based on subjectively scored performance tasks when all the candidates are re-scored, only the failures are re-scored or score only those above a cut score. They conclude that based on this model, only those around the cut score should be re-scored if there is an equal proportion of passes and fails and only the fails should be rescored if the pass rate is really high. Furthermore, random assignment of the pass/fail condition leads to lower error rate than the use of test scores.

Few studies have reported internal consistency measures of portfolio assessments. In the context of classroom-based writing skills assessment, Nystrand, Cohen and Dowling (1993) report low to moderate reliability estimates (near .50s) of total scores judged holistically. They obtain relatively higher values however when judges were asked to mark by task and with a shared understanding about what was expected. However, they recognise several confounding factors, including issues with the domain (expository writing) and the scoring procedures which may affect the reliability estimates. Studies of the reliability of the Vermont Portfolio Assessment Programme have reported similar interrater correlations from .46 to .63 depending on how the scores were aggregated (eg within or across scoring dimensions, by task or across sections of the portfolio) (Koretz et al, 1993). These studies suggest that scores of individual tasks or entries in the portfolio will not have high reliability and that task variation is bound to be large due to the fact that students do not answer prompts but create writing exemplars to meet a number of criteria. In addition, inadequate rubrics, lack of rater training or the opportunity for standardisation can negatively impact reliability of test scores (Koretz et al, 1993; Nystrand, Cohen & Dowling, 1993).

These studies have investigated tests with continuous scores. In the context of standards based criterion referenced assessment programmes where there is no limit in the number of candidates meeting the standards, information provided by standard errors and decision consistency may be more informative (Linn & Burton, 1994). For these assessment types, measurement procedures using categorical scores, leading to a classification into two or more discrete categories, different measurement methods may be appropriate to express reliability. Such methods use different statistics from those described so far, depending on whether the tasks are dichotomously scored (Brennan & Kane, 1977), multiple cut scores are used, or the items are scored dichotomously and are not equally weighted (Livingston & Lewis, 1995), polytomous items or assessments using multiple raters (Brennan & Wan, 2004).

Feldt and Brennan (1989) provide a battery composites scores reliability based on the average reliability of the subtest (portfolio entries) scores and the average inter-correlation between the subtests. Applied in the context of a non-experimental data set, Reckase (1995) concludes that although good reliability of scores is achievable when the entries measure the same thing, this is annihilated by prohibited costs of double marking.

A number of studies use classical test theory to study decision reliability of standard or percentile scores on simulated data sets (Rogosa, 1999; Klein & Orlando, 2000). These studies concluded that higher test score reliability was associated with lower measurement error. The study shows that the classification consistency rises as the cut score moves away from the average score and as the pass rate moves away from .50 (Klein & Orlando, 2000).

Furthermore, it has been suggested that when items require complex responses (including those used in occupational performance), candidates' performances or judgements of their performances will be less stable across multiple occasions or tasks which can be attributed to multiple or different focal constructs (Nichols & Smith, 1998). For these assessment types it can be expected to obtain lower estimates of test-retest (Murphy et al, 1995). The world of workplace assessment is made up of unrepeatable observations as well as non-exchangeable sources which need to be interpreted in light of the standards (Mislevy, 1994).

3.2.2 Strategies for estimating the consistency of categorical pass/fail decisions based on a cut score

Two authoritative reviews by Berk (1980) and Traub and Rawley (1980) present a conceptualisation and associated procedures suitable for reliability of criterion-referenced scores used for decision-making. These studies categorise the ways of assessing the losses due to decision errors. The choice of reliability category is dependent on certain assumptions, interpretations and uses of the indices (Berk, 1980). Thus due to the conditions attached to competence based assessments the procedures requiring two administrations such as threshold loss function p_o and kappa (Hambleton & Novick, 1973; Swaminathan, Hambleton & Algina, 1974) may not be suitable for evaluating dichotomous decisions by assessors but they may be useful when using tests with continuous scores and dichotomous, equally weighted items with a single cutting score that determines the mastery-non-mastery state. The proportion of consistency of false-positive and false-negative classifications is suitable to this context. Reliability then refers to the consistency of mastery-non-mastery decisions over repeated test administrations (Hambleton & Novick, 1973). The administration of two tests, although possible for externally set tests may be difficult due to security and group sizes. Other limitations with the statistical conditions constrain their interpretation however (Berk, 1980). Furthermore, in an application of these approaches, Subkoviak (1980) found that these single methods were difficult to compute and resulted in biased estimates for short tests (n=30) and their properties make them unsuitable to criterion-referenced assessments used for classroom decision making (Berk, 1980).

Correlations of test scores which measure the consistency of mastery-non-mastery classifications that are consistent with two test administrations, while avoiding the necessity of multiple test administrations have also been suggested (Huynh, 1976; Subkoviak, 1976). Huynh's method is based on fitting a two-parameter binomial model to observed scores (Huynh, 1976). Huynh's coefficient uses linear regression true-score estimates, including the internal consistency Kuder-Richardson 20, to calculate the probability of getting an item right for an examinee at each observed score, and then assuming a binomial distribution the probability of getting a specific

number of items right for each observed score, and ultimately obtaining probability of consistent classification for the entire group. However, it assumes that student's knowledge or the administration conditions remain unchanged across testing series and it does not always providing a fit to the observed data (Berk, 1980). Brennan and collegues (Hanson & Brennan, 1990; Brennan, 2004) propose instead a four-parameter beta binomial model to be better for certain score distributions.

The Brennan-Kane index (Brennan & Kane, 1977; Brennan, 1980) provides information about the consistency of scores in relation to a cut score and can be useful for placement tests. It is conceptually similar to true-score variance discussed earlier and it is a useful measure since it reflects measurement error as a result of the position of the cut-score on the observed scores scale. It also weighs differently measurement errors due to false-positive or false-negative classifications. Because it includes distance from the cut score, it is sensitive to the cost associated with misclassifying persons depending on where on the true score scale is their score in relation to the cut score. The index is a function of both the length of the test and the cut score. Due to certain limitations, Brennan and Kane (1977) recommend using other measures such as standard error of measurement.

Estimates of domain score provide useful information about a candidate's level of knowledge or skill (see Berk, 1980). Individual specific statistics consist of two estimates of standard error for each individual that can be used to set up a confidence interval around each individual's observed proportional score. Group specific statistics consist of averages of individual statistics over persons. Estimates for each of these categories requires a large number of items to supply stable measurements and they may not be appropriate on their own, but may be useful to programme evaluation decisions.

Because the statistics reviewed here require an assumption about test forms, in the case of work-based performance or portfolio assessments, the notion of parallel forms should be investigated. Similarly, in the world of occupational assessment, tests cannot be standardised and apply decision rules which may not be suitable in these circumstances.

Macready and Dayton (1977) propose two probabilistic models suitable to dichotomous items that assumes candidates to either be masters with Platonic true scores of 1, and nonmasters with Platonic true scores of 0. Scores to each of the n item form a vector of n item scores for each candidate. The strategy provides probabilities of each vector occurring as the sum between two types of events – the candidate is a master but answered item incorrectly due to forgetting or that the candidate is a non-master but answered items correctly due to guessing. Macready and Dayton consider two options as well – either the probability of forgetting or guessing varies for each item or it is the same for all items. One of the advantages with this proposal is that the model can be made to fit the data while it can also compute the probabilities of classification errors as a measure of the goodness of the decision rule (cut score). However, the model assumes an all or none approach to responding to an item. Wilcox and Yeh (1979) propose estimates of the parameters of a latent trait model when the skills represented by items are hierarchically related.

3.2.3 Univariate Generalizability theory methods

Generalizability theory (GT) is a random sampling model based on assumptions of linear modelling. While the observed score may include a number of components, eg to do with candidate's attitudes, administration occasion, the rater or the test form used, classical test theory does not allow for the measurement of these multiple sources of error. By contrast, the main advantage presented by GT is that it is able to estimate the precision of measurement when this is affected by multiple sources of

error (Haertel, 2006; Johnson & Johnson, 2009) by applying certain ANOVA methods (Brennan, 1992). The G-coefficient is analogous with the reliability coefficient in classical theory. It evaluates the ability of a measurement procedure to locate individuals on an absolute or criterion-referenced scale (although the framework can also be applied to relative scales) and so suitable for criterion-referenced assessment methods of the type used in vocational assessment (Johnson & Johnson, 2009). Reliability for each type of measurement is quantified by G-coefficients and by an index of dependability (Brennan, 1992). In addition, the G-coefficient is able to indicate how reliable the instrument can locate individuals around a cut-score on the measurement scale. The universe score and error score estimates are based on variance of components (the amount of variation within all the values of that variable). In the case of pass-fail decisions this can indicate the level of misclassifications. The coefficient however is affected by the D-study design (Brennan, 1992).

Classical analysis of variance (ANOVA) is used to calculate stardard error of measurement (SEM) and G-coefficient as well as weightings of the influence the facets have on results SEM produces confidence intervals around a person's estimated score (rather than an overall error estimate for the population), considered useful in the case of work-place assessment decisions (Johnson & Johnson, 2009). The universe of generalisation may include facets such as tasks, assessors, IVs, assessment method as well as interactions between them (eg candidate x task x assessor). Because some of the facets relate to the validity of the scores, the model can be used to provide evidence for both the consistency and accuracy of results (Shavelson, Baxter & Gao, 1993). In the end, a decision study allows the researcher to predict the G-coefficient and SEM if any of the facets were to change in future applications of the instrument (eg increase or decrease the number of assessors, assessment occasions, select a particular assessment method over another).

Internal consistency is measured through a cross design of two or more facets, for example persons with tasks and occasion (p x t x o), one associated with stability indicators and another with internal consistency indicators of reliability, based on inter-correlations of the tasks in a test across occasions (Nichols & Smith, 1998). When persons are given the same items on two separate occasions, a person-by-task-by-occasion design allows for the evaluation of variation or error in performance from several sources, such as occasion and task unrelated to the construct.

Investigating variability in science performance assessment scores, Shavelson and colleagues (Shavelson, Baxter & Gao, 1993, Ruiz-Primo, Baxter & Shavelson, 1993) show that task and occasion as well as the choice of certain assessment methods are important sources of error. Since in these studies candidates performed the science assessments on only one occasion however, Cronbach, Linn, Brennan and Haertel (1997) argued that it was not possible to separate error due to task or occasion, or a combination of both. In a subsequent G-study design, correlations across task x person x occasion x rater were performed to show that person performance over occasions is unstable (Shavelson, Ruiz-Primo & Wiley, 1999). This indicates that using several assessments to measure performance in a domain may increase their estimated variance component. The study also highlights that candidate performance of complex tasks may not be stable across occasions due to different response strategies applied by students across occasions and tasks. This variability is not necessarily due to error but due to changes in the construct and this may make the measurement procedure inadequate (Nichols & Smith, 1998).

Some advantages presented by the model, including the absence of the requirement for the assumptions of dimensionality and independence, suitability to criterion-referenced methods, the ability to provide weightings across candidates, assessors,

context or tasks, and their correlations mean that it is an approach embraced by many studies investigating criterion referenced performance assessment carrying similar features to the type described here. GT's other advantageous feature is the possibility to add more levels of one or several facets (which may have been identified with a low G-coefficient) such as assessors or tasks.

Indices reflecting classification accuracy, proportions of agreement among assessors, decision-consistency coefficients and generalizability coefficients (Cronbach et al, 1972) provide important information for these types of assessment that may not be evident from typical indices of reliability, eg Cronbach alpha or Kuder-Richardson 20 coefficient, standard error of measurement (Clauser, Margolis & Case, 2006; Haertel, 2006). Brennan and colleagues (Brennan and Johnson, 1995; Brennan, 2001b) have examined the role of various facets in assessing the generalizability of performance assessments. By applying analysis of variance techniques, generalizability theory allows for the error to be broken down into multiple sources of error, such as person, task and rater. This method has yet shown only a limited degree of across-task generalizability.

Consistency of responses is normally expected in interpreting GT coefficients as well as classical estimates, which may present a number of shortcomings for complex item types (Nichols & Smith, 1998) or where tasks cannot be standardised either by narrowing the domain or adding more tasks (Brookhart, 2003). Nichols and Smith maintain that 'anyone who employs a design within one of these traditional approaches is interpreting consistency or inconsistency in test taker performance using a theory of learning and performance that dictates the conditions over which the focal construct is expected to change or remain unchanged' (Nichols & Smith, 1998, p25).

3.2.4 Multivariate GT

Multivariate GT provides a framework for assessing measures that involve fixed facets (Cronbach et al, 1972; Brennan, 2001b; Clauser, Harik & Margolis, 2006). In addition to the variance components that characterise the universe score and the error score, this method includes both variance and covariance components (ie how much two facets change together). It is a suitable model where the measurement errors are correlated across items or tasks, as for example in situations where multiple scores are influenced by the same response. This correlation is important to both the reliability and the SEM of composite scores produced as a weighted combination of component scores.

Clauser, Harik and Margolis (2006) apply this model to medical clinical and interpersonal skills performance assessment using standardised patients. Examinees are scored on four dimensions using different rating scales. The authors found that this type of analysis provides a more complete analysis of the test scores and conclude that seeking to identify which sources of variance contribute to measurement error is more important than whether the assessment method is absolute or relative. However, a number of other facets not accounted for in the design, such as the patient simulated cases or tasks.

3.2.5 Construct-centred reliability

Based on an application of the generalizability theory, Nichols and Smith (1998) develop the trait and cognitive process models that incorporate different theories of learning and performance in a reliability study of the 1992 NAEP Performance Assessment of Writing. Construct-centred reliability is based on the application of a theory of learning and performance to interpret the consistency or inconsistency of test-taker performance across conditions such as tasks, occasion, content or judges. The focus is on the meaning of test scores to interpret consistency or inconsistency

in observed responses across conditions associated with multiple measurements. Using writing performance tests, Ying Hong and Smith (2000) provide an empirical example of the model to show that high generalizability (or reliability) is possible for these scores where there is good understanding of the construct, and so it can be a useful approach to validate performance assessment.

3.2.6 Baysian Estimation (Hambleton & Novick, 1973)

The reliability of polytomously scored items on a test with multiple cut scores is estimated by Wainer and collegues (2005) using a Bayesian approach (Beguin & Glas, 2001). In this approach, a previous distribution is combined with current data to estimate the probability of future distribution. Based on MCMC procedures, the model can obtain estimates of ability for each examinee, and then depict the probability of passing by proficiency. Given its application with polytomous items, the approach appears to be suitable for estimating classification reliability for multiple measures (Douglas, 2007).

3.2.7 Multiple-attempt, single response IRT models (MASI)

Spray (2007) differentiates between single attempt multiple item tests (SAMI) and multiple attempt, single items (MASI) tests. MASI are binomial trails models in which the number of successes, rather than trials, are fixed and for which the number of successes to failures is essential. Relating to the competence based assessment where learners practice until the standard is achieved, a one-success form of this model is answer until correct (Spray, 2007).

3.2.8 Argument approach to reliability (Parkes, 2007)

Following a conceptualisation of reliability that consists of both social and scientific values, Parkes (2007) follows the concepts developed by Kane (2006; see also Chapelle & Jamieson, 2010) 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 replicability in the context, investigating the evidence and finally constructing an argument for or against the inferences made. The model allows for the assessor for example to sample evidence over a period of time to identify trends in performance, rather than collect responses over two administrations of a test. 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' (p5). The interpretative argument proposed by Parkes certainly merits attention since it 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.

3.2.9 Estimates of the accuracy of decisions

Classification accuracy will be determined by the reliability of scores, the score distribution, the pass rate (proportion of candidates that meet the standard or are at or above the true score) and where the cut-off score is set (Clauser, Margolis & Case, 2006). Techniques used to estimate the accuracy of decisions based on true-score methods would estimate the proportion of candidates whose true score where on either side of a cut score. Livingston and Lewis (1995) suggest a method suitable for estimating both the accuracy and consistency of decisions based on a cut score regardless of the scoring system used. The reliability of the score is used to estimate effective test length in terms of discrete items, then the true-score distribution is estimated by fitting a 4-parameter beta model. The conditional distribution of scores on an alternate form, given the true score, is further estimated from a binomial distribution based on the estimated effective test length. Agreement between classifications on alternate forms is estimated by assuming conditional

independence, given the true score. However, the approach was considered inadequate to situations where complex measures are combined (Douglas, 2007).

Using data from medical licensure, Clauser, Margolis and Case (2006) show that when multiple distinct tests are used, the false-positive rate decrease, but the false-negative rate increase. The example is almost intuitive – the more opportunities there are to prove competence, the more certain we can be that the person's profile is accurate, but at the same time to more opportunity to fail one of the instruments.

Applying the techniques outlined by Livingston and Lewis (1995) and Haertel (1996) to tests other than multiple choice items of Mathematics and English, Young and Yoon (1998) estimate the decision accuracy and consistency for each cut score for each cluster reported in the tests, and for the composite total scores for Mathematics, Writing and Reading. The overall probability of consistent classification is given by the sum of the probabilities as being below and above the cut points for each cluster within the composite (multiple hurdles) by both their true (conditional probabilities) and observed scores on two forms of a test.

Other approaches to misclassification that could be used to address measurement error include the AP Reliability-of-Classification Procedure, which is a variant of the Livingston and Lewis procedure (College Entrance Examination Board, 1988, Appendix A), and applications of classical test theory and extensions of decision theory (Kupermintz, 2004). Kupermintz (2004) proposes a reliability measure based on proportional reduction in loss and applies this to a large scale assessment.

Using item response theory, Rudner (2001) describes an expected classification accuracy index for determining the accuracy of classifications of examinees into score categories when multiple opportunities to pass a test are allowed. Martineau (2007) expand on this proposal by evaluating the index as it is likely to be used in practice (as a point estimate of classification accuracy) to provide a related index of expected proportion in each score category and derive a measurement error of expected proportion in each category. Although Rudner (2001) analysis was found inappropriate (Douglas, 2007), Martineau shows that although Rudner's index as a point estimate is slightly positively biased, it is nevertheless useful for groups with reasonably large numbers of examinees.

3.3 Strategies recommended for estimating reliability and consistency of decisions

Based on the body of research evidence presented here, a number of strategies can be used, depending on assessment type. As noted previously, the reliability of assessor decisions or test scores is influenced by the context of assessment and by particular decision rules, which increase the misclassification rate of these decisions when compared to compensatory rules.

Alternative strategies for estimating decision consistency and reliabilty suitable for multiple measures with these characteristics hold most promise for the vocational assessment context (eg Douglas, 2007; Spray, 2007). The approach suggested by Livigston and Lewis (1995) although difficult to compute may be appropriate where assessments incorporate a number of scoring procedures. Further, the construct-centred approach proposed by Nichols and Smith (1998) based on an application of generalizability theory emphasises that the important role empirical evidence and theoretical rationales have in defending the domain and the expectations regarding the uses and interpretations of scores. Other contributions may also prove useful to the context outlined in this review (eg Douglas, 2007), but this would depend on the type and candidate records available for analysis.

4 CONCLUSIONS AND FURTHER RESEARCH

Characteristics of the vocational education system discussed earlier (continuous assessment, learner-assessor-peers interactions, formative and summative functions) suggest that in this context, tasks or activities are linked together and cannot be assumed to be independent (Murphy et al, 1995). Although affinities between teacher or classroom assessment and competence-based assessment are obviously present, the latter presets different challenges to do with instruction in the work-place or a simulated environment of a practical skill or craft, funding regimes which follow progression, accountability of these qualifications, combining different assessment types and measures into a unit outcome (pass/ fail) using conjunctive decision rules. The learner has a choice to use work-place performance as evidence in their portfolio if deemed sufficient and he or she will continue gathering evidence until sufficiency is achieved. A useful measurement model would therefore need to take account of the context the assessment takes place, the links between tasks, small candidate numbers assessors normally work with, the links between validity and reliability, and the formative and summative assessment purposes.

The models and methods reviewed 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. A number of authors suggest approaches that use statistical models to analyse forms of the consistency of assessor judgements using several measures for describing the precision of assessor decisions, estimate the effects of SEM, indices that reflect the classification accuracy or consistency of individual assessments as well as of the composite scores.

Strategies for investigating the consistency and accuracy of assessor decisions in vocational education are understudied and developments in this area are timely considering changes in the regulatory environment which influence the way professional qualifications are delivered and assessed.

5 REFERENCES

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

Béguin, AA, & Glas, CAW (2001). MCMC estimation and some fit analysis of multidimensional IRT models. Psychometrika, 66, 471-488;

Berk RA (1980). A consumers' guide to criterion referenced test reliability. JEM, 17, 4;

Berk, RA. (1980). A Framework for Methodological Advances in Criterion-Referenced Testing. Applied Psychological Measurement 4: 563-573Baume et al, 2002

Bradlow ET & Wainer H (1998). Some statistical and logical considerations when rescoring items. Statistica Sinica, 8, 713-728;

Brennan RL & Kane M (1977). An index of dependability for mastery tests. J of Ed Meas, 38(4), 295-317;

Brennan RL & Wan L (2004). A bootstrap procedure for estimating decision consistency for single-administration complex assessments. Paper presented at the annual meeting of the NCME, San Diego, CA;

Brennan RL (1992). An NCME Instructional module on Generalizability theory. Retrieved from http://www.ncme.org/pubs/items/21.pdf;

Brennan RL (1998). Misconceptions at the intersection of measurement theory and practice. Educational Measurement: Issues and Practice, 17(1), 5-9;

Brennan RL (2001a). An essay on the history and future of reliability from the perspective of replications. JEM, 38(4), 295-317;

Brennan RL (2001b). 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 http://www.unl.edu/buros/biaco/pdf/pres03buck02.pdf;

Carmines EG and Zeller RA (1979). Reliability and validity assessment. SAGE, Issues 7-17;

Case SM, Swanson DB & Ripkey DR (1994). Comparison of items in five-option and extended-matching formats for assessment of diagnostic skills. Academic Medicine, 69(Suppl10), S1-3;

Chapelle CA, Enright MK & Jamieson J (2010). Does an Argument-Based Approach to Validity Make a Difference? Educational Measurement: Issues and Practice, 29(1), 3-13;

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

Cizek G (2001). Standard setting: Concepts, methods and perspectives. Mahwah, NJ, Lawrence Erlbaum;

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

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;

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;

Daelmans HEM, Albert, Scherpbier JJA, van der Vleuten CPM, Donker AJM (2001). Reliability of clinical oral examinations re-examined. Medical Teacher 2001 23:4, 422-424;

Day SC, Norcini JJ, Deserens D, Cebul RD, Schwartz J, Beck LH et al (1990). The validity of an essay test of clinical judgment. Academic medicine, 65(Suppl 9), S39-S40;

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; doi:10.1523/JNEUROSCI.2895-08.2008

De Martino B, Kumaran D, Seymour B & Dolan RJ (2006). Frames, Biases, and Rational Decision-Making in the Human Brain. Science 313 (5787), 684. [DOI: 10.1126/science.1128356]

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;

Engestrom Y (2001). Expansive learning at work: toward an activity theoretical reconceptualisation. J of Education & Work, 14(1), 133-156;

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

E-scape project, TERU, Retrieved from http://www.gold.ac.uk/teru/projectinfo/;

Evans JS (2003) In two minds: dual-process accounts of reasoning. Trends Cogn Sci 7:454–459
FAB/JCQ (2010). Writing QCF Units: How much detail to provide, Guidance Note 3, Version 1, March;
Feldt LS & Brennan RL (1989). Reliability. In RL Linn (Ed), Educational measurement (3rd ed, p105-146). New York: Macmillan Fletcher, 1991
Fletcher S (1991) NVQs, Standards and Competence, Kogan Page: London.
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;
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;
Haertel EH (2006). Reliability. In RL Brennan (Ed) Educational Measurement, pp65-107;
Hambleton RK & Novik MR (1973). Toward an integration of theory and method for criterion referenced tests. JEM, 10, 159-170;
Huynh H (1976). On the reliability of decision in domain-referenced testing. JEM, 13, 253-264;
Jarjoura D. (1985). Tolerance Intervals for True Scores. Journal of Educational Statistics, 10(1), 1-17;
Jessup G (1989). The concept of Reliability in the assessment of NVQs. In JW Burke, Competency based Education and Training , London, Falmer Press;
Jessup G (1991). Outcomes. NVQs and the emerging model of education and training. London, Falmer;
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;
Kahneman D (2003) Maps of bounded rationality: psychology for behavioral economics. Am Econ Rev 93:1449 –1475.
Kane (1986). The role of reliability in criterion –referenced tests. J of Educational Measurement (JEM), 23(3), 221-224;
Kane MT (1982). A sampling model for validity. Applied Psychological Measurement, 6, 125-160;
Kane MT (2001). Current concerns in validity theory. Journal of Educational Measurement, 38, 319–342
Kane MT (2006). Validation. In RL Brennan (Ed) Educational Measurement, pp17-64;
Kingston P (March 2007). No Accounting for Taste, The Guardian, Retrieved from http://www.guardian.co.uk/education/2007/mar/06/furthereducation.uk2;
Klein SP & Orlando M (2000).CUNY's testing program. Characteristics, results and implications for policy and research. MR-1249-CAE. Santa Monica, RAND;
Kleinmuntz DN (1990). Decomposition and the control of error in decision-analytic models. In RM Hogarth (Ed), Insights in decision making: a tribute to HJ Einhorn, The U of Chicago Press, Chicago and London;
Konrad J (1998). Assessment and Verification of National Vocational Qualifications: a European quality perspective. Education on-line, Retrieved from www.leeds.ac.uk/educol/index.html;
Konrad J. (2000). Assessment and verification of national vocational qualifications: policy and practice. Journal of Vocational Education & Training, 52(2), 225-243. doi:10.1080/13636820000200117
Koretz D, McCaffrey D, Klein S, Bell R & Stecher B (1993). The Reliability of scores from the 1992 Vermont Portfolio Assessment Program. CSE Technical Report 355, RAND Institute on Education and Training/CRESST, Retrieved from http://www.cse.ucla.edu/products/Reports/TECH355.pdf;
Kupermintz H (2004). On the Reliability of Categorically Scored Examinations. JEM. 41(3), 193-204;
Lane S and Stone CA (2006). Performance assessment. In . In RL Brennan (Ed) Educational Measurement, pp387-430;
Linn RL & Burton E (1994). Performance-based assessment: Implications of task specificity. Educational Measurement: Issues and Practice, 13(1), 5-8;
Livingston SA & Lewis C (1995). Estimating the consistency and accuracy of classifications based on test scores. Journal of Educational Measurement, 32, 179-197;
Macready GB & Dayton CM (1977). The use of probabilistic models in the assessment of mastery. J of Ed Statistics, 2, 99-120;
Martineau JA (2003). An Expansion and Practical Evaluation of Expected Classification Accuracy. Applied Psychological Measurement 31, 181-194
Mislevy RJ (1994). Can there be reliability without 'reliability'? ETS, Princeton, NJ;
Mitchell L (1989). The Definition of Standards and their Assessment. In John Burke (Ed), Competency based education and training, p54-65, The Falmer Press;
Moss PA. (1994). Can There Be Validity Without Reliability? Educational Researcher 1994 23: 5-12;

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=
Nichols PD & Smith PL (1998). Contextualising the interpretation of reliability data. Educational Measurement: Issues and Practice, 17(3), 24-36;
Nystrand, M., Cohen, A. S. & Dowling, N. M. (1993). Addressing Reliability Problems in the Portfolio Assessment of College Writing. Educational Assessment, 1(1), 53-70. doi:10.1207/s15326977ea0101_4
Ofqual (2008). Regulatory arrangements for the Qualifications and Credit Framework. Ofqual/08/37/26;
Parkes J (2000). The relationship between the reliability and cost of performance assessments. Ed Policy Analysis Archives, 8(16);
Parkes J (2007). Reliability as argument. Educational Measurement: Issues and Practice, 26(4), 2-10;
Peregrine P, Pedreschi T, Connor J, Thackray D & Wolstencroft T (1994). Effective practice in assessment against the management standards. Research and Development Series, Report No 24, Department of Employment, Sheffield;
Popham WJ & Husek TR (1969). Implications of criterion referenced measurement. JEM, 6, 1-9;
Project e-scape, TERU, 2010. Retrived from http://www.gold.ac.uk/teru/projectinfo/projecttitle,12370,en.php;
QCA (2009). Developing units and qualifications for occupational competence in the Qualifications and Credit Framework Additional guidance to support sector skills councils and standards setting bodies. QCA/09/4261;
QCDA (2008). Guidelines for Writing Credit-Based Units of Assessment, Version 4, QCDA/10/4725;
QCDA (2009). What is QCF? Retrieved from http://www.qcda.gov.uk/19674.aspx;
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;
Reckase MD (1995). Portfolio Assessment: A theoretical estimate of score reliability. Educational Measurement: Issues and Practice, 14(1), 12-14;
Rogosa D (1999). Accuracy of individual scores expressed in percentile ranks: classical test theory calculations. CSE Technical Report 509, NCRE, Standards and Student Testing;
Rogosa, D. R. (1994). Misclassification in student performance categories. Appendix to CLAS Technical Report. Monterey, CA: CTB/McGraw-Hill.
Rudner L (2001). Computing the expected proportions of misclassified examinees. Practical Assessment, Research & Evaluation, 7(14);
Ruiz-Primo MA, Baxter GP, & Shavelson RJ (1993). On the stability of performance assessments. JEM, 30, 215-232;
Ryan JM & Hess RK (1999). Issues, strategies and procedures for combining data from multiple measures. Paper presented at an annual meeting of AERA, Montreal;
SEMPTA, 2010, Retrieved from http://www.semta.org.uk/training_providers__awarding/national_occupational_standard/115_uses_for_nos.aspx;
Shavelson RJ, Baxter GP & GAo X (1993). Sampling variability of performance assessments. JEM, 30, 215-232;
Shavelson RJ, Ruiz-Primo MA, Wiley EW (1999). Note on Sources of Sampling Variability in Science Performance Assessments. JEM, 36(1), 61-71;
Shepard LA (2001). The role of assessment in teaching and learning. In V Richardson (Ed), Handbook of research on teaching (4th Edition), p1066-101. Washington DC: AERA;
Spray JA (1997). Multiple-attempt, single-item response models. In W.J. van der Linden & R.K. Hambleton (Eds.), Handbook of modern item response theory. New York: Springer-Verlag
Subkoviak MJ (1976).Estimating reliability from a single administration of a mastery test. JEM, 13, 265-276;
Subkoviak MJ (1988). A practitioner's guide to computation and interpretation of reliability indices for mastery tests. JEM, 25, 47-55;
Subkoviak, M.J. (1988). A practitioner's guide to computation and interpretation of reliability indices for mastery tests. Journal of Educational Measurement, 25, 47-55
Swaminathan, H., Hambleton, R.K., & Algina, J. (1974). Reliability of Criterion-Referenced Tests: A Decision-Theoretic Formulation. Journal of Educational Measurement, 11(4), 263-267Traub RE & Rawley GL (1980) Reliability of Test Scores and Decisions. Applied Psych Meas 4 (4), 517-545;
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;

Univeristy of Cambridge Local Examinations Syndicate (2009). The Cambridge Approach. Retrieved from http://www.bulats.org/docs/cambridge_approach.pdf;
Vygotsky, L. (1978). Mind in society. Cambridge: Harvard University Press
Walklin L (1991). The assessment of performance and competence. Stanley Thornes Publishers, England;
Wass V, Wakeford R, Neighbour R & Van der Vleuten, C (2003). Achieving acceptable reliability in oral examinations: an analysis of the Royal College of General Practitioners membership examination's oral component. Medical Education, 37(2), 126-131, DOI: 10.1046/j.1365-2923.2003.01417AERA, APA, & NCME, 1999
Webb, N. M., Herman, J. L., & Webb, N. L. (2007). Alignment of mathematics' state-level standards and assessments: The role of reviewer agreement. Educational Measurement: Issues and Practice, 26, 17-29;
Weigel, T., M. Mulder & K. Collins (2007). The concept of competence in the development of vocational education and training in selected EU member states. Journal of Vocational Education and Training, 59(1), 51-64;
Wenger E (1998). Communities of practice: learning, meaning and identity. CUP;
Wilcox RR & YEh JP (1979). On Latent Structure Models for Measuring Achievement on Hierarchically Related Skills. CSE Report 124, Retrieved from http://www.cse.ucla.edu/products/reports/R124.pdf;
Wiliam, D. (1998). Enculturating learners into communities of practice: Raising achievement through classroom assessment. Paper presented at the European Conference for Educational Research, University of Ljubljana, Slovenia;
Wilmut J, Hamer J, Macintosh H, Murphy R & Warmington P (2003). Fair assessment of NVQs. CDELL, U of Nottingham, Nottingham, Retrieved from http://www.nottingham.ac.uk/shared/shared_cdell/pdf-reports/fairassessofnvqs.pdf;
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, http://www.nottingham.ac.uk/shared/shared_cdell/pdf-reports/relexam.pdf;
Wolf A & Silver R (1986). Work based learning: trainee assessment by supervisors. R&D Report 33, Sheffield Manpower Services Commission;
Wolf A (1995). Competence-based Assessment. Open University Press: Buckingham;
Wolf A (1998) Portfolio Assessment as National Policy: the National Council for Vocational Qualifications and its quest for a pedagogical revolution. Assessment in Education, Policy and Practice, 5(3), 413-445;
Ying Hong J & Smith PL (2000). A construct-centered generalizability model. Analysing underlying constructs of cognitively complex performance assessments. Paper presented at the annual meeting of the AERA, New Oreleans, April;